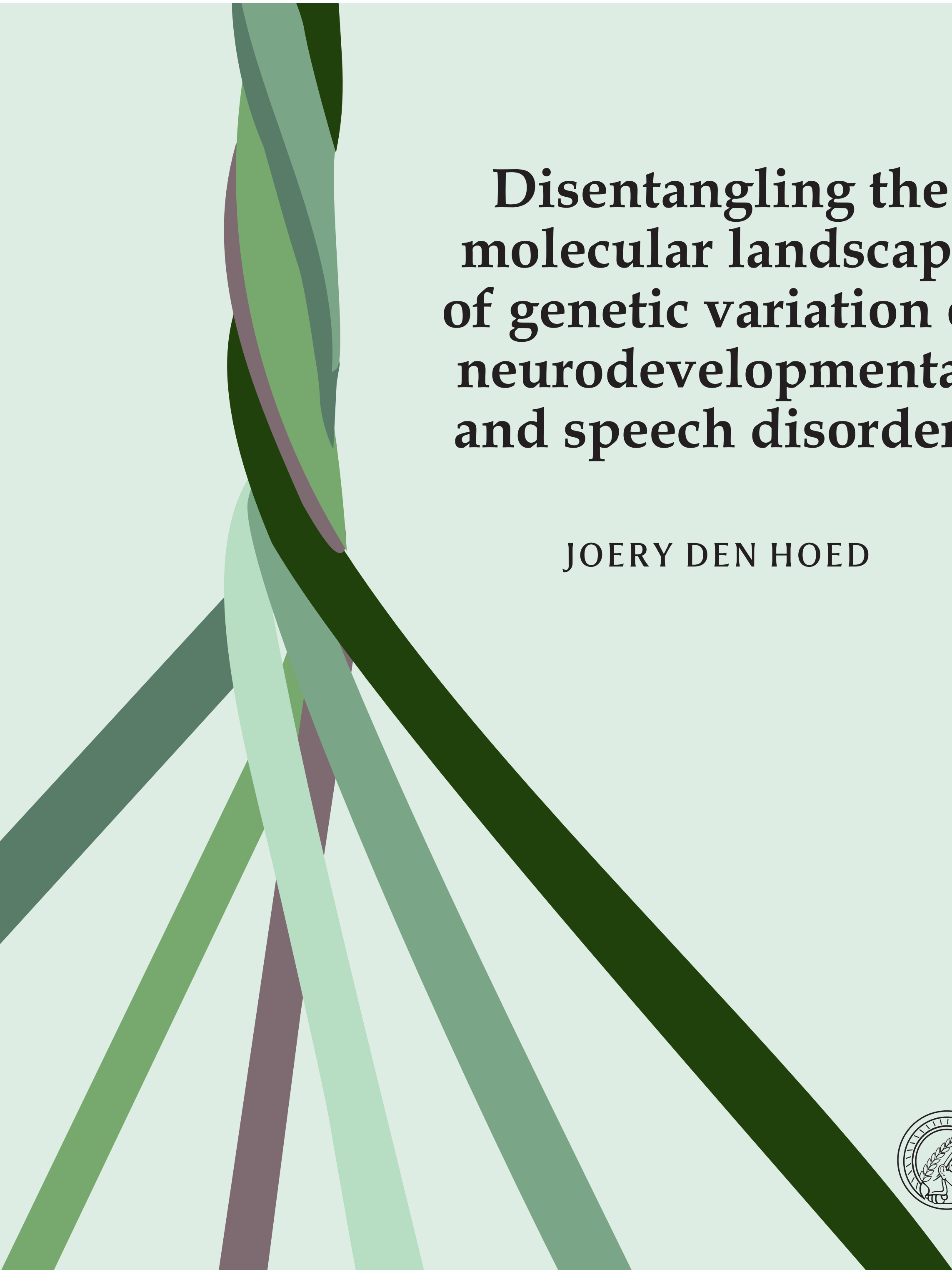
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS



Disentangling the molecular landscape of genetic variation of neurodevelopmental and speech disorders

JOERY DEN HOED

Disentangling the molecular landscape of genetic variation of neurodevelopmental and speech disorders

Funding body
This research was funded by the Max Planck Society for the Advancement of Science (www.mpg.de/en) in the form of an IMPRS for Language Sciences PhD Fellowship (2017-2021) awarded to Joery den Hoed.

International Max Planck Research School (IMPRS) for Language Sciences
The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics
Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

© 2022, Joery den Hoed
ISBN: 978-94-92910-43-1
Cover design by Joery den Hoed and Lucía de Hoyos González
Formatting by Joery den Hoed
Printed and bound by Ipskamp Drukkers, Enschede

All rights reserved. No part of this book may be reproduced, distributed, stored in a retrieval system, or transmitted in any form or by any means, without prior written permission of the author. The research reported in this thesis was conducted at the Max Planck Institute for Psycholinguistics, in Nijmegen, the Netherlands

Disentangling the molecular landscape of genetic variation of neurodevelopmental and speech disorders

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op maandag 14 november 2022
om 12.30 uur precies

door
Joery den Hoed
geboren op 17 februari 1993
te Vlissingen

Promotoren
Prof. dr. S.E. Fisher
Prof. dr. L.E.L.M. Vissers

Manuscriptcommissie
Prof. dr. J.H.L.M. van Bokhoven (voorzitter)
Dr. S. Cappello (MPI, München, Duitsland)
Prof. dr. T. Bourgeron (Institut Pasteur, Parijs, Frankrijk)

Disentangling the molecular landscape of genetic variation of neurodevelopmental and speech disorders

Dissertation

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M van Krieken,
according to the decision of the Doctorate Board
to be defended in public on Monday, November 14, 2022
at 12.30 pm

by
Joery den Hoed
born on February 17, 1993
in Vlissingen, the Netherlands

Supervisors
Prof. dr. S.E. Fisher
Prof. dr. L.E.L.M. Vissers

Manuscript committee
Prof. dr. J.H.L.M. van Bokhoven (chair)
Dr. S. Cappello (MPI, Münich, Germany)
Prof. dr. T. Bourgeron (The Institut Pasteur, Paris, France)

Contents

General introduction 11
Genetics of human speech disorders... 13
How can we connect clinical diversity with genetic variation at molecular level?... 15
The strengths of high-throughput functional assays.................................. 16
Physiologically-relevant cellular model systems... 19
The advent of gene-editing tools 22
Aims and relevance of this thesis... 23
Outline of the thesis.. 23
References 25

Genetic pathways involved in human speech disorders......................33
Abstract 34
Introduction... 35
Molecular perspectives on speech - the example of *FOXP2*...................... 35
Genomic screening of disorder cohorts identifies novel risk variants......... 38
Insights from gene-driven studies.. 39
Effects of speech-related regulatory genes on early brain development.... 41
Acknowledgements 43
References 43

Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction 49
Abstract 50
Main text... 51
Methods.. 58
Acknowledgements 64
References 65
Supplemental information.. 68
Supplemental notes... 88
References supplemental notes.. 95

Inherited variants in *CHD3* show variable expressivity in Snijders Blok-Campeau syndrome 97
Abstract 98
Introduction... 99
Results.. 99
Discussion 106
Methods.. 108
Data availability 112
Acknowledgements 112
Ethics declaration 112
References 112
Supplemental information.. 116

Supplemental notes 1 126
References supplemental notes 1 132
Supplemental notes 2 133
References supplemental notes 2 136
Supplemental notes 3 137
References supplemental notes 3 144

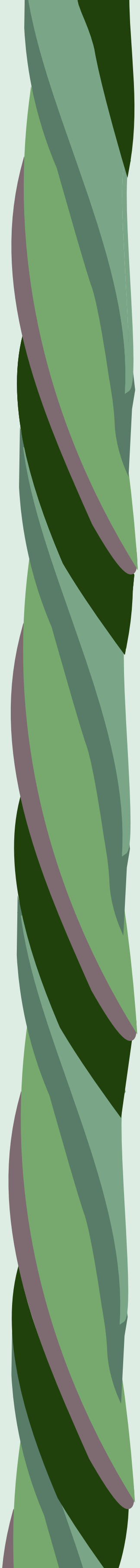
Deciphering the roles of *CHD3* during early human brain development using cerebral organoids 147
Abstract 148
Introduction 149
Results 150
Discussion 160
Methods 163
References 167
Supplemental information 170

Molecular networks of the FOXP2 transcription factor in the brain 207
Abstract 208
Introduction 209
Regulation of *FOXP2* expression 212
Downstream effectors of FOXP2 215
FOXP2 transcriptional regulation 216
Future perspectives 219
Acknowledgements 220
References 220

General discussion 231
Summary of the results 233
Contributions to the field 234
Functional testing of genetic variants 236
Integration of clinical and functional data 239
Future perspectives 240

Appendices 245
English summary 247
Nederlandse samenvatting 251
Acknowledgements 257
Curriculum vitae 263
List of publications 265

1



General introduction

Genetics of human speech disorders

A genetic basis for human language

Language is a complex human-specific ability that is acquired at a young age with little effort and without formal teaching, assuming adequate exposure[1]. Typically, by the age of 2.5 years, a child can speak over a thousand words[2], and by four years is able to produce complex sentences[3]. The drive to communicate is not confined to hearing children, as deaf babies babble with their hands and develop sign language without direct instruction[4; 5]. Language is critical at the level of society for human cultural evolution, allowing the transfer of knowledge over generations[6], while at the individual level, language skills impact socio-emotional development and educational achievement[7]. While some non-human species signal with simple messages, such as alarm or identification calls[8], or are even able to learn to produce more complex sounds from a tutor, a behaviour called vocal learning[9], this contrasts with the human capacity to structure and combine specific sounds to create an infinite number of meanings[8]. The innate character[1], the cognitive skills required for receptive and expressive language[10], and the lack of complex language in other species[8] suggest that these human capacities have a genetic basis.

Family studies support genetic contributions to speech and language acquisition

Initial studies in twins revealed high heritability for speech and language impairments (0.5 or more)[11-14]. Consistently, speech and language disorders have been found to run in families[15], with a positive family history increasing risk of developing speech and language problems, although it is important to recognise that shared environmental factors could also lead to such familial clustering[16]. While individual variability in speech and language skills may partly be explained by multifactorial inheritance with a combination of environmental factors and a complex genetic background of common variants with a small effect size[17], a subset of developmental speech and language disorders are caused by rare high-penetrant variants in single genes[18-21].

Mendelian causes of severe speech disorders

The first discovery of a rare single nucleotide variant causing a severe monogenic speech disorder was made in a large multigenerational family in which the inheritance of the phenotype followed an autosomal-dominant pattern[18]. All affected family members were diagnosed with childhood apraxia of speech, a disorder that is characterised by difficulties with the sequencing of the orofacial movements required for fluent speech. Moreover, individuals with the disorder also had problems with language processing and grammar, affecting both expressive and receptive domains[18]. Linkage analysis pinpointed a region on chromosome 7 which co-segregated with the phenotype and contained multiple genes[22; 23]. In a second independent case with a similar phenotype, a *de novo* balanced reciprocal translocation was found in that same region, disrupting the *FOXP2* gene[18]. FOXP2 is a transcription factor expressed in various brain regions during development[24], and subsequent sequencing of the gene resulted in the discovery of a p.R553H missense variant in all affected family members, predicted to disrupt the DNA binding domain[18]. Follow-up studies showed that this missense variant abolishes DNA binding, and results in a loss-of-function, although these investigations also raised the possibility of dominant-negative effects[25]. Since the initial association of *FOXP2* with human developmental speech disorder, more disruptions in the gene have been identified in speech phenotypes[19; 26]. These only seem to explain a small proportion of Mendelian forms of speech disorders, as putative pathogenic variants in *FOXP2* have not been identified in the most recent phenotype-driven studies of childhood apraxia of speech[20; 21].

Rare high-penetrant variants in severe speech disorders – genetic heterogeneity

The emergence of next-generation sequencing (NGS) strategies to time- and cost-effectively read the full human exome or genome has transformed the field of genetics of developmental disorders in the last two decades – now making it possible to investigate the genetic cause in individuals without a family history[27]. Leveraging the *de novo* paradigm[28; 29] and bi-allelic[30; 31] strategies in large cohorts of individuals with neurodevelopmental disorders (NDDs) such as intellectual disability and autism, has led to the identification of many new genes harbouring rare high-penetrant variants that could be associated with human phenotypes[28; 32]. These studies, and more in-depth follow-up investigations, show that Mendelian NDDs are not always genetically homogeneous, and that overlapping phenotypes can be caused by genetic variants in different genes, such as Noonan (MIM #163950 / #605275 / #609942 / #610733 / #613224 / #613706)[33] and Rett syndrome (MIM #312750 /# 613454)[34].

The field of genetics of speech and language disorders has lagged behind the progress that has been made for NDDs. This may in part be due to the difficulties of comprehensively characterising phenotypes of individuals with speech problems using standardised criteria[15]. However, a small number of phenotype-driven cohort studies on individuals with developmental language disorder and childhood apraxia of speech (some without a family history) have identified novel risk variants in different genes[20; 21; 35] (described in detail in Chapter 2). Although the discovery of aetiological variants in different loci points to genetic heterogeneity for Mendelian speech disorders, several of the affected genes seem to belong to a shared co-expression module with high expression during early human brain development[20; 21], suggesting that they may converge on specific pathways. This is consistent with findings that genes affected in genetically heterogeneous NDDs are implicated in specific and conserved pathways as well[36].

Phenotypic variability – speech disorder versus neurodevelopmental disorder

Interestingly, when genes associated with severe speech disorders are followed up by gene-driven studies (e.g. by using clinical gene-matching strategies to search for probands identified by different teams around the world), additional aetiological variants are discovered in individuals with a wider range of developmental phenotypes. *BCL11A*[37; 38], *CHD3*[20; 39], *POU3F3*[40] and *WDR5*[20; 41], are all examples of genes for which the discovery of a likely pathogenic variant in a case of developmental speech disorder led to the identification of variants associated with broader NDD syndromes in other cases.

So far, the current literature does not show evidence for the existence of genes that are highly specific to human speech and language disorders, but rather supports the idea that the biology underlying speech and language impairments has overlaps with that involved in NDD, in particular with intellectual disability and autism[41]. Although *FOXP2* stands out, with aetiological variants having disproportionate effects on speech and language development, broader effects on cognitive skills have been described in some family and case studies, showing that the associated phenotype is not limited to speech problems in every affected individual[19; 26]. These findings are consistent with recent studies showing the lack of autism-specific genes[42], and suggest that phenotypes such as severe speech problems and monogenic forms of autism should not be considered separate clinical entities with distinct molecular diagnoses, but rather features that fall within the spectrum of NDDs with shared biological underpinnings.

In this thesis, we aimed to study the mechanistic differences that underlie such variability in clinical phenotypes, and to explore functional model systems that could help to better understand the molecular links between human speech disorders and NDDs.

How can we connect clinical diversity with genetic variation at molecular level?

As noted above, rare high-penetrant variation in genes linked to speech disorders seem to cause, in some cases mild phenotypes in which speech and language skills are disproportionately affected, while in other individuals it results in more severe and wide-ranging phenotypes, where speech impairments are one part of a broader syndrome. Aetiological variants in these genes have likely pleiotropic effects, which means that while the phenotype may partially overlap, variation in the same gene can also lead to distinct seemingly unrelated phenotypic features in different individuals. Another factor that could contribute to the wide spectrum of observed phenotypes associated with variants in genes linked to speech disorders is variable expressivity, referring to the variable effects of genetic variation on the phenotype in different individuals with the same genetic condition. However, these phenomena are still poorly understood at the molecular level (and further explored in Chapter 4).

For some genes linked to NDDs, clear genotype-phenotype correlations have been identified. An example is the *POLR2A* gene, which is associated with a disorder that includes mild to severe developmental delay and hypotonia (MIM #618603), with different types of variants resulting in different severities of the phenotype. Using a severity-scoring approach, *POLR2A* protein-truncating variants were found to cause the mildest phenotypes, while individuals with missense variants that exert a dominant-negative effect were identified to present with more severe symptoms[43]. In the *SCRAP* gene, the location of the variants has been shown to be important. *SCRAP* protein-truncating variants in the FLHS domain result in a disorder characterised by a short stature, developmental and speech delay, and facial dysmorphisms (MIM #136140), which is phenotypically and molecularly distinct from disorders caused by truncating variants proximal or distal of that functional domain[44]. Moreover, in the *SCN2A* sodium channel, missense variants cause early infantile/childhood onset epilepsy that is responsive to sodium channel blockers, while protein truncating variants are associated with late onset epilepsy that does not respond to sodium channel blockers (MIM #613721), linking variant type to disease onset and drug response[45].

In addition to variant type, location and specific residue changes, effects on the phenotype could be modulated by general mutational load, sensitivity of the gene for variation, genetic predisposition, a second hit (by both rare or common variants), and/or additional genetic findings (Figure 1), further complicating the molecular picture and its links to the clinical outcome[36].

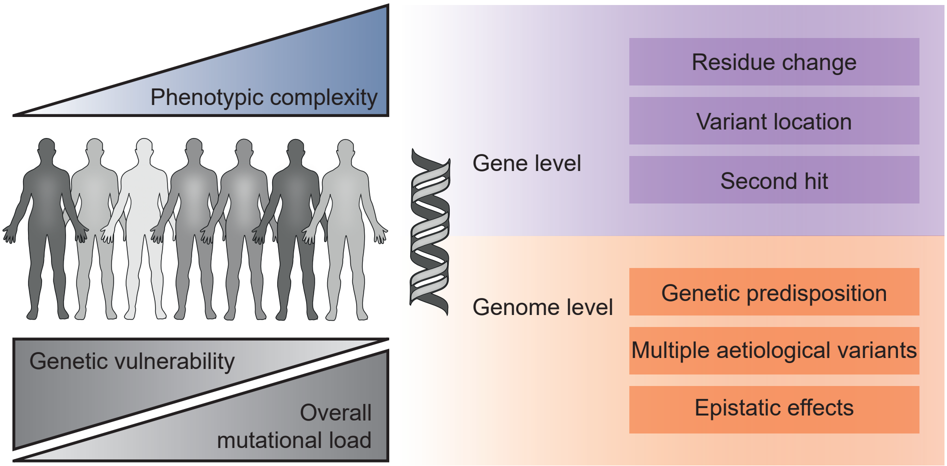


Figure 1. Genetic mechanisms underlying phenotypic complexity of developmental disorders. Partly adapted from Ref. 36

For the earlier mentioned genes initially linked to speech disorders (*BCL11A*, *CHD3*, *POU3F3* and *WDR5*), it has not been possible to identify genotype-phenotype correlations that could explain the identification of individuals with more distinct speech problems compared to cases with syndromic NDD. However, the described individuals have mainly been studied at the phenotypic level so far, and functional data that could be informative about the effects at the molecular level are still limited. Therefore, it remains unknown whether variants may have distinct genotype-specific effects on the molecular landscape, or whether they are modulated by other factors in the genetic background, which could potentially explain the clinical diversity.

In order to disentangle the complex relationships between specific variants and the wide array of associated phenotypes, we will need to develop more standardised and objective methods to register phenotypic data, and link that to information from functional assays that are able to screen for molecular effects of large sets of different variants in relevant model systems (an example of such an approach is described in Chapter 3 of this thesis).

The strengths of high-throughput functional assays

To aid variant interpretation in phenotype- and gene-driven NGS studies, and to better understand possible effects of aetiological variants, *in silico* tools can predict the impact of identified variants based on various criteria, including the evolutionary conservation of the affected residue, the difference of the specific residue change, and possible effects on protein structure (e.g. SIFT, PolyPhen, CADD-Phred)[46-48]. In addition, the intolerance for loss-of-function and missense variation of genes, based on genomic data of large control cohorts, can be summarised in an intolerance score[49], which can provide clues about the possible impact of a newly identified variant. Moreover, modelling the variants in a three-dimensional protein structure can identify hotspots of pathogenic variants, and in some cases may help predict effects on the functions of the protein[50]. However, without experimental validation, it remains challenging to label novel variants as pathogenic, in particular missense variants, and to draw genotype-phenotype correlations.

Immortalised cells to test for pathogenicity of genetic variants

Monogenic disorders that disrupt brain development are rare, such that new syndromes are first described using small cohorts of affected individuals carrying putative pathogenic variants in a gene of interest (with cohort sizes typically ranging between one to forty individuals). In such cohorts most identified variants can be functionally tested using cellular model systems that allow screening of multiple variants in parallel in a matter of weeks. Immortalised cells (over)expressing the variant of interest require little resources, grow indefinitely, and hence provide an accessible tool to examine the effects of variants on specific protein functions (Figure 2).

Immortalised cell lines are widely used in the field of disease modelling. A recent study on aetiological variants in the transcription factor POU3F3, first identified to carry a putative pathogenic variant in a case of severe speech disorder and later associated with a broader NDD including intellectual disability, autism-like features and speech impairments, used an immortalised cell line of embryonic origin (HEK293) to assess the effects of missense variants on its transcriptional activity[40]. All identified missense variants were experimentally tested by transfecting cells with expression constructs, and the variable effects on the transactivation capacity of POU3F3 were compared to the different phenotypic profiles of affected individuals. Another recent relevant example of a study combining clinical and functional data focused on *FOXP4*, a paralogue of *FOXP2*, and identified putative pathogenic variants associated with an NDD including developmental and language delays and congenital abnormalities. In that study, all identified missense variants were examined for their effects on transcriptional activity of FOXP4 in HEK293 cells, and the results allowed to distinguish between pathogenic and likely benign variants[51]. Such studies showcase how cell-based functional experiments can be used to provide additional evidence for pathogenicity of variants when describing novel syndromes.

Immortalised cells derived from patient samples, such as fibroblasts and lymphoblastoid cells could also be used to screen for functional effects of putative pathogenic variants on specific functions, although they are limited by the availability of the material. SETBP1-associated phenotypes provide an example of how patient-derived cell lines can be used to obtain a better understanding of variant-specific mechanisms. SETBP1 is a regulatory protein that is expressed during brain development and has been implicated in Schinzel-Giedion syndrome (MIM #269150), a severe disorder characterised by intellectual disability, multiple congenital malformations and premature death[52]. However, variants in the same gene can also cause SETBP1 haploinsufficiency disorder (MIM #616078), with symptoms that are milder than those observed in Schinzel-Giedion syndrome, and include intellectual disability and speech language difficulties[53]. While the first syndrome is associated with SETBP1 missense variants that cluster in a small region of the protein, suggesting a dominant-negative or gain-of-function effect, the latter is typically caused by loss-of-function variants[52; 53]. *In silico* tools can predict the pathogenicity of these missense variants, but are unable to provide evidence for mechanisms that are different from loss-of-function. Functional assays were carried out using lymphoblastoid cells derived from patient blood samples that were immortalised using Epstein-Barr virus transformation. The experimental data from these studies showed that missense variants in Schinzel-Giedion syndrome block SETBP1 protein degradation and thus impact the protein differently from loss-of-function variants, uncovering distinct disease mechanisms[52].

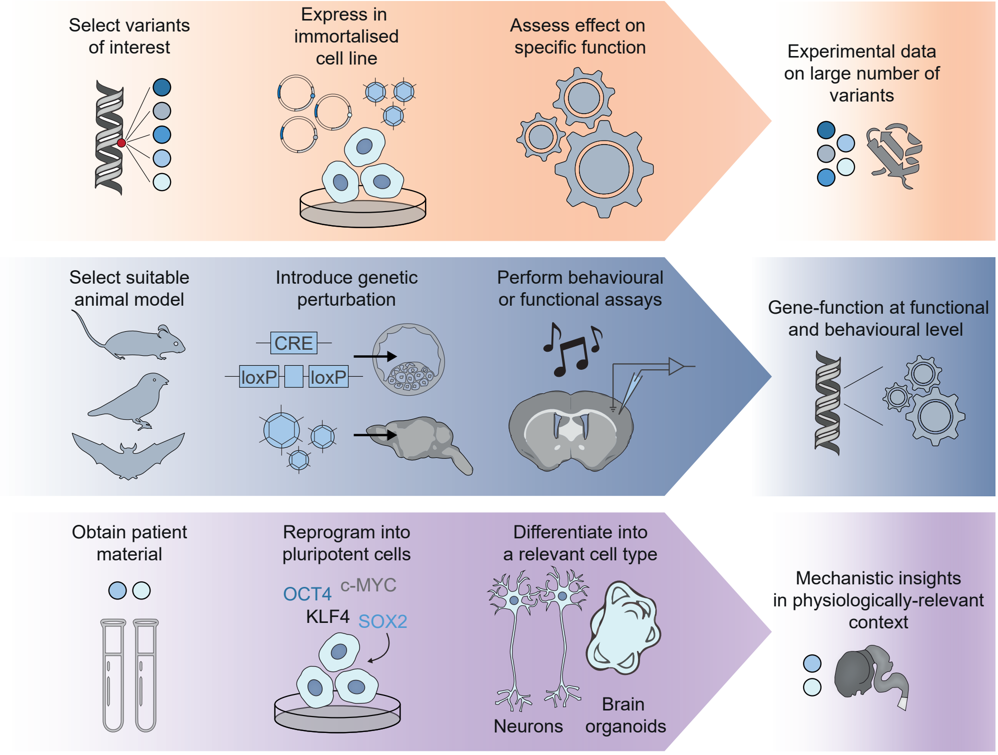

Select variants of interest
Express in immortalised cell line
Assess effect on specific function
Experimental data on large number of variants
Select suitable animal model
Introduce genetic perturbation
CRE
loxP
loxP
Perform behavioural or functional assays
Gene-function at functional and behavioural level
Obtain patient material
Reprogram into pluripotent cells
OCT4
c-MYC
KLF4
SOX2
Differentiate into a relevant cell type
Neurons
Brain organoids
Mechanistic insights in physiologically-relevant context

Figure 2. Functional model systems to study neurodevelopmental disorders.

High-throughput mutagenesis to screen large numbers of variants

In more recent years, methods have been developed that focus on scaling up functional screening to allow assessment of large numbers of variants, making it possible to test the effects of every possible amino acid change in a protein or peptide on a selected function simultaneously, generally called deep mutational scanning[54; 55]. Such studies rely on high-throughput mutagenesis, generating a library of DNA sequences that is then delivered to cells as expression plasmids using transfection or viral transduction[56], or expressed *in vitro* using bacteriophages[56; 57]. The cells expressing the proteins, or the proteins displayed on phages, are then sorted for a selectable phenotype (e.g. a post-translational modification, ligand binding, drug resistance, growth rate). The DNA library is sequenced before and after the selection step to infer the effect of each variant on the studied phenotype.

An example of a successful deep mutational scan is a study on *PTEN*[58], a tumour suppressor gene in which somatic mutations increase cancer susceptibility[59], while germline variants are associated with a spectrum of clinical phenotypes including autism spectrum disorder[60], macrocephaly[61] and tumour predisposition disorders[62-64]. The deep mutation scan study assessed the effect of 7,244 single amino acid variants (86% of all possible amino acid changes) on its lipid phosphatase enzyme activity in a yeast cell model[58]. Cells were selected based on growth rate, and the results were combined with information on protein structure to create a sequence-function map. These data provided insights into the genotype-phenotype correlations of symptoms associated with *PTEN*, identified regions of mutational tolerance and residues critical for PTEN function, and hence may eventually help to discriminate likely pathogenic variants from benign variants in the future.

Extrapolation of experimental data from such systematic high-throughput functional testing to clinical phenotypes will help increase understanding of phenotypic complexity, and therefore such assays could play an important role in the field of speech disorders and NDDs.

Physiologically-relevant cellular model systems

Cell systems based on immortalised cells or transformed patient-derived cells can provide functional read-outs within a time span of weeks under affordable conditions, without the need for complex mixtures of reagents and technical skills, and therefore allow for high-throughput functional screening experiments (Figure 2). However, such systems often lack physiological relevance.

In contrast, in animal models more complex mechanisms can be studied in the physiological context of brain development and function. As one of the most commonly used animal models, laboratory mice come with a rich genetic toolkit that offers possibilities of easy genetic manipulation and breeding. Hundreds of mouse strains have been well-documented and systematically phenotyped. The generation and crossing of mice with modular gene targeting alleles makes for a flexible system that allows the engineering of different types of conditional knockout models[65]. For example, to study the functions of *Foxp2* and its roles in behaviour, conditional knockouts were made that lacked expression of the gene in three different brain regions, the cortex, striatum and cerebellum, by crossing a *Foxp2-flox* strain with mouse lines that have a *Cre* gene under different promoters[66]. Other notable animal models have well-established genetic toolkits. For instance, in zebrafish, genes can be knocked down using morpholinos (DNA antisense oligomers)[67], while in fruit flies the gal4/UAS system is often used to overexpress or knockdown genes, or to rescue knockouts[68]. These animal model systems make it possible to study aspects of gene function in the context of development, the involvement in different cell types and cell functions, and the effects on behaviour (Figure 2). However, thorough characterisation of such models is usually labour- and time-intensive, and therefore screening for effects of multiple genetic variants is rarely done.

Moreover, human-specific processes in studies relying on animal models may be missed and the complex genetic background of patients cannot be taken into account[69].

Therefore, the field of functional genetics of NDDs also requires physiologically-relevant human-specific models that allow researchers to study the full genetic make-up of patients.

Induced pluripotent stem cells as a tool for disease modelling

Neural progenitors and neurons make up the most relevant cell types for studying NDDs. However, access to these cell types is extremely limited, as neurons are a post-mitotic cell type that cannot be kept and expanded in culture, and the pools of actively dividing neural stem cells in the adult brain are small and cannot be reached. Therefore, it is not possible to use primary patient-derived neuronal cells, and we need other methods to study human neurons in a laboratory setting.

Already in the 1980s, researchers discovered that cells from the inner cell mass of the blastocyst, a stage during early mammalian development, have the ability to grow indefinitely and maintain pluripotency, meaning that they can differentiate into cells of all three germ layers[70; 71]. These cells are called embryonic stem cells (ESCs). When ESCs were pushed to differentiate, by changing the culture conditions, cells with neural epithelial characteristics were identified[72], showing that ESCs could provide a cellular system to generate cells from the neuronal lineage *in vitro*. However, ESCs come with limitations. Ethical issues make it difficult to harvest and use human embryonic material, and although they could provide a tool to grow different cell types *in vitro*, it is not possible to obtain ESCs from patients.

To circumvent these issues, in 2007, researchers reported a method to generate cells with ESC-like pluripotent characteristics from somatic cells, called induced pluripotent stem cells (iPSCs). The development of this procedure, based on overexpressing a small set of transcription factors that normally have high expression in the inner cell mass of the blastocyst[73], revolutionised the field of disease-modelling: it was now possible to reprogramme somatic cells from patients into pluripotent cells that can then be differentiated into cell types relevant to the studied disease[69]. In subsequent years, protocols were optimised to generate iPSCs from easily accessible types of cells, such as skin fibroblasts, blood and urine[74-76]. Nowadays, iPSCs are the preferred cell type for studying human genetic diseases using stem cell-based models[69].

Growing neurons in a dish

When neuronal cell types are generated from ESCs and iPSCs using undirected differentiation protocols, the differentiated cells are part of a heterogeneous cell population including many non-neuronal cell types[72]. In order to study relevant disease mechanisms, homogeneous neuronal cell cultures are required.

In the 2000s-2010s, the first methods were described to generate neurons from pluripotent cells taking a directed differentiation approach. These protocols were based on insights from the field of human embryonic development, using small molecules and growth factors that could guide stem cells into the neuronal lineage[77-81]. Such methods pushed ESCs and iPSCs towards a neural progenitor identity with WNT-, BMP- and TGFβ-inhibitors[77; 80; 81], and later differentiated and matured the cells into neurons by withdrawal of the FGF2 growth factor[79]. The advantage of these directed differentiation models is that they follow the normal developmental trajectory.

Nowadays many variations of these protocols exist with different combinations of patterning molecules to direct cells into different neuronal identities, including ventral or dorsal forebrain and midbrain dopaminergic neurons[82; 83]. However, directed differentiation requires long-time culturing in order to obtain mature neurons (two weeks of neural induction, followed by 20-100 days of differentiation[79]), making it laborious and cost-intensive.

In order to speed up the process of generating *in vitro* neurons, methods were developed that skip the initial step from stem cell to neural progenitor cell, and immediately generate neurons from ESCs and iPSCs, by overexpressing a proneural master transcription factor delivered using viral transduction, such as *NGN2*[84] or *ASCL1*[85]. With these protocols, called direct reprogramming, homogeneous neuronal cell populations can be grown with mature characteristics more efficiently and in a shorter time (approximately two to three weeks)[84; 85].

Both directed differentiation and direct reprogramming of neurons have been successfully used to study how neuronal differentiation[86; 87], morphology[86; 88; 89], activity[88-90], and other cellular phenotypes[91; 92], are affected in cells derived from individuals with neurodevelopmental or neurodegenerative disorders, or in cells with genetic disruptions associated with disease, introduced using gene-editing tools.

Moving beyond the cellular monolayer

The developing human brain is a complex and dynamic structure, consisting of multiple types of neural progenitors and a large number of different neuronal cell types in a spatial organisation. The stem cell-based directed differentiation and direct reprogramming methods, described above, differentiate pluripotent cells into a homogeneous monolayer of neuronal cells. Therefore, while these models are good tools to study certain (isolated) aspects of neuronal development and function, they do not accurately recapitulate processes that involve multiple different cell types, or pathways that underlie the spatial organisation of the brain, such as cell migration and cell-cell interactions.

In 2013, it was demonstrated that non-adherent aggregates of human stem cells, grown in culture conditions without patterning growth factors, resulted in the development of three-dimensional neuroectodermal tissues[93]. The discovery led to a large number of studies that further optimised methods to grow three-dimensional stem cell-derived neuronal cultures[94-97]. These three-dimensional cell models are called brain organoids or spheroids, of which some rely on the default developmental differentiation programme of stem cells towards neuroectoderm (referred to as unpatterned)[93; 97], while others direct the cells into more specific areas of the brain using sets of signalling molecules (referred to as patterned)[94; 95]. Depending on the protocol used, these cultures contain multiple neuronal cell types[93-95] with regions that resemble the organisation of the human foetal brain, called ventricles[93; 94]. The brain organoid ventricles are organised around fluid-filled cavities, with neural progenitors at the apical side forming a ventricular zone, and post-mitotic neurons on the outside in a region similar to the cortical plate[93]. Brain organoids have transcriptomic profiles that overlap with early human foetal expression programmes (postconceptional weeks 12-13)[98], are able to reveal human-specific developmental trajectories[99] and show reproducible results when grown from different pluripotent cell lines[99; 100]. Moreover, studies focusing on neuronal defects in brain organoids grown from iPSCs derived from individuals with NDDs have shown that this model system is able to uncover developmental abberations[86; 101]. However, the existing methods to grow brain organoids are not able to mature the three-dimensional cultures to recapitulate later stages of brain development (> postconceptional week 16)[102], and brain organoids only contain broad cell classes of neural progenitors and neurons that in some cases intermix, both molecularly and spatially[103]. Moreover, in a recent study, organoids were found with increased activation of cellular stress pathways impairing cell-type specification[103], although others could not confirm such progressive stress effects in their culture system[104]. Overall, brain organoids represent a promising model to study variant-specific disease mechanisms in neurodevelopmental and speech disorders (Figure 2).

The advent of gene-editing tools

Alternative to heterologous expression of a variant of interest in a host cell line, or the use of patient material as a starting point for functional follow up studies, the genetic variant of interest can also be directly introduced into the genome of an appropriate cell model. Such a gene-editing approach makes it possible to study the effects of a variant in endogenous expression conditions, circumvents the need of patient material, and allows researchers to study different genetic variants in the same genetic background.

Early work in the 1990s and 2000s identified various nucleases that could be targeted to regions in the DNA where it would cleave the phosphodiester bonds of nucleotides, introducing a double-stranded break (homing endonucleases[105], zinc finger nucleases[106], TALENs[107]). Subsequently, these breaks are repaired by the cells default DNA repair mechanisms: 1) end-joining or 2) homology-directed repair[108]. End-joining is the most efficient repair mechanism, but often causes the addition of short nucleotide deletions or insertions, resulting in a frameshift variant, useful to generate knockout alleles. Homology-directed repair is less error-prone and allows the introduction of a specific variant of interest using a template, but only occurs in actively dividing cells. However, the specificity of these targeted nucleases for DNA sequences is encoded in the amino acid sequence[105-107], limiting their flexibility to guide them to a region of interest.

The discovery of a bacterial endonuclease, Cas9, that is guided to the DNA using a small RNA molecule[109], revolutionised the field of gene editing. The sole necessity of a short twenty-nucleotide target sequence allows for an easy and flexible but also highly specific system that is suitable for high-throughput experiments and multiplexing with libraries of guide-RNA molecules[110]. In the 2010s and 2020s, many adaptations of the CRISPR-Cas9 system have been developed, including an enzymatically inactive Cas9 protein (dCas9) that can be guided to a promoter area to block gene expression (CRISPRi)[111], a dCas9 protein fused to the transcriptional activator VP64 to increase gene expression (CRISPRa)[112], a dCas9 protein fused to single-stranded DNA deaminases to introduce specific nucleotide changes (base-editing)[108] and a dCas9 protein fused to a reverse transcriptase to introduce a specific variant from an RNA template (prime-editing)[108], among many others.

Combining the CRISPR-Cas gene-editing toolkit to introduce and/or repair genetic variants or modulate gene expression, with high-throughput functional analyses, or with sophisticated cell model systems such as brain organoids makes it possible to develop powerful study designs to uncover functional effects of putative pathogenic variants in speech disorders and NDDs.

Aims and relevance of this thesis

Currently, we do not understand how high-penetrant variants in the same gene can result in mild developmental phenotypes with prominent speech problems in some individuals, while causing broader neurodevelopmental syndromes characterised by intellectual disability/developmental delays, affected motor skills and/or autism, in others. We have not yet observed clear-cut genotype-phenotype correlations for genes initially identified in cases of speech disorder (e.g. see literature on *de novo* variants in *CHD3* (MIM #618205)[39] and *POU3F3* (MIM #618604)[40]). Nonetheless, it is plausible that differences in variant location, residue change, allele-specific events, and/or combinations with epistatic effects may explain differences in the molecular landscape (Figure 1), resulting in distinct mechanistic effects that could eventually impact the phenotype differently.

This thesis explores the potential of using functional cell-based assays to study complex human phenotypes, and discusses how such approaches could help to uncover in which ways different (types of) variants in a gene can cause different symptoms at the phenotypic level. Both high-throughput tests focusing on specific protein functions and physiologically-relevant functional assays modelling early brain development could map out this molecular space, to ultimately increase our understanding of the links between speech disorders and broader neurodevelopmental phenotypes (Figure 3).

Furthermore, disentangling the mechanistic differences of genetic variants is important to comprehensively understand the aetiology of the associated neurodevelopmental (speech) disorders. Such knowledge is not only critical in understanding the fundamental genetic complexity of dominant Mendelian phenotypes caused by high-penetrant variants, but also has direct applications in a clinical setting, improving genetic counselling and prenatal diagnostics, and providing insights in variant interpretation and disease prognosis.

Outline of the thesis

In **Chapter 2**, we provide a comprehensive overview of the status of the field on the genetics of Mendelian speech disorders caused by rare high-penetrant variants. We discuss how studies of rare variants in speech disorders often result in the identification of broader neurodevelopmental phenotypes, and argue that investigating such broader NDDs that share aspects of disrupted speech can still converge on common pathways that are affected. So far, genes linked to Mendelian forms of speech disorders are highly co-expressed during human foetal brain development and the large majority are involved in regulation of gene expression[20; 21]. These regulatory co-expression networks could subsequently be studied in both cell-based and physiologically-relevant model systems to learn more about neurobiological pathways important for human speech and language.

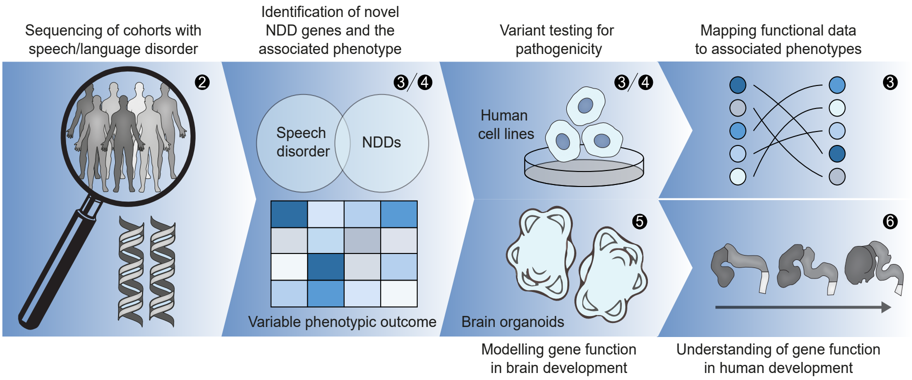
Chapter 3 shows the value of integrating cell-based functional screening with clinical information. We describe a cohort of individuals with rare heterozygous variants in SATB1, a protein that was previously identified to be part of the FOXP2-interactome[113]. Moreover, SATB1 is a homologue and interaction partner of SATB2, associated with a neurodevelopmental disorder characterised by severe speech problems (MIM #612313)[114; 115]. Using a combination of objectified in-depth clinical analysis and cell-based functional follow-up, we show that rare pathogenic variants in SATB1 have distinct consequences on protein function and lead to at least two different clinically distinguishable NDDs.

Although diagnostic strategies focusing on variants with full clinical penetrance are successful in identifying causal variants in individuals with NDDs, for a large number of cases no genetic explanation can be found. In **Chapter 4**, we combined objectified clinical analysis with functional assays in patient-derived cell lines to show variable expressivity for rare (likely) pathogenic inherited variants in *CHD3*. Previously, almost all described putative pathogenic *CHD3* variants were confirmed to have arisen *de novo* (55 out of 57)[39; 116]. The first *de novo CHD3* missense variant was identified in a child with childhood apraxia of speech[20], while a gene-driven follow-up study identified 35 individuals with *de novo CHD3* variants with broader NDD symptoms that included intellectual disability, macrocephaly, characteristic facial features and speech impairments[39]. In our study, we characterised twenty-one families with an affected NDD proband with an inherited *CHD3* missense or protein-truncating variant, transmitted from a healthy or only mildly affected carrier parent.

Evident from its links with speech apraxia and NDD, CHD3 seems critical for human brain development and/or function. However, little is known about the roles of CHD3 at the molecular level. In **Chapter 5** we further explore the functions of CHD3 during early stages of human brain development. After establishing and characterising a method to grow cerebral organoids to model brain development, we introduced heterozygous and homozygous loss-of-function variants in *CHD3* using CRISPR-Cas9, to knock-out the gene. Functional consequences were assessed with the latest transcriptomic approaches, including single-cell RNA sequencing.

Since links of speech disorders and NDDs to variants in genes like *CHD3* have only been recently established, investigations of their functional consequences are still at an early stage. **Chapter 6** further examines the potential for these types of studies to yield insights into aetiological pathways, by discussing FOXP2, a transcription factor that was first implicated in speech disorders more than two decades ago. This chapter describes what we have learned so far about the roles of FOXP2 in brain development and function integrating findings from clinical phenotype studies, research using animal models and work with human cell-based assays.

Chapter 7 summarises the studies of this thesis and the direct contributions to the field. Furthermore, it discusses how and which type of cell-based functional assays can provide critical insights in understanding mutation-specific mechanisms and the link to variability in complex human neurodevelopmental phenotypes.


Sequencing of cohorts with speech/language disorder
Identification of novel NDD genes and the associated phenotype
Variant testing for pathogenicity
Mapping functional data to associated phenotypes
❷
❸/❹
Speech disorder
NDDs
Variable phenotypic outcome
❸/❹
Human cell lines
Brain organoids
❺
❸
❻
Modelling gene function in brain development
Understanding of gene function in human development

❷ Chapter 2: Genetic pathways disrupted in human speech disorder
❸ Chapter 3: Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction
❹ Chapter 4: Inherited variants in *CHD3* demonstrate variable expressivity in Snijders Blok-Campeau syndrome
❺ Chapter 5: Using brain organoids to study the role of *CHD3* in early human brain development
❻ Chapter 6: Molecular networks of the FOXP2 transcription factor in the brain

Figure 3. Studying the genetics of neurodevelopmental and speech disorders. Studies to identify genetic factors underlying neurodevelopmental and speech disorders, disentangle variant-specific mechanisms, and map those back to phenotypic data to better understand phenotypic complexity. The areas covered by the different chapters of this thesis are indicated.

References

1. Kuhl, P.K. (2004). Early language acquisition: cracking the speech code Nature reviews Neuroscience 5, 831-843.
2. Mayor, J., and Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis. Developmental science 14, 769-785.
3. Dosman, C.F., Andrews, D., and Goulden, K.J. (2012). Evidence-based milestone ages as a framework for developmental surveillance. Paediatr Child Health 17, 561-568.
4. Petitto, L.A., and Marentette, P.F. (1991). Babbling in the manual mode: evidence for the ontogeny of language. Science (New York, NY) 251, 1493-1496.
5. Senghas, A., Kita, S., and Ozyürek, A. (2004). Children creating core properties of language: evidence from an emerging sign language in Nicaragua. Science (New York, NY) 305, 1779-1782.
6. Smith, K., and Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 3591-3603.
7. Van Agt, H., Verhoeven, L., Van Den Brink, G., and De Koning, H. (2011). The impact on socio-emotional development and quality of life of language impairment in 8-year-old children. Developmental medicine and child neurology 53, 81-88.
8. Fisher, S.E., and Marcus, G.F. (2006). The eloquent ape: genes, brains and the evolution of language. Nature reviews Genetics 7, 9-20.
9. Lattenkamp, E.Z., and Vernes, S.C. (2018). Vocal learning: a language-relevant trait in need of a broad cross-species approach. Current Opinion in Behavioral Sciences 21, 209-215.
10. Graham, S.A., Deriziotis, P., and Fisher, S.E. (2015). Insights into the Genetic Foundations of Human Communication. Neuropsychology Review 25, 3-26.
11. Bishop, D.V., North, T., and Donlan, C. (1995). Genetic basis of specific language impairment: evidence from a twin study. Developmental medicine and child neurology 37, 56-71.
12. Tomblin, J.B., and Buckwalter Paula, R. (1998). Heritability of Poor Language Achievement Among Twins. Journal of Speech, Language, and Hearing Research 41, 188-199.
13. Lewis, B.A., and Thompson, L.A. (1992). A study of developmental speech and language

disorders in twins. Journal of speech and hearing research 35, 1086-1094.
14. Bishop, D.V.M., and Hayiou-Thomas, M.E. (2008). Heritability of specific language impairment depends on diagnostic criteria. Genes, brain, and behavior 7, 365-372.
15. Bishop, D.V., Snowling, M.J., Thompson, P.A., and Greenhalgh, T. (2016). CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. PloS one 11, e0158753.
16. Stromswold, K. (1998). Genetics of spoken language disorders. Human biology 70, 297-324.
17. Gialluisi, A., Newbury, D.F., Wilcutt, E.G., Olson, R.K., DeFries, J.C., Brandler, W.M., Pennington, B.F., Smith, S.D., Scerri, T.S., Simpson, N.H., et al. (2014). Genome-wide screening for DNA variants associated with reading and language traits. Genes, brain, and behavior 13, 686-701.
18. Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 519-523.
19. Reuter, M.S., Riess, A., Moog, U., Briggs, T.A., Chandler, K.E., Rauch, A., Stampfer, M., Steindl, K., Gläser, D., and Joset, P. (2017). FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. Journal of medical genetics 54, 64-72.
20. Eising, E., Carrion-Castillo, A., Vino, A., Strand, E.A., Jakielski, K.J., Scerri, T.S., Hildebrand, M.S., Webster, R., Ma, A., Mazoyer, B., et al. (2019). A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. Molecular psychiatry 24, 1065-1078.
21. Hildebrand, M.S., Jackson, V.E., Scerri, T.S., Van Reyk, O., Coleman, M., Braden, R.O., Turner, S., Rigbye, K.A., Boys, A., Barton, S., et al. (2020). Severe childhood speech disorder: Gene discovery highlights transcriptional dysregulation. Neurology 94, e2148-e2167.
22. Fisher, S.E., Vargha-Khadem, F., Watkins, K.E., Monaco, A.P., and Pembrey, M.E. (1998). Localisation of a gene implicated in a severe speech and language disorder. Nature genetics 18, 168-170.
23. Lai, C.S.L., Fisher, S.E., Hurst, J.A., Levy, E.R., Hodgson, S., Fox, M., Jeremiah, S., Povey, S., Jamison, D.C., Green, E.D., et al. (2000). The SPCH1 Region on Human 7q31: Genomic Characterization of the Critical Interval and Localization of Translocations Associated with Speech and Language Disorder. The American Journal of Human Genetics 67, 357-368.
24. Co, M., Anderson, A.G., and Konopka, G. (2020). FOXP transcription factors in vertebrate brain development, function, and disorders. Wiley Interdisciplinary Reviews: Developmental Biology 9, e375.
25. Vernes, S.C., Nicod, J., Elahi, F.M., Coventry, J.A., Kenny, N., Coupe, A.M., Bird, L.E., Davies, K.E., and Fisher, S.E. (2006). Functional genetic analysis of mutations implicated in a human speech and language disorder. Human molecular genetics 15, 3154-3167.
26. Morgan, A., Fisher, S.E., Scheffer, I., and Hildebrand, M. (2016 [updated 2017 Feb 2]). FOXP2-Related Speech and Language Disorders. In GeneReviews, M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, andA. Amemiya, eds. (Seattle (WA)).
27. de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. New England Journal of Medicine 367, 1921-1929.
28. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433-438.
29. Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nature genetics 42, 1109-1112.
30. Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A.F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T.W., et al. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nature genetics 47, 1363-1369.
31. Martin, H.C., Jones, W.D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson, J.D., Jones, C.P., Handsaker, J., Gallone, G., Bruntraeger, M., et al. (2018). Quantifying the contribution of recessive coding variation to developmental disorders. Science (New York, NY) 362, 1161-1164.
32. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature 586, 757-762.
33. Allanson, J.E., and Roberts, A.E. (2001 - [updated 2019 Aug 8]). Noonan Syndrome. In

GeneReviews, M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (University of Washington, Seattle (WA)).

34. Skjeldal, O., and Henriksen, M.W. (2020). Rett syndrome: A more Heterogeneous group than previously thought? (907). Neurology 94, 907.

35. Chen, X.S., Reader, R.H., Hoischen, A., Veltman, J.A., Simpson, N.H., Francks, C., Newbury, D.F., and Fisher, S.E. (2017). Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. Scientific reports 7, 46105.

36. Parenti, I., Rabaneda, L.G., Schoen, H., and Novarino, G. (2020). Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends in Neurosciences 43, 608-621.

37. Peter, B., Matsushita, M., Oda, K., and Raskind, W. (2014). De novo microdeletion of BCL11A is associated with severe speech sound disorder. American journal of medical genetics Part A 164, 2091-2096.

38. Dias, C., Estruch, S.B., Graham, S.A., McRae, J., Sawiak, S.J., Hurst, J.A., Joss, S.K., Holder, S.E., Morton, J.E., and Turner, C. (2016). BCL11A haploinsufficiency causes an intellectual disability syndrome and dysregulates transcription. The American Journal of Human Genetics 99, 253-274.

39. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nature communications 9, 4619.

40. Snijders Blok, L., Kleefstra, T., Venselaar, H., Maas, S., Kroes, H.Y., Lachmeijer, A.M.A., van Gassen, K.L.I., Firth, H.V., Tomkins, S., Bodek, S., et al. (2019). De Novo Variants Disturbing the Transactivation Capacity of POU3F3 Cause a Characteristic Neurodevelopmental Disorder. The American Journal of Human Genetics 105, 403-412.

41. Snijders Blok, L. (2021). Let the genes speak! De novo variants in developmental disorders with speech and language impairment.

42. Myers, S.M., Challman, T.D., Bernier, R., Bourgeron, T., Chung, W.K., Constantino, J.N., Eichler, E.E., Jacquemont, S., Miller, D.T., Mitchell, K.J., et al. (2020). Insufficient Evidence for "Autism-Specific" Genes. American Journal of Human Genetics 106, 587-595.

43. Haijes, H.A., Koster, M.J.E., Rehmann, H., Li, D., Hakonarson, H., Cappuccio, G., Hancarova, M., Lehalle, D., Reardon, W., Schaefer, G.B., et al. (2019). De Novo Heterozygous POLR2A Variants Cause a Neurodevelopmental Syndrome with Profound Infantile-Onset Hypotonia. The American Journal of Human Genetics 105, 283-301.

44. Rots, D., Chater-Diehl, E., Dingemans, A.J.M., Goodman, S.J., Siu, M.T., Cytrynbaum, C., Choufani, S., Hoang, N., Walker, S., Awamleh, Z., et al. (2021). Truncating SRCAP variants outside the Floating-Harbor syndrome locus cause a distinct neurodevelopmental disorder with a specific DNA methylation signature. The American Journal of Human Genetics 108, 1053-1068.

45. Wolff, M., Johannesen, K.M., Hedrich, U.B.S., Masnada, S., Rubboli, G., Gardella, E., Lesca, G., Ville, D., Milh, M., Villard, L., et al. (2017). Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. Brain : a journal of neurology 140, 1316-1336.

46. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature methods 7, 248-249.

47. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nature Protocols 11, 1-9.

48. Rentzsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine 13, 31.

49. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434-443.

50. Lelieveld, S.H., Wiel, L., Venselaar, H., Pfundt, R., Vriend, G., Veltman, J.A., Brunner, H.G., Vissers, L., and Gilissen, C. (2017). Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. American journal of human genetics 101, 478-484.

51. Snijders Blok, L., Vino, A., den Hoed, J., Underhill, H.R., Monteil, D., Li, H., Reynoso Santos, F.J., Chung, W.K., Amaral, M.D., Schnur, R.E., et al. (2021). Heterozygous variants that disturb the transcriptional repressor activity of FOXP4 cause a developmental disorder with speech/

language delays and multiple congenital abnormalities. Genetics in medicine : official journal of the American College of Medical Genetics 23, 534-542.

52. Acuna-Hidalgo, R., Deriziotis, P., Steehouwer, M., Gilissen, C., Graham, S.A., van Dam, S., Hoover-Fong, J., Telegrafi, A.B., Destree, A., Smigiel, R., et al. (2017). Overlapping SETBP1 gain-of-function mutations in Schinzel-Giedion syndrome and hematologic malignancies. PLoS Genet 13, e1006683-e1006683.
53. Jansen, N.A., Braden, R.O., Srivastava, S., Otness, E.F., Lesca, G., Rossi, M., Nizon, M., Bernier, R.A., Quelin, C., van Haeringen, A., et al. (2021). Clinical delineation of SETBP1 haploinsufficiency disorder. European journal of human genetics : EJHG 29, 1198-1205.
54. Araya, C.L., and Fowler, D.M. (2011). Deep mutational scanning: assessing protein function on a massive scale. Trends in biotechnology 29, 435-442.
55. Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nature methods 7, 741-746.
56. Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature methods 11, 801-807.
57. Starita, L.M., Pruneda, J.N., Lo, R.S., Fowler, D.M., Kim, H.J., Hiatt, J.B., Shendure, J., Brzovic, P.S., Fields, S., and Klevit, R.E. (2013). Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proceedings of the National Academy of Sciences of the United States of America 110, E1263-1272.
58. Mighell, T.L., Evans-Dutson, S., and O'Roak, B.J. (2018). A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. The American Journal of Human Genetics 102, 943-955.
59. Alimonti, A., Carracedo, A., Clohessy, J.G., Trotman, L.C., Nardella, C., Egia, A., Salmena, L., Sampieri, K., Haveman, W.J., and Brogi, E. (2010). Subtle variations in Pten dose determine cancer susceptibility. Nature genetics 42, 454-458.
60. Varga, E.A., Pastore, M., Prior, T., Herman, G.E., and McBride, K.L. (2009). The prevalence of PTEN mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly. Genetics in Medicine 11, 111-117.
61. Mester, J.L., Tilot, A.K., Rybicki, L.A., Frazier, T.W., and Eng, C. (2011). Analysis of prevalence and degree of macrocephaly in patients with germline PTEN mutations and of brain weight in Pten knock-in murine model. European Journal of Human Genetics 19, 763-768.
62. Liaw, D., Marsh, D.J., Li, J., Dahia, P.L., Wang, S.I., Zheng, Z., Bose, S., Call, K.M., Tsou, H.C., and Peacoke, M. (1997). Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nature genetics 16, 64-67.
63. Marsh, D.J., Dahia, P.L., Zheng, Z., Liaw, D., Parsons, R., Gorlin, R.J., and Eng, C. (1997). Germline mutations in PTEN are present in Bannayan-Zonana syndrome. Nature genetics 16, 333-334.
64. Padberg, G.W., Schot, J.D., Vielvoye, G.J., Bots, G.T.A., and De Beer, F.C. (1991). Lhermitte-Duclos disease and Cowden disease: a single phakomatosis. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 29, 517-523.
65. van der Weyden, L., White, J.K., Adams, D.J., and Logan, D.W. (2011). The mouse genetics toolkit: revealing function and mechanism. Genome Biology 12, 224.
66. French, C.A., Vinueza Veloz, M.F., Zhou, K., Peter, S., Fisher, S.E., Costa, R.M., and De Zeeuw, C.I. (2019). Differential effects of Foxp2 disruption in distinct motor circuits. Molecular Psychiatry 24, 447-462.
67. Eisen, J.S., and Smith, J.C. (2008). Controlling morpholino experiments: don't stop making antisense. Development 135, 1735-1743.
68. Southall, T.D., Elliott, D.A., and Brand, A.H. (2008). The GAL4 System: A Versatile Toolkit for Gene Expression in Drosophila. CSH protocols 2008, pdb.top49.
69. Shi, Y., Inoue, H., Wu, J.C., and Yamanaka, S. (2017). Induced pluripotent stem cell technology: a decade of progress. Nature Reviews Drug Discovery 16, 115-130.
70. Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature 292, 154-156.
71. Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proceedings of the National Academy of Sciences 78, 7634-7638.
72. Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science (New York, NY) 282, 1145-1147.

73. Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663-676.
74. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131, 861-872.
75. Zhou, T., Benda, C., Duzinger, S., Huang, Y., Li, X., Li, Y., Guo, X., Cao, G., Chen, S., Hao, L., et al. (2011). Generation of induced pluripotent stem cells from urine. J Am Soc Nephrol 22, 1221-1228.
76. Staerk, J., Dawlaty, M.M., Gao, Q., Maetzel, D., Hanna, J., Sommer, C.A., Mostoslavsky, G., and Jaenisch, R. (2010). Reprogramming of Human Peripheral Blood Cells to Induced Pluripotent Stem Cells. Cell Stem Cell 7, 20-24.
77. Watanabe, K., Kamiya, D., Nishiyama, A., Katayama, T., Nozaki, S., Kawasaki, H., Watanabe, Y., Mizuseki, K., and Sasai, Y. (2005). Directed differentiation of telencephalic precursors from embryonic stem cells. Nature Neuroscience 8, 288-296.
78. Eiraku, M., Watanabe, K., Matsuo-Takasaki, M., Kawada, M., Yonemura, S., Matsumura, M., Wataya, T., Nishiyama, A., Muguruma, K., and Sasai, Y. (2008). Self-Organized Formation of Polarized Cortical Tissues from ESCs and Its Active Manipulation by Extrinsic Signals. Cell Stem Cell 3, 519-532.
79. Shi, Y., Kirwan, P., and Livesey, F.J. (2012). Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. Nature Protocols 7, 1836-1846.
80. Shi, Y., Kirwan, P., Smith, J., Robinson, H.P., and Livesey, F.J. (2012). Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. Nat Neurosci 15, 477-486, s471.
81. Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nature biotechnology 27, 275-280.
82. Li, X.-J., Zhang, X., Johnson, M.A., Wang, Z.-B., LaVaute, T., and Zhang, S.-C. (2009). Coordination of sonic hedgehog and Wnt signaling determines ventral and dorsal telencephalic neuron types from human embryonic stem cells. Development 136, 4055-4063.
83. Kriks, S., Shim, J.-W., Piao, J., Ganat, Y.M., Wakeman, D.R., Xie, Z., Carrillo-Reid, L., Auyeung, G., Antonacci, C., Buch, A., et al. (2011). Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson’s disease. Nature 480, 547-551.
84. Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., et al. (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron 78, 785-798.
85. Chanda, S., Ang, C.E., Davila, J., Pak, C., Mall, M., Lee, Q.Y., Ahlenius, H., Jung, S.W., Südhof, T.C., and Wernig, M. (2014). Generation of induced neuronal cells by the single reprogramming factor ASCL1. Stem cell reports 3, 282-296.
86. Schafer, S.T., Paquola, A.C., Stern, S., Gosselin, D., Ku, M., Pena, M., Kuret, T.J., Liyanage, M., Mansour, A.A., and Jaeger, B.N. (2019). Pathological priming causes developmental gene network heterochronicity in autistic subject-derived neurons. Nature neuroscience 22, 243-255.
87. Zhang, K., Yu, F., Zhu, J., Han, S., Chen, J., Wu, X., Chen, Y., Shen, T., Liao, J., Guo, W., et al. (2020). Imbalance of Excitatory/Inhibitory Neuron Differentiation in Neurodevelopmental Disorders with an NR2F1 Point Mutation. Cell Reports 31, 107521.
88. Frega, M., Linda, K., Keller, J.M., Gümüş-Akay, G., Mossink, B., van Rhijn, J.-R., Negwer, M., Klein Gunnewiek, T., Foreman, K., Kompier, N., et al. (2019). Neuronal network dysfunction in a model for Kleefstra syndrome mediated by enhanced NMDAR signaling. Nature communications 10, 4928.
89. Klein Gunnewiek, T.M., Van Hugte, E.J.H., Frega, M., Guardia, G.S., Foreman, K., Panneman, D., Mossink, B., Linda, K., Keller, J.M., Schubert, D., et al. (2020). m.3243A > G-Induced Mitochondrial Dysfunction Impairs Human Neuronal Development and Reduces Neuronal Network Activity and Synchronicity. Cell Reports 31, 107538.
90. Mossink, B., Verboven, A.H.A., van Hugte, E.J.H., Klein Gunnewiek, T.M., Parodi, G., Linda, K., Schoenmaker, C., Kleefstra, T., Kozicz, T., van Bokhoven, H., et al. (2021). Human neuronal networks on micro-electrode arrays are a highly robust tool to study disease-specific genotype-phenotype correlations in vitro. Stem cell reports 16, 2182-2196.
91. Chung Chee, Y., Khurana, V., Auluck Pavan, K., Tardiff Daniel, F., Mazzulli Joseph, R., Soldner, F., Baru, V., Lou, Y., Freyzon, Y., Cho, S., et al. (2013). Identification and Rescue of α-Synuclein Toxicity in Parkinson Patient–Derived Neurons. Science (New York, NY) 342, 983-987.
92. Bell, S., McCarty, V., Peng, H., Jefri, M., Hettige, N., Antonyan, L., Crapper, L., O'Leary, L.A.,

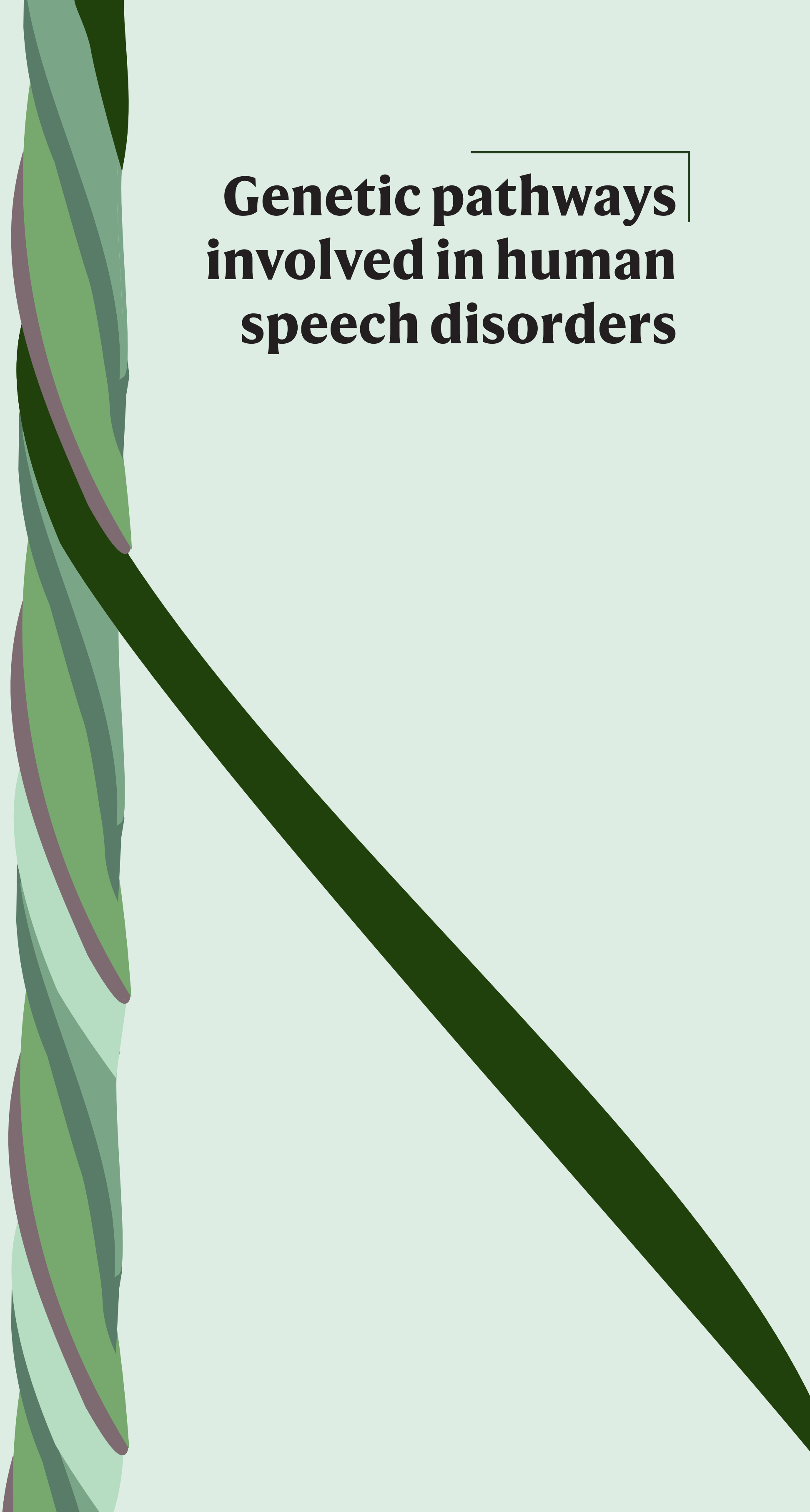
Zhang, X., Zhang, Y., et al. (2021). Lesch-Nyhan disease causes impaired energy metabolism and reduced developmental potential in midbrain dopaminergic cells. Stem cell reports 16, 1749-1762.

93. Lancaster, M.A., Renner, M., Martin, C.A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature 501, 373-379.
94. Qian, X., Nguyen, H.N., Song, M.M., Hadiono, C., Ogden, S.C., Hammack, C., Yao, B., Hamersky, G.R., Jacob, F., Zhong, C., et al. (2016). Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. Cell 165, 1238-1254.
95. Birey, F., Andersen, J., Makinson, C.D., Islam, S., Wei, W., Huber, N., Fan, H.C., Metzler, K.R.C., Panagiotakos, G., Thom, N., et al. (2017). Assembly of functionally integrated human forebrain spheroids. Nature 545, 54-59.
96. Sloan, S.A., Darmanis, S., Huber, N., Khan, T.A., Birey, F., Caneda, C., Reimer, R., Quake, S.R., Barres, B.A., and Paşca, S.P. (2017). Human Astrocyte Maturation Captured in 3D Cerebral Cortical Spheroids Derived from Pluripotent Stem Cells. Neuron 95, 779-790.e776.
97. Lancaster, M.A., Corsini, N.S., Wolfinger, S., Gustafson, E.H., Phillips, A.W., Burkard, T.R., Otani, T., Livesey, F.J., and Knoblich, J.A. (2017). Guided self-organization and cortical plate formation in human brain organoids. Nature biotechnology 35, 659-666.
98. Camp, J.G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M., et al. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proceedings of the National Academy of Sciences 112, 15672.
99. Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., et al. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature 574, 418-422.
100. Velasco, S., Kedaigle, A.J., Simmons, S.K., Nash, A., Rocha, M., Quadrato, G., Paulsen, B., Nguyen, L., Adiconis, X., Regev, A., et al. (2019). Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. Nature 570, 523-527.
101. Khan, T.A., Revah, O., Gordon, A., Yoon, S.-J., Krawisz, A.K., Goold, C., Sun, Y., Kim, C.H., Tian, Y., Li, M.-Y., et al. (2020). Neuronal defects in a human cellular model of 22q11.2 deletion syndrome. Nature Medicine 26, 1888-1898.
102. Amiri, A., Coppola, G., Scuderi, S., Wu, F., Roychowdhury, T., Liu, F., Pochareddy, S., Shin, Y., Safi, A., Song, L., et al. (2018). Transcriptome and epigenome landscape of human cortical development modeled in organoids. Science 362, eaat6720.
103. Bhaduri, A., Andrews, M.G., Mancia Leon, W., Jung, D., Shin, D., Allen, D., Jung, D., Schmunk, G., Haeussler, M., Salma, J., et al. (2020). Cell stress in cortical organoids impairs molecular subtype specification. Nature 578, 142-148.
104. Gordon, A., Yoon, S.-J., Tran, S.S., Makinson, C.D., Park, J.Y., Andersen, J., Valencia, A.M., Horvath, S., Xiao, X., Huguenard, J.R., et al. (2021). Long-term maturation of human cortical organoids matches key early postnatal transitions. Nature Neuroscience 24, 331-342.
105. Stoddard, B.L. (2011). Homing Endonucleases: From Microbial Genetic Invaders to Reagents for Targeted DNA Modification. Structure 19, 7-15.
106. Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory, P.D. (2010). Genome editing with engineered zinc finger nucleases. Nature Reviews Genetics 11, 636-646.
107. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. Science 326, 1509-1512.
108. Anzalone, A.V., Koblan, L.W., and Liu, D.R. (2020). Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. Nature Biotechnology 38, 824-844.
109. Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468, 67-71.
110. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nature Protocols 8, 2281-2308.
111. Larson, M.H., Gilbert, L.A., Wang, X., Lim, W.A., Weissman, J.S., and Qi, L.S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nature Protocols 8, 2180-2196.
112. Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517, 583-588.

113. Estruch, S.B., Graham, S.A., Quevedo, M., Vino, A., Dekkers, D.H.W., Deriziotis, P., Sollis, E., Demmers, J., Poot, R.A., and Fisher, S.E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Human molecular genetics 27, 1212-1227.
114. Bengani, H., Handley, M., Alvi, M., Ibitoye, R., Lees, M., Lynch, S.A., Lam, W., Fannemel, M., Nordgren, A., Malmgren, H., et al. (2017). Clinical and molecular consequences of disease-associated de novo mutations in SATB2. Genetics in medicine : official journal of the American College of Medical Genetics 19, 900-908.
115. Snijders Blok, L., Goosen, Y.M., van Haaften, L., van Hulst, K., Fisher, S.E., Brunner, H.G., Egger, J.I.M., and Kleefstra, T. (2021). Speech-language profiles in the context of cognitive and adaptive functioning in SATB2-associated syndrome. Genes, Brain and Behavior n/a, e12761.
116. Drivas, T.G., Li, D., Nair, D., Alaimo, J.T., Alders, M., Altmüller, J., Barakat, T.S., Bebin, E.M., Bertsch, N.L., Blackburn, P.R., et al. (2020). A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome. European journal of human genetics: EJHG 28, 1422-1431.

2

This chapter has been published as:
Den Hoed, J., & Fisher, S. E. (2020). Genetic pathways involved in human speech disorders. Current Opinion in Genetics & Development, 65, 103-111.



Genetic pathways involved in human speech disorders

Abstract

Rare genetic variants that disrupt speech development provide entry points for deciphering the neurobiological foundations of key human capacities. The value of this approach is illustrated by *FOXP2*, a transcription factor gene that was implicated in speech apraxia, and subsequently investigated using human cell-based systems and animal models. Advances in next-generation sequencing, coupled to *de novo* paradigms, facilitated discovery of aetiological variants in additional genes in speech disorder cohorts. As for other neurodevelopmental syndromes, gene-driven studies show blurring of boundaries between diagnostic categories, with some risk genes shared across speech disorders, intellectual disability and autism. Convergent evidence hints at involvement of regulatory genes co-expressed in early human brain development, suggesting that aetiological pathways could be amenable for investigation in emerging neural models such as cerebral organoids.

Introduction

Following decades of speculation over genetic contributions to distinctive human communication skills, advances in molecular methods enabled scientists to begin identifying critical genomic factors[1]. Much research so far focused on linkage mapping and association screening of developmental speech and language impairments, revealing that while such disorders have a complex genetic architecture, a significant subset of cases involve rare high-penetrance variants disrupting single genes[2]. Here, we discuss the importance of rare variants as entry points for studying neurobiological pathways, describe how next-generation sequencing and gene-driven studies are transforming this field, and argue that emerging cell-based models of human brain development will be crucial for a fuller understanding of how gene disruptions yield speech disorders.

Molecular perspectives on speech - the example of *FOXP2*

FOXP2 was the first gene for which rare variants could be implicated in a monogenic speech disorder (primarily characterised by childhood apraxia of speech; CAS; Table 1). Since the initial report describing a causative point mutation in a multigenerational family, as well as a translocation disturbing the gene in an independent case[3], different genetic disruptions of *FOXP2* have been identified in multiple cases of speech/language disorder, both inherited and *de novo*[4; 5]. The discovery of *FOXP2* led to an array of studies of its functions in the brain (Figure 1)[2; 5].

FOXP2 encodes a transcription factor with a high degree of evolutionary conservation (both for protein sequences and neural expression patterns), facilitating functional analyses in animal models[6]. Conditional knockout and targeted knockdown/overexpression strategies in mice and birds are being used to dissect roles of FoxP2 in different parts of the brain (Figure 1). Studies of mouse models build on a well-established genetic toolkit, as well as rich literature on brain development, and can therefore teach us about gene function for conserved molecular mechanisms and behaviours. Mice are known to produce sequences of ultrasonic vocalisations, but their abilities to learn these appear limited, and the relevance of such behaviours for gaining insights into biology of human speech is much debated[7]. In contrast, although birds are more distantly related to humans than are mice, some species of songbird have sophisticated skills for auditory-guided vocal learning, which involves integration of auditory processing and motor learning, showing parallels to processes underlying speech. Moreover, there is evidence that birdsong and speech are coded in somewhat analogous brain circuitries[8].

Recent work in murine and avian models has largely (though not exclusively) centred on neuronal subpopulations of the cortex, striatum and cerebellum, three key sites where the gene is expressed[9], which have been independently highlighted by neuroimaging of humans with *FOXP2*-related speech disorder[10; 11]. Although it is an established marker of deep cortical layers, selective *Foxp2* deletion from the developing mouse cortex does not disturb lamination[12; 13]. Even so, mice lacking cortical *Foxp2* show abnormalities in tests of social behaviour and cognitive flexibility[13; 14]. Single-cell transcriptomics in cortical-specific mouse knockouts suggests that the gene

Table 1. Brief description of the main neurodevelopmental disorders mentioned in this review.

Disorder	**Description**
Childhood apraxia of speech (CAS)	Developmental deficits in speech motor planning and programming. Diagnostic symptoms include inconsistent speech errors, difficulties in speech sequencing that worsen with increased complexity of the utterance, and disrupted rhythm and intonation. Also known as developmental verbal dyspraxia (DVD).
Stuttering	Speech fluency disorder that involves interruptions in the flow of speaking, characterised by involuntary repetitions (of individual sounds, syllables, words, or phrases), sound prolongations, blocks, interjections, and revisions.
Developmental language disorder (DLD)	Delayed or impaired acquisition and use of language in the absence of a clear biomedical cause, with a poor prognosis and interfering with daily life (according to CATALISE-2 definition from 2017[33]). Prior to CATALISE-2 study, other terms were commonly used to classify these kinds of problems, in particular Specific Language Impairment (SLI).
Intellectual disability (ID)	Heterogeneous group of disorders involving general cognitive impairments that significantly affect both intellectual (learning, problem solving, judgement) and adaptive functioning (communication, independent living).
Autism spectrum disorder (ASD)	Range of developmental conditions characterised by impaired skills for communication/interaction with others, and restricted interests and repetitive behaviours, impacting on the ability to function in every-day life contexts (school, work etc.)

contributes to development and function of dopamine-receptor expressing neurons[13].

Within the rodent striatum, *Foxp2* is predominantly expressed in D1-receptor-positive medium spiny neurons; studies of global heterozygous knockout mice revealed effects on inhibitory presynaptic strength of these cells, implicating the gene in excitation/inhibition balance of pathways underlying motor-skill learning[15]. Striatal-specific *Foxp2* knockouts show increased variability in skilled motor behaviours, assessed via operant lever-pressing tasks[9]. Viral-based manipulations (knockdown versus overexpression) of this brain region in adult mice demonstrate post-developmental roles of *Foxp2* in regulating corticostriatal synapse functions and associated behaviours[16]. Moreover, knockdown/overexpression experiments targeting Area X (a striatal nucleus involved in vocal production learning of male zebra finches) underline the importance of this gene for learning of song by juvenile birds[17], and its maintenance in adulthood[18]. Regarding cerebellar functions, mice with Purkinje-cell specific knockouts of *Foxp2* display slower sequencing in lever-pressing tasks , and reduced performance on tests of skilled locomotion. *In vivo* electrophysiology indicates that *Foxp2*-deficient Purkinje cells have increased intrinsic excitability, and show abnormal firing properties during limb movement[9].

According to the latest human cell-based studies (Figure 1), FOXP2 is part of a broader interacting network of brain-expressed transcription factors[19], promoting pathways for neuronal maturation via chromosomal remodelling, while repressing genes that would maintain a neural progenitor state[20]. Of the molecules known to be regulated by and/or interact with FOXP2, many are themselves associated with brain-related disorders[19; 20]. Therefore, the FOXP2 interactome could provide useful inroads for defining and characterising neurobiological pathways involved in speech development. An example is the close paralogue FOXP1, which is co-expressed with FOXP2 in a subset of brain structures, where the transcription factors can heterodimerise to potentially co-regulate targets. Rare variants disrupting human FOXP1 cause a phenotype that is broader and more severe than *FOXP2*-related disorder, including features of autism and/or intellectual disability (ID)[21]. Human cell-based analyses of an aetiological missense variant in the DNA-binding domain of *FOXP1*, equivalent to the most studied mutation of *FOXP2*, showed comparable functional effects, suggesting that it is the differences in neural expression patterns of the two paralogues that account for distinctive phenotypes of the associated disorders[22].

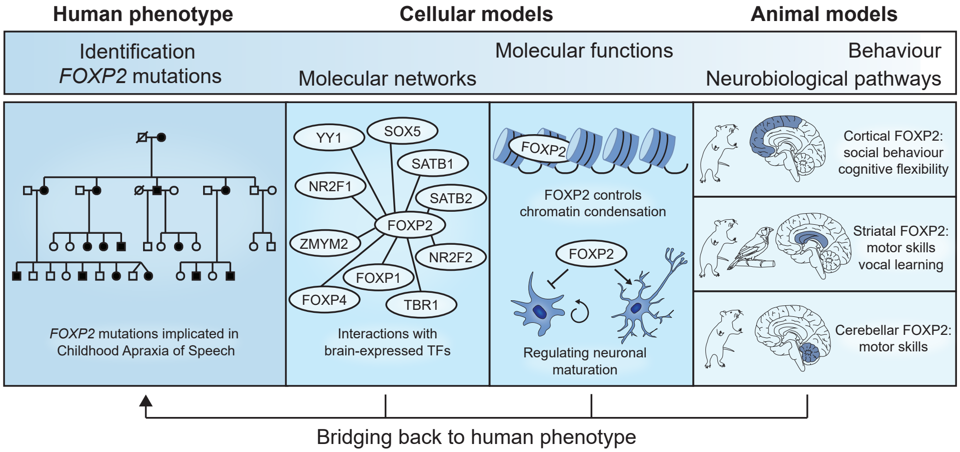

Human phenotype
Cellular models
Animal models
Identification
FOXP2 mutations
Molecular networks
Molecular functions
Behaviour
Neurobiological pathways
FOXP2 mutations implicated in Childhood Apraxia of Speech
YY1
SOX5
SATB1
NR2F1
SATB2
FOXP2
ZMYM2
NR2F2
FOXP1
FOXP4
TBR1
Interactions with brain-expressed TFs
FOXP2
FOXP2 controls chromatin condensation
FOXP2
Regulating neuronal maturation
Cortical FOXP2: social behaviour cognitive flexibility
Striatal FOXP2: motor skills vocal learning
Cerebellar FOXP2: motor skills
Bridging back to human phenotype

Figure 1. Multiple approaches for studying the role of *FOXP2* in the brain.

Taken together, these molecular studies uncover distinct roles for *FOXP2* in different brain regions that implicate the gene in development and function of cortico-striatal and -cerebellar circuitries[9-20], converging with identification of subtle cortical, striatal and cerebellar abnormalities in patients with *FOXP2* disruptions[10; 11]. For example, integrating data from different model systems, a recurrent finding is that striatal FoxP2 helps modulate neuronal plasticity involved in complex motor skills of various kinds (locomotor behaviours, manual skills and/or vocalisations)[9; 15; 16], consistent with cell-based studies showing roles of this transcription factor in neuronal differentiation and maturation[20]. Hence, the development, plasticity and maturation of the relevant circuits may be crucial for proficient speech, not only during early development[9; 15; 17], but also at post-developmental stages[16; 18]. Of note, FoxP2 is also expressed in other brain structures that have been less well studied, including the thalamus[23] and amygdala[24]. Moreover, the demonstration that this transcription factor belongs within a strongly interconnected network of brain-expressed regulatory proteins[19] underscores the complexity of pathways underlying human speech, and emphasises that we can only reach a better understanding by taking other genetic factors into account.

Given its links to human speech disorder, and its high conservation throughout the animal kingdom, *FOXP2* has also received attention from the field of evolutionary biology. One prominent focus has been on two amino-acid substitutions which occurred on the human lineage after splitting from the chimpanzee, and which have been shown to impact on striatal-dependent neurophysiology and behaviours when introduced into transgenic mice[25]. However, initial evidence of positive selection acting on intronic regulatory sequences of *FOXP2* in recent hominin evolution[26] has not been supported by subsequent systematic next-generation sequencing of global populations[27]. The details are beyond the scope of the current article, but are discussed further elsewhere (e.g. Ref. 28).

Genomic screening of disorder cohorts identifies novel risk variants

As illustrated by *FOXP2*, initial insights into the roles of rare DNA variants in developmental speech disorders came from analyses of pedigrees with multiple affected relatives across successive generations[3]. In another example of this strategy, genetic mapping in families with multiple cases of persistent stuttering (Table 1) has implicated variants in genes involved in intracellular trafficking[29] followed up further using animal models[30].

The past decade has seen emergence of another way to identify high-penetrance variants disrupting human brain development, relying not on multiplex pedigrees, but instead based around affected probands with a normal family history. Large-scale genomic screening revealed that *de novo* mutations (disruptive DNA variants found in an affected child, but absent from unaffected parents) account for a substantive proportion of cases of severe undiagnosed developmental disorders, ID, and autism spectrum disorders (ASD), among other major human disease phenotypes[31; 32]. For speech/language traits, progress has lagged behind, in part because challenges for disorder ascertainment and diagnosis have precluded systematic recruitment of large well-phenotyped cohorts[2]. Lack of consistency in criteria for detecting and classifying childhood language disorders led to establishment of a special initiative, CATALISE, in which experts worked toward consensus for the field[33]. However, issues continue to be debated by some researchers/practitioners, for example over relevance of information on general cognitive performance when diagnosing language difficulties. For disorders severely affecting speech production, like CAS, best-practice diagnostic guidelines are available from professional societies, like the American Speech-Language Hearing Association (e.g. https://www.asha.org/Practice-Portal/Clinical-Topics/Childhood-Apraxia-of-Speech/) but there remains considerable variation in how such terms are applied in practice, both clinically and for research. Identification of rare causal DNA variants could also be enhanced incorporating data from quantitative phenotyping, as has proved effective for other developmental disorders[34].

So far, a handful of phenotype-driven genome-screening studies reported rare variants in speech/language disorder cohorts, including developmental language disorder (DLD, previously often referred to as SLI) and CAS (Table 1; Figure 2). With modest sample sizes, the number of causal variants identified is small. For example, the SLI consortium performed whole exome sequencing (WES) in 43 unrelated DLD probands from the UK, identifying a *de novo* missense variant in *GRIN2A*, inherited co-segregating stop-gain variants in *OXR1* and *MUC6*, and putative pathogenic variants in a few other genes, including *SRPX2* and *ERC1*, previously implicated in speech-related disorders[35]. WES was applied only to probands, not parents; testing for *de novo*/inherited status was performed *post-hoc* using Sanger sequencing. An earlier study of this cohort used SNP-array data to investigate copy number variants (CNVs) in 127 cases, 385 first-degree relatives and 269 population controls. DLD cases carried more CNVs than controls, and the CNVs were of higher average size, but this overall increased burden was mainly driven by common events[36]. Subsequent array-based analyses of 58 severe DLD probands, 159 relatives and 76 controls, from Sweden, found that rare CNVs tended to be larger in probands, and that (both for probands and siblings) more coding genes were affected[37]. 4.8% of cases (2 of 42 tested) carried *de novo* CNVs, and 6.9% (4 of 58) had clinically significant rearrangements[37], including two cases of 16p11.2 deletion, a CNV originally identified in ASD, which has since been linked to speech/language deficits[38].

The first whole genome sequencing (WGS) study of a speech disorder investigated nineteen probands from the USA with a primary diagnosis of CAS[39]. For nine probands, WGS could also be carried out for unaffected parents, leading to identification of *de novo* single-nucleotide variants disrupting *CHD3*, *SETD1A* and *WDR5* in three cases. In the other ten probands (for whom parental DNA was unavailable) novel loss-of-function variants were found in *KAT6A*, *SETBP1*, *ZFHX4*, *TNRC6B* and *MKL2*. Through analyses of Brainspan RNA-sequencing data these CAS-related genes were found to belong to a co-expression module with high expression during early human brain development (Figure 2)[39]. More recently, WES and WGS in 34 Australian probands ascertained for CAS identified twelve rare high-confidence aetiological variants, nine of which were *de novo*[40]. In co-expression analyses using Brainspan, the ten genes highlighted in this later study (*DDX3X*, *EBF3*, *GNB1*, *MEIS2*, *SETBP1*, *UPF2*, *ZNF142*, *GNAO1*, *CDK13*, *POGZ*) showed strong overlap with the early brain-expressed gene network from the earlier WGS study of CAS, consistent with a shared pathway[39; 40].

Insights from gene-driven studies

Genome screening of CAS/DLD cohorts uncover novel genetic disruptions linked to speech disorders, but initial evidence implicating a particular gene may come from one or perhaps a few index cases. Such findings are followed-up with a gene-first approach, using information-sharing across global networks of clinical geneticists to identify independent high-risk variants in that gene, ideally regardless of routes used for proband recruitment. These efforts increase understanding of the consequences of gene disruption, evaluating variant pathogenicity through *in silico* analyses and lab-based experiments (e.g. in cellular models), and gathering data on phenotypic profiles observed in people who carry them (Figure 2).

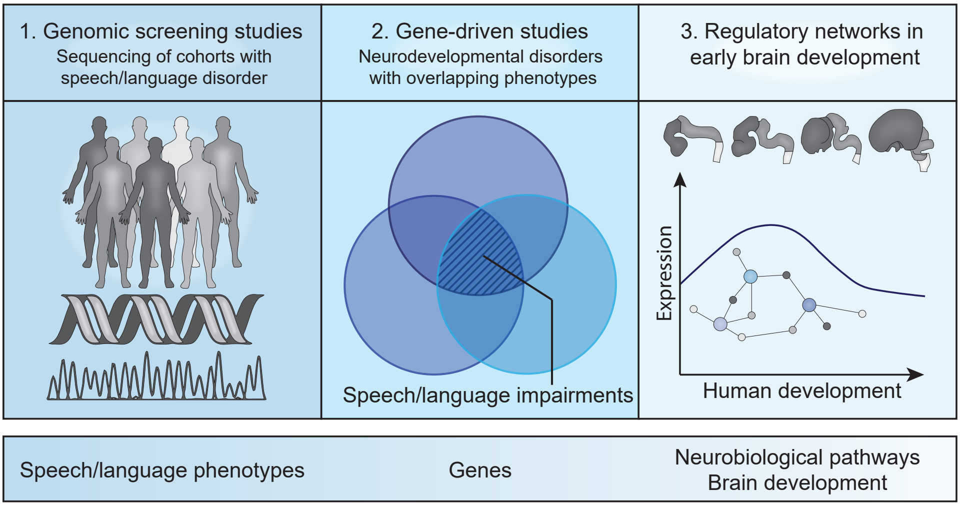

1. Genomic screening studies
Sequencing of cohorts with speech/language disorder
2. Gene-driven studies
Neurodevelopmental disorders with overlapping phenotypes
3. Regulatory networks in early brain development
Speech/language impairments
Expression
Human development
Speech/language phenotypes
Genes
Neurobiological pathways
Brain development

Figure 2. Genetic studies to identify the contribution of rare variants disrupting single genes in human developmental speech disorders.

Often when a mutation is found in an index case with a speech disorder, analyses of additional aetiological variants through gene-driven studies reveal a variable spectrum of phenotypic consequences in different individuals, including those with more severe impairments affecting multiple cognitive domains, evidence of both heterogeneity and pleiotropy (Table 2). For instance, following identification of a *de novo* microdeletion spanning *BCL11A* in a child with severe speech impairments and mild intellectual delays[41], heterozygous missense, nonsense, and frameshift variants were shown to cause a distinct syndrome involving ID (mild to severe; most cases showing moderate dysfunction) and global developmental delays, with persistence of haemoglobin representing a non-neural biomarker[42]. More recently, a *de novo* missense variant of *POU3F3* in a child with severe developmental speech/language disorder, ASD, and mild ID, led to a gene-driven study of 19 mutation cases, who showed a wide range of functioning, most having borderline-to-moderate levels of ID and/or developmental delays[43]. All had delayed expressive language, and almost all had received speech therapy; oral motor problems, word-finding difficulties, and social communication issues were common.

Variants uncovered in WGS/WES screens of CAS cohorts[39; 40] have facilitated subsequent gene-driven studies defining novel syndromes that were not previously described. Identification of a missense variant disrupting the helicase domain of *CHD3* in a proband from the first WGS screen of CAS[39] led researchers to gather 34 other individuals with *de novo* variants in the gene; overlapping features included global developmental delay and/or ID, with many showing macrocephaly and a distinctive facial phenotype[44]. Speech/language problems were common, but occurred against a wide background of levels of general cognitive dysfunction, without an obvious relationship between the specific mutation and severity.

Next-generation sequencing of CAS cohorts also identified variants in genes already investigated in earlier gene-driven studies, for which loss-of-function variants had been linked to an array of neurodevelopmental disorders, such as *SETD1A*[45]. Aetiological variants found in probands ascertained for CAS thus expand the phenotypic spectrum associated with several known neurodevelopmental disorder genes. These observations are in line with a broad consensus that single-gene disorders often show variable co-occurrence of diverse neurodevelopmental features, and that pleiotropy is a major theme – the same gene being implicated across multiple different syndromes, in ways that are not yet fully understood[46]. Curiously, *FOXP2* appears to stand out somewhat; while new cases have expanded the profile of deficits and range of severity associated with rare disruptions[4; 5], disproportionate effects on speech and language skills are consistently noted. We argue that valuable insights about speech neurobiology can be gleaned from an integrated approach - one that not only focuses on the most specific cases of disorder, but also considers data from genes linked to distinct speech phenotype profiles in only a subset of the affected people, and/or genes in shared neuromolecular pathways. Table 2 gives selected examples from the literature, with explanations of why each gene could be of interest [3; 21; 22; 35; 39-41; 43-45; 47-57].

Effects of speech-related regulatory genes on early brain development

The number of genes implicated in developmental speech disorders is still too low for comprehensive enrichment analysis, but it is intriguing that unbiased screening of CAS cohorts converged on regulatory genes co-expressed during early brain development[39; 40], with transcription factors and chromatin remodellers being prominent in gene-driven studies in this area[42-45; 50; 53]. Moreover, proteomic analyses of FOXP transcription factors identified protein-protein interactions with other brain-expressed regulatory molecules linked to neurodevelopmental diseases[19]. Involvement of regulatory genes is a common theme in aetiology of brain-related disorders, including ID[58], and experimental studies show that chromatin remodelling is crucial for differentiation and maturation of the developing brain[59-61]. So far, searches for rare gene disruptions underlying speech disorders have mainly focused on protein-coding variants, but the field could benefit from newly emerging deep-learning tools to help identify potential risk variants affecting chromatin state (DeepSEA[62]; ExPecto[63]).

As shown for *FOXP2*, animal models and cellular assays can increase understanding of gene (dys)function. Nonetheless, for disorders disturbing human capacities like speech, and that involve regulatory genes with impacts on early brain development, it could be especially valuable to also adopt more physiologically-relevant models. Brain organoids[64], grown in the lab from human stem cells, display species-specific developmental programmes[65] and capture the complex cellular diversity of the developing human cortex[66], although see Ref. 67 for important limitations. Applying such methods to patient-derived cells is illuminating pathogenic mechanisms in neurodevelopmental disorders, including idiopathic autism[68]. Long-term and pre-patterned cultures can model complex events, including neuronal activity and cellular

Table 2. Selected examples of genes that could be of interest for studying the neurobiology of human speech, including information on gene function, phenotypes associated with gene disruption in humans, and rationale for highlighting. This is not intended as a comprehensive list of all potentially relevant genes, but an illustration of the broader approach discussed in the text.

Gene	Gene function	Phenotypic profile associated with gene disruptions	Reason for highlighting
BCL11A (MIM *606557)	Transcriptional regulator - BAF complex member	ID, variable dysmorphic features including microcephaly, persistence of fetal hemoglobin (MIM #617101)	Gene disruption initially identified in a case of speech-sound disorder[41]
CHD3 (MIM *602120)	Chromatin remodeller - NuRD complex member	Mild to severe ID, dysmorphic features, macrocephaly (MIM #618205)	Index case identified in genome-wide screening of a CAS cohort[39; 44]. CHD3 is part of the FOXP2 interactome[47]
DDX3X (MIM *300160)	DEAD-box RNA helicase – involved in transcription, splicing, RNA transport, and translation	Mild to severe ID, variable dysmorphic features including microcephaly and polymicrogyria (MIM #300958)	Disruptions well established as one cause of ID[48]. A mutation case recently identified in unbiased screening of a CAS cohort[40]
ERC1 (MIM *607127)	RIM-binding protein – regulating neurotransmitter release	CAS, ID, psychiatric phenotypes	Cases of overlapping 12p13.33 microdeletions involving *ERC1* identified, with variable phenotype that includes CAS[49]
FOXP1 (MIM *605515)	Transcription factor	Mild to severe ID, dysmorphic features, speech delay, ASD (MIM #613670)	Close paralogue of FOXP2, with partially overlapping neural co-expression and potential to form heterodimers[21; 22]
FOXP2 (MIM *605317)	Transcription factor	CAS, developmental delay (MIM #602081)	First gene to be implicated in a monogenic speech disorder[3]
GATAD2B (MIM *614998)	Transcriptional regulator - NuRD complex member	ID, dysmorphic features, severe speech delay (MIM #615074)	Clinical features of mutation cases include CAS[50]. Interaction partner of CHD3[50] and FOXP proteins[51]
GRIN2A (MIM *138253)	N-methyl-D-aspartate (NMDA) receptor subunit - expressed in excitatory synapses	Focal epilepsy with speech disorder with or without ID (MIM #245570)	Putative risk variant identified in a DLD cohort study[35]. Speech disruptions described by gene-driven studies[52]
POU3F3 (MIM *602480)	Transcription factor	Mild to moderate ID, dysmorphic features, impaired speech and language acquisition (MIM #618604)	Index case identified with a severe developmental speech/language disorder[43]. Binds a regulatory region of the *FOXP2* locus
SATB2 (MIM *608148)	Transcription factor	Mild to severe ID, dysmorphic features, teeth abnormalities, severe speech delay (MIM #612313)	Disruptions well established as one cause of ID[53]. Gene-driven studies have described speech deficits[54]. Part of the FOXP interactome[19]
SCN3A (MIM *182391)	Voltage-gated sodium channel subunit - expressed in central nervous system	Familial focal epilepsy (MIM #617935), Infantile-onset refractory epilepsy, polymicrogyria (MIM #617938), prominent speech and oral motor dysfunction	Variants identified in cases with prominent speech problems[55]
SETBP1 (MIM *611060)	DNA-binding regulatory protein	Mild to severe ID with speech delay (MIM #616078), or severe ID with multiple congenital malformations (MIM #269150)	Variants identified in two independent genome-wide CAS cohort screens[39; 40]
SETD1A (MIM *611052)	H3K4 methyl transferase - HMT complex member	Range of neurodevelopmental disorders including severe developmental problems and neuropsychiatric phenotypes, including schizophrenia	Implicated In developmental disorder with a broad phenotype[45]. A mutation case identified in unbiased screening of a CAS cohort[39]
SLC6A8 (MIM *300036)	Creatine transporter - creatine transport into the brain and muscle	Mild to severe ID, severe speech delay, seizures (MIM #300352)	Described in a case of mild ID with severely affected speech[56]
UPF2 (MIM *605529)	Regulating nonsense-mediated decay - Exon Junction Complex member	Mild to severe ID, varied speech and language deficits	A mutation case identified in unbiased screening of a CAS cohort[40]. Speech phenotypes further described in a recent gene-driven study[57]

migration[69; 70], with recent studies demonstrating neuronal network formation[71; 72]. Ever more sophisticated gene-editing technologies (CRISPR and beyond) allow researchers to insert causal variants into isogenic cell-lines and/or repair mutations in patient-derived tissue, while single-cell transcriptomics facilitates systematic analyses of molecular and cellular consequences. Application of this powerful new tool-kit to rare variants implicated in developmental speech disorders could shed light on fundamental neurogenetic pathways underlying unique aspects of human biology.

Acknowledgements

We thank Lot Snijders Blok for valuable discussions. This work was supported by the Max Planck Society.

References

1. Graham, S.A., and Fisher, S.E. (2015). Understanding Language from a Genomic Perspective. Annual review of genetics 49, 131-160.
2. Deriziotis, P., and Fisher, S.E. (2017). Speech and Language: Translating the Genome. Trends in genetics : TIG 33, 642-656.
3. Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 519-523.
4. Reuter, M.S., Riess, A., Moog, U., Briggs, T.A., Chandler, K.E., Rauch, A., Stampfer, M., Steindl, K., Glaser, D., Joset, P., et al. (2017). FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. Journal of medical genetics 54, 64-72.
5. Morgan, A., Fisher, S.E., Scheffer, I., and Hildebrand, M. (2016 [updated 2017]). FOXP2-Related Speech and Language Disorders. In GeneReviews, M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, K. Stephens, and A. Amemiya, eds. (Seattle (WA), University of Washington, Seattle.
6. Fisher, S.E., and Scharff, C. (2009). FOXP2 as a molecular window into speech and language. Trends in genetics : TIG 25, 166-177.
7. French, C.A., and Fisher, S.E. (2014). What can mice tell us about Foxp2 function? Current opinion in neurobiology 28, 72-79.
8. Scharff, C., and Adam, I. (2013). Neurogenetics of birdsong. Current opinion in neurobiology 23, 29-36.
9. French, C.A., Vinueza Veloz, M.F., Zhou, K., Peter, S., Fisher, S.E., Costa, R.M., and De Zeeuw, C.I. (2019). Differential effects of Foxp2 disruption in distinct motor circuits. Mol Psychiatry 24, 447-462.
10. Argyropoulos, G.P.D., Watkins, K.E., Belton-Pagnamenta, E., Liegeois, F., Saleem, K.S., Mishkin, M., and Vargha-Khadem, F. (2019). Neocerebellar Crus I Abnormalities Associated with a Speech and Language Disorder Due to a Mutation in FOXP2. Cerebellum (London, England) 18, 309-319.
11. Liégeois, F.J., Hildebrand, M.S., Bonthrone, A., Turner, S.J., Scheffer, I.E., Bahlo, M., Connelly, A., and Morgan, A.T. (2016). Early neuroimaging markers of FOXP2 intragenic deletion. Scientific reports 6, 35192.
12. Kast, R.J., Lanjewar, A.L., Smith, C.D., and Levitt, P. (2019). FOXP2 exhibits projection neuron class specific expression, but is not required for multiple aspects of cortical histogenesis. eLife 8, e42012.
13. Co, M., Hickey, S.L., Kulkarni, A., Harper, M., and Konopka, G. (2019). Cortical Foxp2 Supports Behavioral Flexibility and Developmental Dopamine D1 Receptor Expression. Cerebral Cortex. 30.3, 1855-1870.
14. Medvedeva, V.P., Rieger, M.A., Vieth, B., Mombereau, C., Ziegenhain, C., Ghosh, T., Cressant, A., Enard, W., Granon, S., Dougherty, J.D., et al. (2018). Altered social behavior in mice carrying a cortical Foxp2 deletion. Human molecular genetics 28, 701-717.
15. van Rhijn, J.R., Fisher, S.E., Vernes, S.C., and Nadif Kasri, N. (2018). Foxp2 loss of function increases striatal direct pathway inhibition via increased GABA release. Brain structure & function 223, 4211-4226.

16. Hachigian, L.J., Carmona, V., Fenster, R.J., Kulicke, R., Heilbut, A., Sittler, A., Pereira de Almeida, L., Mesirov, J.P., Gao, F., Kolaczyk, E.D., et al. (2017). Control of Huntington's Disease-Associated Phenotypes by the Striatum-Enriched Transcription Factor Foxp2. Cell reports 21, 2688-2695.
17. Norton, P., Barschke, P., Scharff, C., and Mendoza, E. (2019). Differential song deficits after lentivirus-mediated knockdown of FoxP1, FoxP2 or FoxP4 in Area X of juvenile zebra finches. The Journal of Neuroscience, 1250-1219.
18. Day, N.F., Hobbs, T.G., Heston, J.B., and White, S.A. (2019). Beyond Critical Period Learning: Striatal FoxP2 Affects the Active Maintenance of Learned Vocalizations in Adulthood. eneuro 6, ENEURO.0071-0019.2019.
19. Estruch, S.B., Graham, S.A., Quevedo, M., Vino, A., Dekkers, D.H.W., Deriziotis, P., Sollis, E., Demmers, J., Poot, R.A., and Fisher, S.E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Human molecular genetics 27, 1212-1227.
20. Hickey, S.L., Berto, S., and Konopka, G. (2019). Chromatin Decondensation by FOXP2 Promotes Human Neuron Maturation and Expression of Neurodevelopmental Disease Genes. Cell reports 27, 1699-1711.e1699.
21. Sollis, E., Graham, S.A., Vino, A., Froehlich, H., Vreeburg, M., Dimitropoulou, D., Gilissen, C., Pfundt, R., Rappold, G.A., Brunner, H.G., et al. (2016). Identification and functional characterization of de novo FOXP1 variants provides novel insights into the etiology of neurodevelopmental disorder. Human molecular genetics 25, 546-557.
22. Sollis, E., Deriziotis, P., Saitsu, H., Miyake, N., Matsumoto, N., Hoffer, M.J.V., Ruivenkamp, C.A.L., Alders, M., Okamoto, N., Bijlsma, E.K., et al. (2017). Equivalent missense variant in the FOXP2 and FOXP1 transcription factors causes distinct neurodevelopmental disorders. Human mutation 38, 1542-1554.
23. Ebisu, H., Iwai-Takekoshi, L., Fujita-Jimbo, E., Momoi, T., and Kawasaki, H. (2017). Foxp2 Regulates Identities and Projection Patterns of Thalamic Nuclei During Development. Cerebral cortex (New York, NY : 1991) 27, 3648-3659.
24. Kuerbitz, J., Arnett, M., Ehrman, S., Williams, M.T., Vorhees, C.V., Fisher, S.E., Garratt, A.N., Muglia, L.J., Waclaw, R.R., and Campbell, K. (2018). Loss of Intercalated Cells (ITCs) in the Mouse Amygdala of Tshz1 Mutants Correlates with Fear, Depression, and Social Interaction Phenotypes. The Journal of neuroscience : the official journal of the Society for Neuroscience 38, 1160-1177.
25. Schreiweis, C., Irinopoulou, T., Vieth, B., Laddada, L., Oury, F., Burguière, E., Enard, W., and Groszer, M. (2019). Mice carrying a humanized Foxp2 knock-in allele show region-specific shifts of striatal Foxp2 expression levels. Cortex; a journal devoted to the study of the nervous system and behavior 118, 212-222.
26. Maricic, T., Gunther, V., Georgiev, O., Gehre, S., Curlin, M., Schreiweis, C., Naumann, R., Burbano, H.A., Meyer, M., Lalueza-Fox, C., et al. (2013). A recent evolutionary change affects a regulatory element in the human FOXP2 gene. Molecular biology and evolution 30, 844-852.
27. Atkinson, E.G., Audesse, A.J., Palacios, J.A., Bobo, D.M., Webb, A.E., Ramachandran, S., and Henn, B.M. (2018). No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. Cell 174, 1424-1435 e1415.
28. Fisher, S.E. (2019). Human Genetics: The Evolving Story of FOXP2. Current biology : CB 29, R65-r67.
29. Frigerio-Domingues, C., and Drayna, D. (2017). Genetic contributions to stuttering: the current evidence. Molecular genetics & genomic medicine 5, 95-102.
30. Han, T.U., Root, J., Reyes, L.D., Huchinson, E.B., Hoffmann, J.D., Lee, W.S., Barnes, T.D., and Drayna, D. (2019). Human GNPTAB stuttering mutations engineered into mice cause vocalization deficits and astrocyte pathology in the corpus callosum. Proceedings of the National Academy of Sciences of the United States of America 116, 17515-17524.
31. Deciphering Developmental Disorders, S., McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433.
32. Turner, T.N., Coe, B.P., Dickel, D.E., Hoekzema, K., Nelson, B.J., Zody, M.C., Kronenberg, Z.N., Hormozdiari, F., Raja, A., Pennacchio, L.A., et al. (2017). Genomic Patterns of De Novo Mutation in Simplex Autism. Cell 171, 710-722.e712.
33. Bishop, D.V.M., Snowling, M.J., Thompson, P.A., Greenhalgh, T., and and the, C.-c. (2017). Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. Journal of Child Psychology and Psychiatry 58, 1068-

1080.

34. Aitken, S., Firth, H.V., McRae, J., Halachev, M., Kini, U., Parker, M.J., Lees, M.M., Lachlan, K., Sarkar, A., Joss, S., et al. (2019). Finding Diagnostically Useful Patterns in Quantitative Phenotypic Data. The American Journal of Human Genetics 105, 933-946.
35. Chen, X.S., Reader, R.H., Hoischen, A., Veltman, J.A., Simpson, N.H., Francks, C., Newbury, D.F., and Fisher, S.E. (2017). Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. Scientific reports 7, 46105.
36. Simpson, N.H., Ceroni, F., Reader, R.H., Covill, L.E., Knight, J.C., Nudel, R., Monaco, A.P., Simonoff, E., Bolton, P.F., Pickles, A., et al. (2015). Genome-wide analysis identifies a role for common copy number variants in specific language impairment. European Journal of Human Genetics 23, 1370-1377.
37. Kalnak, N., Stamouli, S., Peyrard-Janvid, M., Rabkina, I., Becker, M., Klingberg, T., Kere, J., Forssberg, H., and Tammimies, K. (2018). Enrichment of rare copy number variation in children with developmental language disorder. Clin Genet 94, 313-320.
38. Mei, C., Fedorenko, E., Amor, D.J., Boys, A., Hoeflin, C., Carew, P., Burgess, T., Fisher, S.E., and Morgan, A.T. (2018). Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. European journal of human genetics : EJHG 26, 676-686.
39. Eising, E., Carrion-Castillo, A., Vino, A., Strand, E.A., Jakielski, K.J., Scerri, T.S., Hildebrand, M.S., Webster, R., Ma, A., Mazoyer, B., et al. (2019). A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. Molecular psychiatry 24, 1065-1078.
40. Hildebrand, M.S., Jackson, V.E., Scerri, T.S., Van Reyk, O., Coleman, M., Braden, R.O., Turner, S., Rigbye, K.A., Boys, A., Barton, S., et al. (2020). Severe childhood speech disorder: Gene discovery highlights transcriptional dysregulation. Neurology. 94.20, e2148-e2167.
41. Peter, B., Matsushita, M., Oda, K., and Raskind, W. (2014). De novo microdeletion of BCL11A is associated with severe speech sound disorder. American journal of medical genetics Part A 164, 2091-2096.
42. Dias, C., Estruch, S.B., Graham, S.A., McRae, J., Sawiak, S.J., Hurst, J.A., Joss, S.K., Holder, S.E., Morton, J.E., Turner, C., et al. (2016). BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. American journal of human genetics 99, 253-274.
43. Snijders Blok, L., Kleefstra, T., Venselaar, H., Maas, S., Kroes, H.Y., Lachmeijer, A.M.A., van Gassen, K.L.I., Firth, H.V., Tomkins, S., Bodek, S., et al. (2019). De Novo Variants Disturbing the Transactivation Capacity of POU3F3 Cause a Characteristic Neurodevelopmental Disorder. American journal of human genetics 105, 403-412.
44. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nature Communications 9, 4619.
45. Singh, T., Kurki, M.I., Curtis, D., Purcell, S.M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., et al. (2016). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci 19, 571-577.
46. Sanders, S.J., Sahin, M., Hostyk, J., Thurm, A., Jacquemont, S., Avillach, P., Douard, E., Martin, C.L., Modi, M.E., Moreno-De-Luca, A., et al. (2019). A framework for the investigation of rare genetic disorders in neuropsychiatry. Nature Medicine. 25.10, 1477-1487.
47. Estruch, S.B., Graham, S.A., Deriziotis, P., and Fisher, S.E. (2016). The language-related transcription factor FOXP2 is post-translationally modified with small ubiquitin-like modifiers. Scientific reports 6, 20911.
48. Lennox, A.L., Hoye, M.L., Jiang, R., Johnson-Kerner, B.L., Suit, L.A., Venkataramanan, S., Sheehan, C.J., Alsina, F.C., Fregeau, B., Aldinger, K.A., et al. (2020). Pathogenic DDX3X Mutations Impair RNA Metabolism and Neurogenesis during Fetal Cortical Development. Neuron. 106.3, 404-420.e8.
49. Thevenon, J., Callier, P., Andrieux, J., Delobel, B., David, A., Sukno, S., Minot, D., Mosca Anne, L., Marle, N., Sanlaville, D., et al. (2013). 12p13.33 microdeletion including ELKS/ERC1, a new locus associated with childhood apraxia of speech. European journal of human genetics : EJHG 21, 82-88.
50. Shieh, C., Jones, N., Vanle, B., Au, M., Huang, A.Y., Silva, A.P.G., Lee, H., Douine, E.D., Otero, M.G., Choi, A., et al. (2020). GATAD2B-associated neurodevelopmental disorder (GAND): clinical and molecular insights into a NuRD-related disorder. Genetics in medicine : official journal of the American College of Medical Genetics. 22, 878–888.

51. Chokas, A.L., Trivedi, C.M., Lu, M.M., Tucker, P.W., Li, S., Epstein, J.A., and Morrisey, E.E. (2010). Foxp1/2/4-NuRD interactions regulate gene expression and epithelial injury response in the lung via regulation of interleukin-6. The Journal of biological chemistry 285, 13304-13313.
52. Turner, S.J., Mayes, A.K., Verhoeven, A., Mandelstam, S.A., Morgan, A.T., and Scheffer, I.E. (2015). GRIN2A: an aptly named gene for speech dysfunction. Neurology 84, 586-593.
53. Bengani, H., Handley, M., Alvi, M., Ibitoye, R., Lees, M., Lynch, S.A., Lam, W., Fannemel, M., Nordgren, A., Malmgren, H., et al. (2017). Clinical and molecular consequences of disease-associated de novo mutations in SATB2. Genetics in Medicine 19, 900-908.
54. Thomason, A., Pankey, E., Nutt, B., Caffrey, A.R., and Zarate, Y.A. (2019). Speech, language, and feeding phenotypes of SATB2-associated syndrome. Clinical Genetics 96, 485-492.
55. Smith, R.S., Kenny, C.J., Ganesh, V., Jang, A., Borges-Monroy, R., Partlow, J.N., Hill, R.S., Shin, T., Chen, A.Y., Doan, R.N., et al. (2018). Sodium Channel SCN3A (NaV1.3) Regulation of Human Cerebral Cortical Folding and Oral Motor Development. Neuron 99, 905-913.e907.
56. Battini, R., Chilosi, A.M., Casarano, M., Moro, F., Comparini, A., Alessandri, M.G., Leuzzi, V., Tosetti, M., and Cioni, G. (2011). Language disorder with mild intellectual disability in a child affected by a novel mutation of SLC6A8 gene. Molecular genetics and metabolism 102, 153-156.
57. Johnson, J.L., Stoica, L., Liu, Y., Zhu, P.J., Bhattacharya, A., Buffington, S.A., Huq, R., Eissa, N.T., Larsson, O., Porse, B.T., et al. (2019). Inhibition of Upf2-Dependent Nonsense-Mediated Decay Leads to Behavioral and Neurophysiological Abnormalities by Activating the Immune Response. Neuron. 104.4, 665-679. e8.
58. Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, Merel A.W., Kleefstra, T., Kramer, Jamie M., et al. (2016). Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. The American Journal of Human Genetics 98, 149-164.
59. Nord, A.S., Pattabiraman, K., Visel, A., and Rubenstein, J.L.R. (2015). Genomic perspectives of transcriptional regulation in forebrain development. Neuron 85, 27-47.
60. de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell 172, 289-304.e218.
61. Ronan, J.L., Wu, W., and Crabtree, G.R. (2013). From neural development to cognition: unexpected roles for chromatin. Nature reviews Genetics 14, 347-359.
62. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 12, 931-934.
63. Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics 50, 1171-1179.
64. Marton, R.M., and Pasca, S.P. (2019). Organoid and Assembloid Technologies for Investigating Cellular Crosstalk in Human Brain Development and Disease. Trends Cell Biol. 30.2, 133-143.
65. Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., et al. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature 574, 418-422.
66. Velasco, S., Kedaigle, A.J., Simmons, S.K., Nash, A., Rocha, M., Quadrato, G., Paulsen, B., Nguyen, L., Adiconis, X., Regev, A., et al. (2019). Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. Nature 570, 523-527.
67. Bhaduri, A., Andrews, M.G., Mancia Leon, W., Jung, D., Shin, D., Allen, D., Jung, D., Schmunk, G., Haeussler, M., Salma, J., et al. (2020). Cell stress in cortical organoids impairs molecular subtype specification. Nature 578, 142-148.
68. Schafer, S.T., Paquola, A.C.M., Stern, S., Gosselin, D., Ku, M., Pena, M., Kuret, T.J.M., Liyanage, M., Mansour, A.A., Jaeger, B.N., et al. (2019). Pathological priming causes developmental gene network heterochronicity in autistic subject-derived neurons. Nature Neuroscience 22, 243-255.
69. Quadrato, G., Nguyen, T., Macosko, E.Z., Sherwood, J.L., Min Yang, S., Berger, D.R., Maria, N., Scholvin, J., Goldman, M., Kinney, J.P., et al. (2017). Cell diversity and network dynamics in photosensitive human brain organoids. Nature 545, 48-53.
70. Birey, F., Andersen, J., Makinson, C.D., Islam, S., Wei, W., Huber, N., Fan, H.C., Metzler, K.R.C., Panagiotakos, G., Thom, N., et al. (2017). Assembly of functionally integrated human forebrain spheroids. Nature 545, 54-59.
71. Trujillo, C.A., Gao, R., Negraes, P.D., Gu, J., Buchanan, J., Preissl, S., Wang, A., Wu, W., Haddad, G.G., Chaim, I.A., et al. (2019). Complex Oscillatory Waves Emerging from Cortical Organoids Model Early Human Brain Network Development. Cell stem cell 25, 558-569.e557.

72. Sakaguchi, H., Ozaki, Y., Ashida, T., Matsubara, T., Oishi, N., Kihara, S., and Takahashi, J. (2019). Self-Organized Synchronous Calcium Transients in a Cultured Human Neural Network Derived from Cerebral Organoids. Stem cell reports 13, 458-473.

3

This chapter has been published as:
Den Hoed, J.*, de Boer, E.*, Voisin, N.*, Dingemans, A. J., Guex, N., Wiel, L., ... & Vissers, L. E. (2021). Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction. The American Journal of Human Genetics, 108(2), 346-356.

* authors contributed equally

Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction

Abstract

Whereas large-scale statistical analyses can robustly identify disease-gene relationships, they do not accurately capture genotype-phenotype correlations or disease mechanisms. We use multiple lines of independent evidence to show that different variant types in a single gene, *SATB1*, cause clinically overlapping but distinct neurodevelopmental disorders. Clinical evaluation of 42 individuals carrying *SATB1* variants identified overt genotype-phenotype relationships, associated with different pathophysiological mechanisms, established by functional assays. Missense variants in the CUT1 and CUT2 DNA-binding domains result in stronger chromatin binding, increased transcriptional repression and a severe phenotype. In contrast, variants predicted to result in haploinsufficiency are associated with a milder clinical presentation. A similarly mild phenotype is observed for individuals with premature protein truncating variants that escape nonsense-mediated decay, which are transcriptionally active but mislocalised in the cell. Our results suggest that in-depth mutation-specific genotype-phenotype studies are essential to capture full disease complexity and to explain phenotypic variability.

Main text

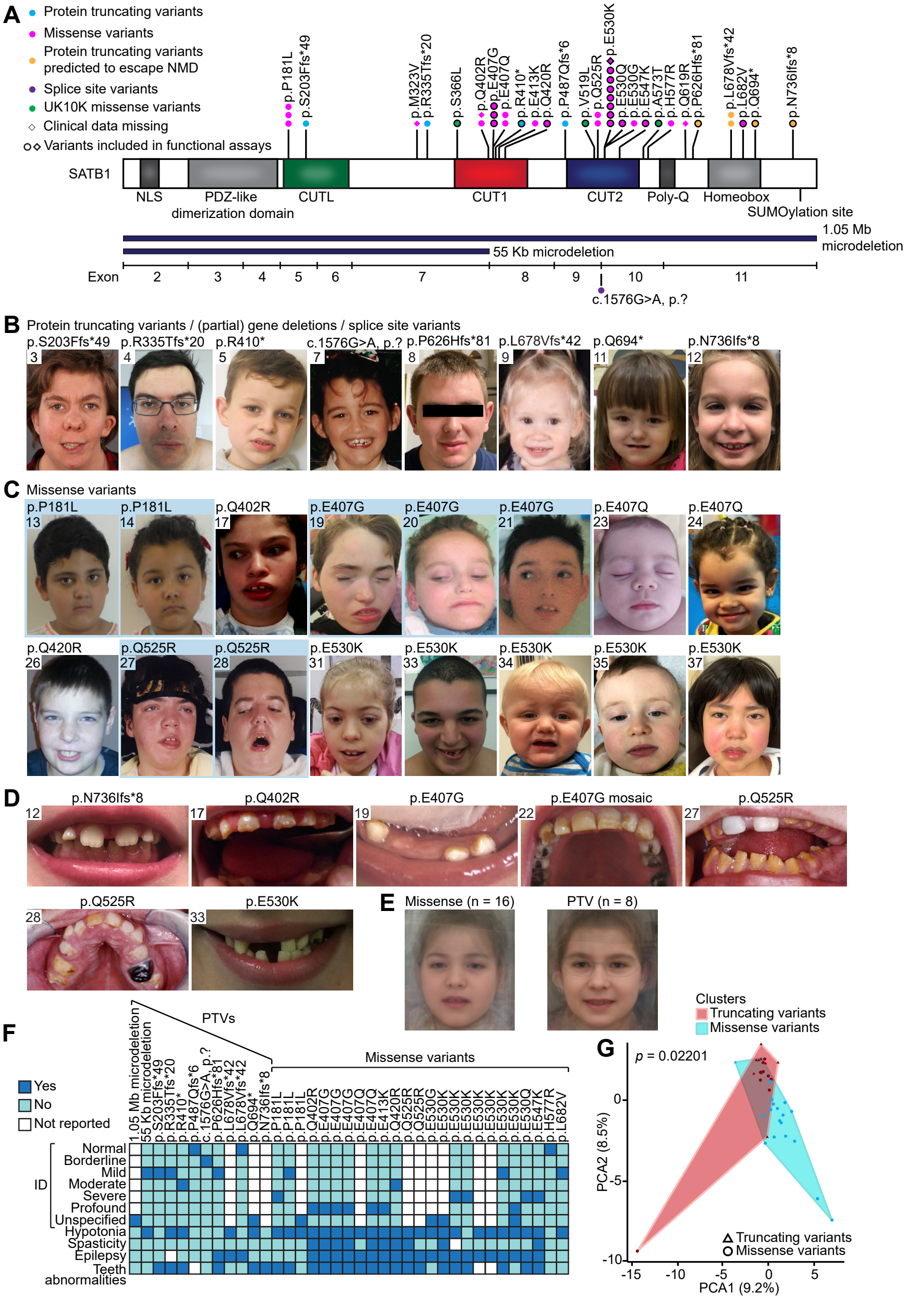
SATB1 encodes a dimeric/tetrameric transcription factor[1] with crucial roles in development and maturation of T-cells[2-4]. Recently, a potential contribution of *SATB1* to brain development was suggested by statistically significant enrichment of *de novo* variants in two large neurodevelopmental disorder (NDD) cohorts[5; 6], although its functions in the central nervous system are poorly characterised.

Through international collaborations[7-9] conforming to local ethical guidelines and the declaration of Helsinki, we identified 42 individuals with a rare (likely) pathogenic variant in *SATB1* (NM_001131010.4), a gene under constraint against loss-of-function and missense variation (pLoF: o/e=0.15 (0.08-0.29); missense: o/e=0.46 (0.41-0.52); gnomAD v2.1.1)[10]. Twenty-eight of the SATB1 variants occurred *de novo*, three were inherited from an affected parent, and five resulted from (suspected) parental mosaicism (Figure S1). Reduced penetrance is suggested by two variants inherited from unaffected parents (identified in individuals 2 and 12), consistent with recent predictions of incomplete penetrance being more prevalent in novel NDD syndromes[6]. Inheritance status of the final four could not be established. Of note, two individuals also carried a (likely) pathogenic variant affecting other known disease genes, including *NF1* (MIM #162200; individual 27) and *FOXP2* (MIM #602081; individual 42) which contributed to (individual 27) or explained (individual 42) the observed phenotype.

Thirty individuals carried 15 unique *SATB1* missense variants, including three recurrent variants (Figure 1A), significantly clustering in the highly homologous DNA-binding domains CUT1 and CUT2 (p=1.00e-7; Figure 2A, Figure S2)[11; 12]. Ten individuals carried premature protein truncating variants (PTVs; two nonsense, seven frameshift, one splice site; Table S1), and two individuals had a (partial) gene deletion (Figure S3). For 38 affected individuals and one mosaic parent, clinical information was available. Overall, we observed a broad phenotypic spectrum, characterised by neurodevelopmental delay (35/36, 97%), intellectual disability (ID) (28/31, 90%), muscle tone abnormalities (abnormal tone 28/37, 76%; hypotonia 28/37, 76%; spasticity 10/36, 28%), epilepsy (22/36, 61%) behavioural problems (24/34, 71%), facial dysmorphisms (24/36, 67%; Figure 1B-1C, Figure S4A), and dental abnormalities (24/34, 71%) (Figure 1D, Table 1, Figure S4B). Individuals with missense variants were globally more severely affected than those with PTVs: 57% of individuals with a missense variant had severe/profound ID whereas this level of ID was not observed for any individuals with PTVs. Furthermore, hypotonia, spasticity and (severe) epilepsy were more common in individuals with missense variants than in those with PTVs (92% versus 42%, 42% versus 0%, 80% versus 18%, respectively) (Figure 1F, Table 1). To objectively quantify these observations, we divided our cohort into two variant-specific clusters (missense versus PTVs) and assessed the two groups using a Partitioning Around Medoids clustering algorithm[13] on 100 features derived from standardised clinical data (Human Phenotype Ontology (HPO); Figure S5A)[14]. Thirty-eight individuals were subjected to this analysis, of which 27 were classified correctly as either belonging to the PTV or missense variant group (p=0.022), confirming the existence of at least two separate clinical entities (Figure 1G, Figure S5B). Moreover, computational averaging of facial photographs[15] revealed differences between the average facial gestalt for individuals with missense variants when compared to

Table 1. Summary of clinical characteristics associated with (*de novo*) *SATB1* variants

	All individuals		**Individuals with PTVs and (partial) gene deletions**		**Individuals with missense variants**	
	%	**Present / total assessed**	**%**	**Present / total assessed**	**%**	**Present / total assessed**
Neurologic						
Intellectual disability	90	28/31	80	8/10	95	20/21
Normal	10	3/31	20	2/10	5	1/21
Borderline	0	0/31	0	0/10	0	0/21
Mild	26	8/31	60	6/10	10	2/21
Moderate	10	3/31	10	1/10	10	2/21
Severe	19	6/31	0	0/10	29	6/21
Profound	19	6/31	0	0/10	29	6/21
Unspecified	16	5/31	10	1/10	19	4/21
Developmental delay	97	35/36	100	12/12	96	23/24
Motor delay	92	34/37	92	11/12	92	23/25
Speech delay	89	32/36	83	10/12	92	22/24
Dysarthria	30	6/20	9	1/11	56	5/9
Epilepsy	61	22/36	18	2/11	80	20/25
EEG abnormalities	79	19/24	29	2/7	100	17/17
Hypotonia	76	28/37	42	5/12	92	23/25
Spasticity	28	10/36	0	0/12	42	10/24
Ataxia	22	6/27	17	2/12	27	4/15
Behavioural disturbances	71	24/34	58	7/12	77	17/22
Sleep disturbances	41	12/29	27	3/11	50	9/18
Abnormal brain imaging	55	17/31	43	3/7	58	14/24
Regression	17	6/35	8	1/12	22	5/23
Growth						
Abnormalities during pregnancy	24	8/33	27	3/11	23	5/22
Abnormalities during delivery	32	10/31	55	6/11	20	4/20
Abnormal term of delivery	6	2/31	10	1/10	5	1/21
Preterm (<37 weeks)	6	2/31	10	1/10	5	1/21
Postterm (>42 weeks)	0	0/31	0	0/10	0	0/21
Abnormal weight at birth	16	5/32	22	2/9	13	3/23
Small for gestational age (<p10)	9	3/32	11	1/9	9	2/23
Large for gestational age (>p90)	6	2/32	11	1/9	4	1/23
Abnormal head circumference at birth	7	1/14	17	1/6	0	0/8
Microcephaly (<p3)	0	0/14	0	0/6	0	0/8
Macrocephaly (>p97)	7	1/14	17	1/6	0	0/8
Abnormal height	21	6/29	9	1/11	28	5/18
Short stature (<p3)	14	4/29	0	0/11	22	4/18
Tall stature (>p97)	7	2/29	9	1/11	6	1/18
Abnormal head circumference	23	7/31	11	1/9	27	6/22
Microcephaly (<p3)	23	7/31	11	1/9	27	6/22
Macrocephaly (>p97)	0	0/31	0	0/9	0	0/22
Abnormal weight	48	13/27	11	1/9	67	12/18
Underweight (<p3)	22	6/27	11	1/9	28	5/18
Overweight (>p97)	26	7/27	0	0/9	39	7/18
Other phenotypic features						
Facial dysmorphisms	67	24/36	64	7/11	68	17/25
Dental/oral abnormalities	71	24/34	55	6/11	78	18/23
Drooling/dysphagia	38	12/32	25	3/12	45	9/20
Hearing abnormalities	7	2/30	18	2/11	0	0/19
Vision abnormalities	55	17/31	73	8/11	45	9/20
Cardiac abnormalities	19	6/32	27	3/11	14	3/21
Skeleton/limb abnormalities	38	13/34	18	2/11	48	11/23
Hypermobility of joints	30	8/27	30	3/10	29	5/17
Gastrointestinal abnormalities	53	17/32	27	3/11	67	14/21
Urogenital abnormalities	17	5/30	0	0/11	26	5/19
Endocrine/metabolic abnormalities	30	9/30	0	0/11	47	9/19
Immunological abnormalities	32	8/25	25	2/8	35	6/17
Skin/hair/nail abnormalities	24	8/34	9	1/11	30	7/23
Neoplasms in medical history	0	0/34	0	0/11	0	0/23



F

■ Yes
▨ No
□ Not reported

		PTVs												Missense variants																								
		1.05 Mb microdeletion	55 Kb microdeletion	p.S203Ffs*49	p.R335Tfs*20	p.R410*	p.P487Qfs*6	c.1576G>A, p.?	p.P626Hfs*81	p.L678Vfs*42	p.L678Vfs*42	p.Q694*	p.N736Ifs*8	p.P181L	p.P181L	p.P181L	p.Q402R	p.E407G	p.E407G	p.E407G	p.E407Q	p.E407Q	p.E413K	p.Q420R	p.Q525R	p.Q525R	p.E530G	p.E530K	p.E530K	p.E530K	p.E530K	p.E530K	p.E530K	p.E530K	p.E530Q	p.E547K	p.H577R	p.L682V
ID	Normal	□	▨	▨	▨	▨	■	▨	▨	□	■	□	□	▨	▨	□	▨	▨	▨	▨	□	▨	▨	▨	□	□	□	□	▨	▨	□	□	▨	▨	▨	▨	■	▨
	Borderline	□	▨	▨	▨	▨	▨	■	▨	□	▨	□	□	▨	▨	□	▨	▨	▨	▨	□	▨	▨	▨	□	□	□	□	▨	▨	□	□	▨	▨	▨	▨	▨	▨
	Mild	□	■	■	■	▨	▨	▨	■	□	▨	□	□	▨	■	□	▨	▨	▨	▨	□	▨	▨	▨	□	□	□	□	▨	▨	□	□	■	▨	▨	▨	▨	■
	Moderate	□	▨	▨	▨	■	▨	▨	▨	□	▨	□	□	▨	▨	□	▨	▨	▨	▨	□	▨	▨	■	□	□	□	□	▨	▨	□	□	▨	▨	▨	▨	▨	▨
	Severe	□	▨	▨	▨	▨	▨	▨	▨	□	▨	□	□	■	▨	□	▨	▨	▨	▨	□	▨	▨	▨	□	□	□	□	■	■	□	□	▨	▨	■	■	▨	▨
	Profound	□	▨	▨	▨	▨	▨	▨	▨	□	▨	□	□	▨	▨	□	■	■	■	■	□	■	■	▨	□	□	□	□	▨	▨	□	□	▨	■	▨	▨	▨	▨
	Unspecified	■	▨	▨	▨	▨	▨	▨	▨	□	▨	■	□	▨	▨	■	▨	▨	▨	▨	□	▨	▨	▨	□	□	■	■	▨	▨	□	□	▨	■	▨	▨	▨	▨
Hypotonia		▨	■	▨	■	■	▨	▨	▨	■	▨	■	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	■	■	▨	■
Spasticity		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	■	■	■	▨	■	■	■	■	▨	▨	■	□	▨	▨	▨	▨	▨	▨	■	▨	▨
Epilepsy		▨	▨	▨	□	▨	▨	▨	■	■	■	▨	▨	▨	▨	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	▨
Teeth abnormalities		▨	▨	■	■	■	□	▨	■	▨	▨	■	■	■	■	▨	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	□	□	■	▨	■	■	▨	▨

Figure 1. Clinical evaluation of *SATB1* variants in neurodevelopmental disorders. A) Schematic representation of the SATB1 protein (NM_001131010.4/NP_001124482.1), including functional domains, with truncating variants labelled in cyan, truncating variants predicted to escape NMD in orange, splice site variants in purple, missense variants in magenta, and UK10K rare control missense variants in green. Deletions are shown in dark blue below the protein schematic, above a diagram showing the exon boundaries. We obtained clinical data for all individuals depicted by a circle.

legend continues on next page

individuals with PTVs or deletions (Figure 1B-E, Figure S4).

We performed functional analyses assessing consequences of different types of SATB1 variants for cellular localisation, transcriptional activity, overall chromatin binding, and dimerisation capacity. Based on protein modelling (Figure 2, Suppl. Notes), we selected five missense variants (observed in 14 individuals) in CUT1 and CUT2 affecting residues that interact with, or are close to, the DNA backbone (mosaic variant c.1220A>G; p.Glu407Gly and *de novo* variants c.1259A>G; p.Gln420Arg, c.1588G>A; p.Glu530Lys, c.1588G>C; p.Glu530Gln, c.1639G>A; p.Glu547Lys), as well as the only homeobox domain variant (c.2044C>G; p.Leu682Val, *de novo*). As controls, we selected three rare missense variants from the UK10K consortium, identified in healthy individuals with a normal IQ: c.1097C>T; p.Ser366Leu (gnomAD allele frequency 6.61e-4), c.1555G>C; p.Val519Leu (8.67e-6) and c.1717G>A; p.Ala573Thr (1.17e-4) (Figure 1A, Table S2)[16]. When overexpressed as YFP-fusion proteins in HEK293T/17 cells, wildtype SATB1 localised to the nucleus in a granular pattern, with an intensity profile inverse to the DNA-binding dye Hoechst 33342 (Figure 3A-B). In contrast to wildtype and UK10K control missense variants, the p.Glu407Gly, p.Gln420Arg, p.Glu530Lys/p.Glu530Gln and p.Glu547Lys variants displayed a cage-like clustered nuclear pattern, strongly co-localising with the DNA (Figure 3A-B, Figure S6).

To assess the effects of SATB1 missense variants on transrepressive activity, we used a luciferase reporter system with two previously established downstream targets of SATB1, the IL2-promoter and IgH-MAR (matrix associated region)[17-19]. All five functionally assessed CUT1 and CUT2 missense variants demonstrated increased transcriptional repression of the IL2-promoter, while the UK10K control variants did not differ from wildtype (Figure 3C). In assays using IgH-MAR, increased repression was seen for both CUT1 variants, and for one of the CUT2 variants (Figure 3C). The latter can be explained by previous reports that the CUT1 domain is essential for binding to MARs, whereas the CUT2 domain is dispensable[20; 21]. Taken together,

---

B-C) Facial photographs of individuals with (partial) gene deletions and truncations (B), and of individuals with missense variants (C). All depicted individuals show facial dysmorphisms and although overlapping features are seen, no consistent facial phenotype can be observed for the group as a whole. Overlapping facial dysmorphisms include facial asymmetry, high forehead, prominent ears, straight and/or full eyebrows, puffy eyelids, downslant of palpebral fissures, low nasal bridge, full nasal tip and full nasal alae, full lips with absent cupid's bow, prominent cupid's bow or thin upper lip vermilion. Individuals with missense variants are more alike than individuals in the truncating cohorts, and we observed recognizable overlap between several individuals in the missense cohort (individuals 17, 27, 31, 37, the siblings 19, 20 and 21, and to a lesser extent individuals 24 and 35). A recognizable facial overlap between individuals with (partial) gene deletions and truncations could not be observed. Related individuals are marked with a blue box. **D**) Photographs of teeth abnormalities observed in individuals with *SATB1* variants. Dental abnormalities are seen for all variant types and include widely spaced teeth, dental fragility, missing teeth, disorganised teeth positioning, and enamel discolouration. **E**) Computational average of facial photographs of 16 individuals with a missense variant (left) and 8 individuals with PTVs or (partial) gene deletions (right). **F**) Mosaic plot presenting a selection of clinical features. **G**) The Partitioning Around Medoids analysis of clustered HPO-standardised clinical data from 38 individuals with truncating (triangle) and missense variants (circle) shows a significant distinction between the clusters of individuals with missense variants (blue) and individuals with PTVs (red). Applying Bonferroni correction, a p-value smaller than 0.025 was considered significant. For analyses displayed in (F) and (G), individuals with absence of any clinical data and/or low level mosaicism for the *SATB1* variant were omitted (for details, see Methods section).

these data suggest that aetiological SATB1 missense variants in CUT1 and CUT2 lead to stronger binding of the transcription factor to its targets.

To study whether SATB1 missense variants affect the dynamics of chromatin binding more globally, we employed fluorescent recovery after photobleaching (FRAP) assays. Consistent with the luciferase reporter assays, all CUT1 and CUT2 missense variants, but not the UK10K control variants, affected protein mobility in the nucleus. The CUT2 variant p.Gln420Arg demonstrated an increased half time, but showed a maximum recovery similar to wildtype, while the other CUT1 and CUT2 variants demonstrated both increased halftimes and reduced maximum recovery. These results suggest stabilisation of SATB1 chromatin binding for all tested CUT1 and CUT2 variants (Figure 3D).

In contrast to the CUT1 and CUT2 missense variants, the homeobox variant p.Leu682Val did not show functional differences from wildtype (Figure 3A-D, Figure S6), suggesting that, although it is absent from gnomAD, and the position is highly intolerant to variation and evolutionarily conserved (Figure S2, Figure S7A-B), this variant is unlikely to be pathogenic. This conclusion is further supported by the presence of a valine residue at the equivalent position in multiple homologous homeobox domains (Figure S7C). Additionally, the mild phenotypic features in this individual (individual 42) can be explained by the fact that the individual carries an out-of-frame *de novo* intragenic duplication of *FOXP2*, a gene known to cause NDD through haploinsufficiency[22].

We went on to assess the impact of the CUT1 and CUT2 missense variants (p.Glu407Gly, p.Gln420Arg, p.Glu530Lys, p.Glu547Lys) on protein interaction capacities using bioluminescence resonance energy transfer (BRET). All tested variants retained the ability to interact with wildtype SATB1 (Figure 3E), with the potential to yield dominant-negative dimers/tetramers *in vivo* and to disturb normal activity of the wildtype protein.

The identification of *SATB1* deletions suggests that haploinsufficiency is a second underlying disease mechanism. This is supported by the constraint of *SATB1* against loss-of-function variation, and the identification of PTV carriers that are clinically distinct from individuals with missense variants. PTVs are found throughout the locus and several are predicted to undergo NMD by *in silico* models of NMD efficacy (Table S3)[23]. In contrast to these predictions, we found that one of the PTVs, c.1228C>T; p.Arg410*, escapes NMD (Figure S8A-B). However, the p.Arg410* variant would lack critical functional domains (CUT1, CUT2, homeobox) and indeed showed reduced transcriptional activity in luciferase reporter assays when compared to wildtype protein (Figure S8), consistent with the haploinsufficiency model.

Four unique PTVs that we identified were located within the final exon of *SATB1* (Figure 1A) and predicted to escape NMD (Table S3). Following experimental validation of NMD escape (Figure 4A-B), three such variants (c.1877delC; p.Pro626Hisfs*81, c.2080C>T; p.Gln694* and c.2207delA; p.Asn736Ilefs*8) were assessed with the same functional assays that we used for missense variants. When overexpressed as YFP-fusion proteins, the tested variants showed altered subcellular localisation,

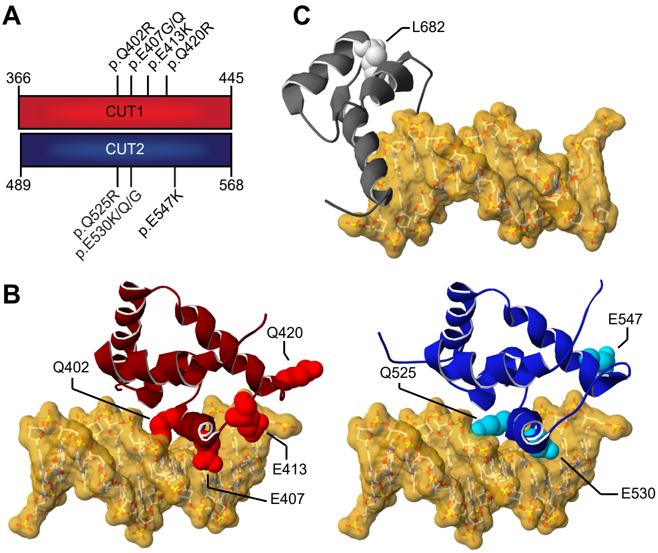


Figure 2. 3D protein modelling of SATB1 missense variants in DNA-binding domains. **A**) Schematic representation of the aligned CUT1 and CUT2 DNA-binding domains. CUT1 and CUT2 domains have a high sequence identity (40%) and similarity (78%). Note that the recurrent p.Q402R, p.E407G/p.E407Q and p.Q525R, p.E530G/p.E530K/p.E530Q variants affect equivalent positions within the respective CUT1 and CUT2 domains, while p.Q420R in CUT1 and p.E547K in CUT2 affect cognate regions. **B**) 3D-model of the SATB1 CUT1 domain (left; PDB 2O4A) and CUT2 domain (right; based on PDB 2CSF) in interaction with DNA (yellow). Mutated residues are highlighted in red for CUT1 and cyan for CUT2, along the ribbon visualisation of the corresponding domains in burgundy and dark blue, respectively. **C**) 3D-homology model of the SATB1 homeobox domain (based on PDB 1WI3 and 2D5V) in interaction with DNA (yellow). The mutated residue is shown in light grey along the ribbon visualisation of the corresponding domain in dark grey. **B-C**) For more detailed descriptions of the different missense variants in our cohort, see Suppl. Notes.

forming nuclear puncta or (nuclear) aggregates, different from patterns observed for missense variants (Figure 4C, Figure S9A-B). In luciferase reporter assays, the p.Pro626Hisfs*81 variant showed increased repression of both the IL2-promoter and IgH-MAR, whereas p.Gln694* only showed reduced repression of IgH-MAR (Figure 4D). The p.Asn736Ilefs*8 variant showed repression comparable to that of wildtype protein for both targets (Figure 4D). In further pursuit of pathophysiological mechanisms, we tested protein stability and SUMOylation, as the previously described p.Lys744 SUMOylation site is missing in all assessed NMD-escaping truncated proteins (Figure 4A)[24]. Our observations suggest the existence of multiple SATB1 SUMOylation sites (Figure S10) and no effect of NMD-escaping variants on SUMOylation of the encoded proteins (Figure S10) nor any changes in protein stability (Figure S9C). Although functional assays with NMD-escaping PTVs hint towards additional disease mechanisms, HPO-based phenotypic analysis or qualitative evaluation could not confirm a third distinct clinical entity (*p*=0.932; Figure S5, Figure S11, Table S4).

Our study demonstrates that while statistical analyses[5; 6] can provide the first step towards identification of new NDDs, a mutation-specific functional follow-up is required to gain insight into the underlying mechanisms and to understand phenotypic differences within patient cohorts. Multiple mechanisms and/or more complex genotype-phenotype correlations are increasingly appreciated in newly described NDDs, such as those associated with *RAC1*, *POLR2A*, *KMT2E* and *PPP2CA*[25-28]. Interestingly, although less often explored, such mechanistic complexity might also underlie well-known (clinically recognizable) NDDs. For instance, a CUT1 missense variant in SATB2, a paralogue of SATB1 that causes Glass syndrome through

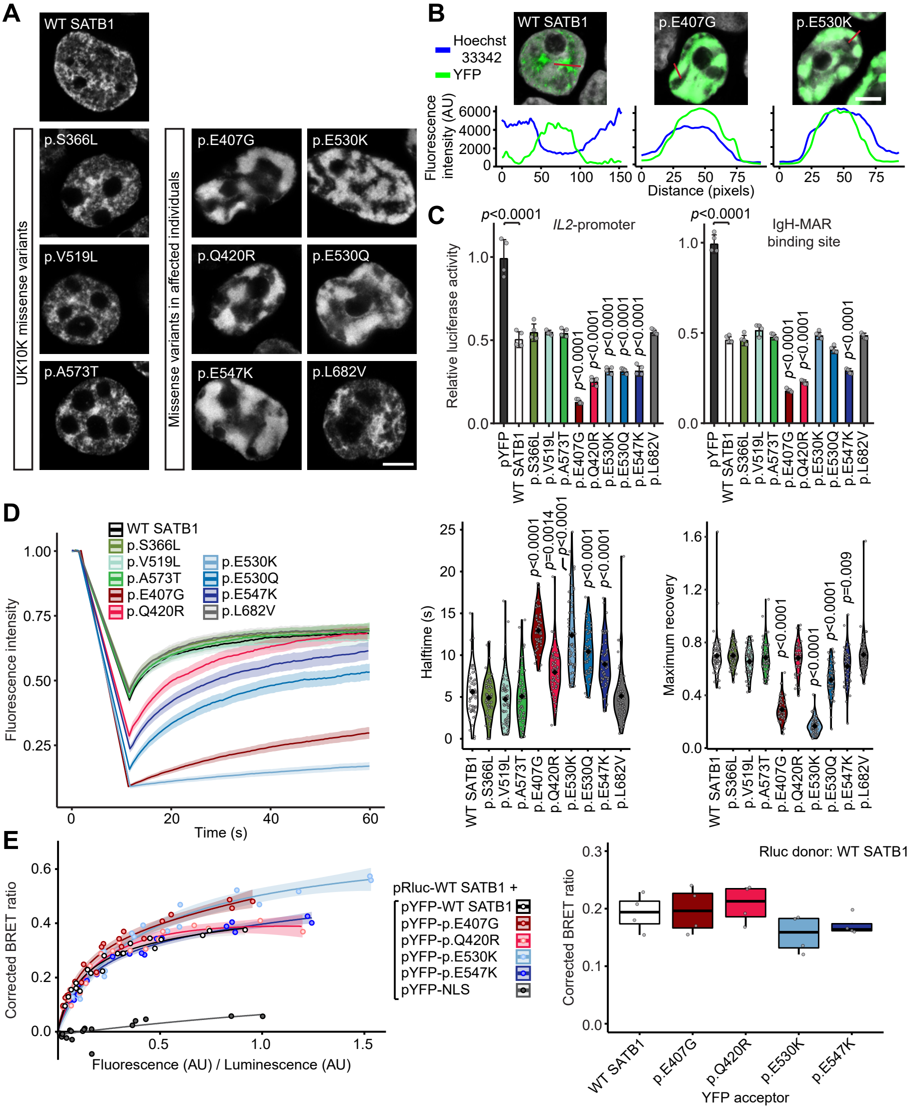


Figure 3. SATB1 missense variants stabilise DNA binding and show increased transcriptional repression. A) Direct fluorescence super-resolution imaging of nuclei of HEK293T/17 cells expressing YFP-SATB1 fusion proteins. Scale bar = 5 µm. **B**) Intensity profiles of YFP-tagged SATB1 and variants, and the DNA binding dye Hoechst 33342. The graphs represent the fluorescence intensity values of the position of the red lines drawn in the micrographs on the top (SATB1 proteins in green, Hoechst 33342 in white, scale bar = 5 µm). For each condition a representative image and corresponding intensity profile plot is shown. **C**) Luciferase reporter assays using reporter constructs containing the IL2-promoter region and the IgH matrix associated region (MAR) binding site. UK10K control variants are shaded in green, CUT1 domain variants in red, CUT2 domain variants in blue and the homeobox variant in grey. Values are expressed relative to the control (pYFP; black) and represent the mean ± S.E.M. (*n* = 4, *p*-values compared to wildtype SATB1 (WT; white), one-way ANOVA and *post-hoc* Bonferroni test). D) FRAP experiments to assess the dynamics of SATB1 chromatin binding in live cells. Left, mean recovery curves ± 95% C.I. recorded in HEK293T/17 cells expressing YFP-SATB1 fusion proteins. Right, violin plots with median of the halftime (central panel) and maximum recovery values (right panel) based on single-term exponential

legend continues on next page

haploinsufficiency (MIM #612313)[29], affects protein localisation and nuclear mobility in a similar manner to the corresponding SATB1 missense variants (Figure S12, Figure S13)[30]. Taken together, these observations suggest that mutation-specific mechanisms await discovery both for new and well-established clinical syndromes.

In summary, we demonstrate that at least two different previously uncharacterised NDDs are caused by distinct classes of rare (*de novo*) variation at a single locus. We combined clinical investigation, *in silico* models and cellular assays to characterise the phenotypic consequences and functional impacts of a large patient series uncovering distinct pathophysiological mechanisms of the *SATB1*-associated NDDs. This level of combined analyses is recommended for known and yet undiscovered NDDs to fully understand disease aetiology.

Methods

Individuals and consent

For all individuals reported in this study, informed consent was obtained to publish unidentifiable data. When applicable, specific consent was obtained for publication of clinical photographs and inclusion of photographs in facial analysis. All consent procedures are in accordance with both the local ethical guidelines of the participating centres, and the Declaration of Helsinki. Individuals with possible (likely) pathogenic *SATB1* variants were identified through international collaborations facilitated by MatchMakerExchange[7], GPAP of RD-connect[8], the Solve-RD consortium, the Decipher Database[9], and through searching literature for cohort-studies for NDD[5; 6]. Clinical characterisation was performed by reviewing the medical files and/or revising the phenotype of the individuals in the clinic. A summary of clinical characteristics is provided in Table 1, including 38 of 42 individuals: individual 16, 32 and 41 were excluded because no clinical data were available, individual 22 was excluded as she is (low) mosaic for the SATB1 variant (~1%). In Figure 1G, 37 of 42 individuals were included: in addition to individuals 16, 22, 32, and 41, we also excluded individual 18, for whom only very limited clinical information was available.

Next generation sequencing

For all individuals except individual 1, 2, and 28, SATB1 variants were identified by whole exome sequencing after variant filtering as previously described[11; 31-36]. Information on inheritance was obtained after parental confirmation, either from

---

curve fitting of individual recordings (*n* = 60 nuclei from three independent experiments, p-values compared to WT SATB1, one-way ANOVA and *post-hoc* Bonferroni test). Colour code as in C. **E**) BRET assays for SATB1 dimerisation in live cells. Left, mean BRET saturation curves ± 95% C.I. fitted using a non-linear regression equation assuming a single binding site (y = BRETmax * x / (BRET50 / x); GraphPad). The corrected BRET ratio is plotted against the ratio of fluorescence/luminescence (AU) to correct for expression level differences between conditions. Right, corrected BRET ratio values at mean BRET50 level of WT SATB1, based on curve fitting of individual experiments (*n* = 4, one-way ANOVA and *post-hoc* Bonferroni test, no significant differences). Colour code as in C. **A-E**) When compared to WT YFP-SATB1 or UK10K variants, most variants identified in affected individuals show a nuclear cage-like localisation (A), stronger co-localisation with the DNA-binding dye Hoechst 33342 (B), increased transcriptional repression (C), reduced protein mobility (D) and unchanged capacity of interaction with WT SATB1 (E).

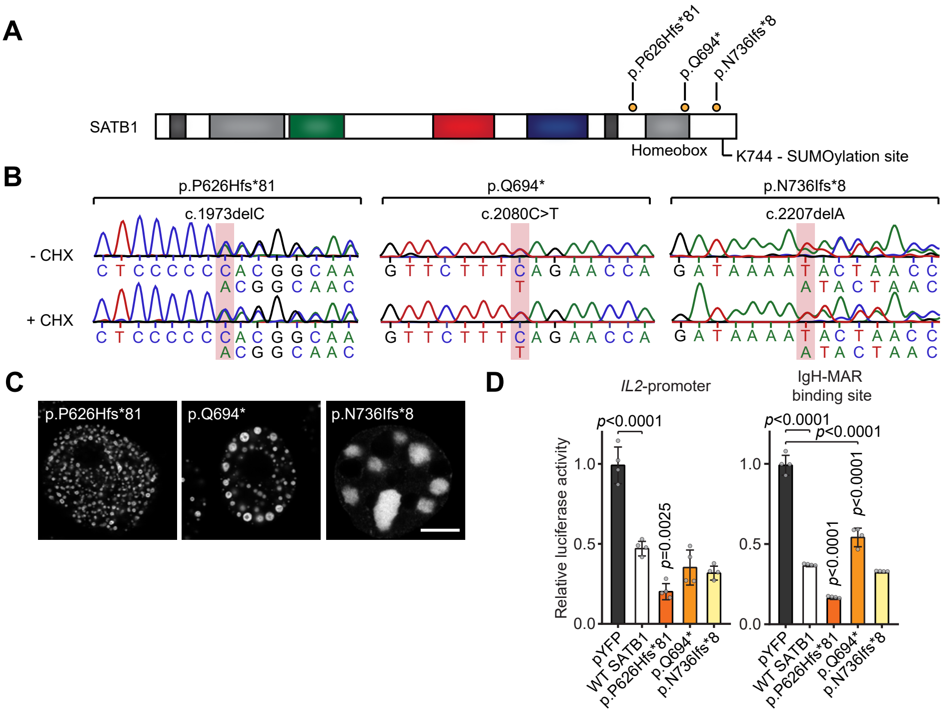


Figure 4. *SATB1* protein-truncating variants in the last exon escape NMD. A) Schematic overview of the SATB1 protein, with truncating variants predicted to escape NMD that are included in functional assays labelled in orange. A potential SUMOylation site at position p.K744 is highlighted. **B**) Sanger sequencing traces of patient-derived EBV-immortalised lymphoblastoid cell lines treated with or without cycloheximide (CHX) to test for NMD. The mutated nucleotides are shaded in red. Transcripts from both alleles are present in both conditions showing that these variants escape NMD. **C**) Direct fluorescence super-resolution imaging of nuclei of HEK293T/17 cells expressing SATB1 truncating variants fused with a YFP-tag. Scale bar = 5 µm. Compared to WT YFP-SATB1, NMD-escaping variants show altered localisation forming nuclear puncta or aggregates. **D**) Luciferase reporter assays using reporter constructs containing the IL2-promoter and the IgH matrix associated region (MAR) binding site. Values are expressed relative to the control (pYFP; black) and represent the mean ± S.E.M. (*n* = 4, *p*-values compared to WT SATB1 (white), one-way ANOVA and *post-hoc* Bonferroni test). All NMD-escaping variants are transcriptionally active and show repression of the IL2-promoter and IgH-MAR binding site.

parental exome sequencing data or through targeted Sanger sequencing. For individual 1 the *SATB1* variant was identified by array-CGH and for individual 2 an Affymetrix Cytoscan HD array was performed in addition to whole exome sequencing. For individual 28 targeted Sanger sequencing was performed after identification of the variant in his similarly affected sister. To predict deleteriousness of variants, CADD-PHRED V1.4 scores and SpliceAI scores (VCFv4.2; dated 20191004) were obtained for all variants identified in affected individuals[37; 38]. In addition, for all nonsense, frameshift and splice site variants, NMDetective scores were obtained (v2)[23]. For all missense variants, we analysed the mutation tolerance of the site of the affected residue using Metadome[39].

UK10K controls for functional assays

Genome sequence data from 1,867 ALSPAC[40; 41] individuals in the UK10K[16] dataset were annotated in ANNOVAR[42] and filtered to identify individuals carrying rare coding variants (gnomAD genome_ALL frequency<0.1%) within *SATB1*. In total six rare variants were identified. These variants were carried by 13 individuals, all in a heterozygous state. Three variants (one in the CUT1 domain, one in the CUT2 domain and one outside of critical domains) were selected for functional studies. These variants were carried by nine individuals. Phenotypic data of carriers and non-carriers were available through the ALSPAC cohort, an epidemiological study of pregnant women who were resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992. This dataset included 13,988 children who were alive at 1 year of age, 1,867 of whom underwent genome sequencing as part of the UK10K project. Of the UK10K individuals, 1,741 children had measures of IQ (WISC) collected at age 8 years providing an indication of cognitive development. The ALSPAC study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/)

Human Phenotype Ontology (HPO)-based phenotype clustering analysis

All clinical data were standardised using HPO terminology[14]. Thirty-eight of 42 individuals were included in analysis: individual 16, 32 and 41 were excluded because no clinical data were available, individual 22 was excluded as she is (low) mosaic for the SATB1 variant (~1%). The semantic similarity between all the HPO terms used in this cohort (356 features) was calculated using the Wang algorithm in the HPOSim package[43; 44] in R. HPO terms with at least a 0.5 similarity score were grouped (Figure S5): a new feature was created as a replacement, which was the sum of the grouped features. For eleven terms, the HPO semantic similarity could not be calculated using HPOSim. Seven of those could be manually assigned to a group, since the feature clearly matched (for instance: nocturnal seizures with the seizure/epilepsy group). HPO terms that could not be grouped were added as separate features, as was severity of intellectual disability. This led to 100 features for every individual, instead of the previous 356 separate HPO terms. To quantify the possible genotype/phenotype correlation in the cohort, we used Partitioning Around Medoids (PAM) clustering[13] dividing our cohort into two groups (missense variants versus truncating variants), followed by a permutations test (n=100,000) and relabelling based on variant types, while keeping the original distribution of variant types into account. The same clustering and permutations test was performed when dividing our cohort into three groups. For both analyses, Bonferroni correction for multiple testing was applied and a p-value smaller than 0.025 was considered significant.

Average face analysis

For 24 of 42 individuals facial 2D-photographs were available for facial analysis. As previously described, average faces were generated while allowing for asymmetry preservation and equal representation by individuals[15].

Three-dimensional protein modelling

The crystal structure of the CUT1 domain of SATB1 bound to Matrix Attachment Region DNA (PDB entry 2O4A[45]) was used to contextualise the SATB1 CUT1 variants with respect to DNA using Swiss-PdbViewer[46]. The solution structure of the CUT2 domain of human SATB2 (first NMR model of the PDB entry 2CSF[47]) was used as a template to align the SATB1 residues T491 to H577 (Uniprot entry Q01826), and to build a model using Swiss-PdbViewer[46]. The model of the CUT2 domain was superposed onto the SATB1 CUT1 domain bound to Matrix Attachment Region DNA (PDB entry 2O4A[45] using the "magic fit" option of Swiss-PdbViewer[46]) to contextualise the SATB1 CUT2 variants with respect to DNA. The solution structure of the homeodomain of human SATB2 (second NMR model of the PDB entry 1WI348 was used as a template to align SATB1 residues P647 to G704 (Uniprot entry Q01826), and to build a model using Swiss-PdbViewer[46]. Chains A, C and D of the crystal structure of HNF-6alpha DNA-binding domain in complex with the TTR promoter (PDB entry 2D5V)[49], which has a DNA binding domain similar to the CUT2 domain of SATB1 and a second DNA binding domain similar to the homeobox of SATB1, was used as a template to superpose the model of the SATB2 homeobox domain onto the HNF-6alpha structure using the "magic fit" option of Swiss-PdbViewer to contextualise the SATB1 homeobox variant with respect to DNA.

Spatial clustering analysis of missense variants

Twenty-four of the observed 30 missense variants were included in the spatial clustering analysis. We excluded 6 variants, to correct for familial occurrence. The geometric mean was computed over the locations of observed (*de novo*) missense variants in the cDNA of SATB1 (NM_001131010.4). This geometric mean was then compared to 1,000,000 permutations, by redistributing the (*de novo*) variant locations over the total size of the coding region of SATB1 (2,388 bp) and calculating the resulting geometric mean from each of these permutations. The p-value was then computed by checking how often the observed geometric mean distance was smaller than the permutated geometric mean distance. This approach was previously used to identify cDNA clusters of variants[11; 12].

DNA expression constructs and site-directed mutagenesis

The cloning of SATB1 (NM_001131010.4), SATB2 (NM_001172509) and SUMO1 (NM_003352.4), has been described previously[50; 51]. Variants in SATB1 and SATB2 were generated using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent). The primers used for site-directed mutagenesis are listed in Table S5. cDNAs were subcloned using *BamHI*/*XbaI* (SATB1 and SUMO1) and *BclI*/*XbaI* (SATB2) restriction sites into pRluc and pYFP, created by modification of the pEGFP-C2 vector (Clontech) as described before[52]. To generate a UBC9-SATB1 fusion, the UBC9 (NM_194260.2) and SATB1 coding sequences were amplified using primers listed in Table S6, and subcloned into the pHisV5 vector (a modified pEGFP-C2 vector adding an N-terminal His- and V5-tag) using *BamHI*/*SmaI* (UBC9) and *HindIII*/*XhoI* (SATB1) restriction sites. All constructs were verified by Sanger sequencing.

Cell culture

HEK293T/17 cells (CRL-11268, ATCC) were cultured in DMEM supplemented with 10% foetal bovine serum and 1x penicillin-streptomycin (all Invitrogen) at 37°C with 5% CO2. Transfections for functional assays were performed using GeneJuice (Millipore) following the manufacturer's protocol. Lympoblastoid cell lines (LCLs) were established by Epstein-Barr virus transformation of peripheral lymphocytes from blood samples collected in heparin tubes, and maintained in RPMI medium (Sigma) supplemented with 15% foetal bovine serum and 5% HEPES (both Invitrogen).

Testing for nonsense mediated decay of truncating variants

Patient-derived LCLs were grown for 4 h with 100 µg/ml cycloheximide (Sigma) to block NMD. After treatment, cell pellets (10*106 cells) were collected and RNA was extracted using the RNeasy Mini Kit (Qiagen). RT-PCR was performed using SuperScriptIII Reverse Transcriptase (ThermoFisher) with random primers, and regions of interest were amplified from cDNA using primers listed in Table S7.

Fluorescence microscopy

HEK293T/17 cells were grown on coverslips coated with poly-D-lysine (Sigma). Cells were fixed with 4% paraformaldehyde (PFA, Electron Microscopy Sciences) 48h after transfection with YFP-tagged SATB1 and SATB2 variants. Nuclei were stained with Hoechst 33342 (Invitrogen). Fluorescence images were acquired with a Zeiss LSM880 confocal microscope and ZEN Image Software (Zeiss). For images of single nuclei, the Airyscan unit (Zeiss) was used with a 4.5 zoom factor. All other images were acquired with a 2.0 zoom factor. Intensity profiles were plotted using the 'Plot Profile' tool in Fiji - ImageJ

FRAP assays

HEK293T/17 cells were transfected in clear-bottomed black 96-well plates with YFP-tagged SATB1 and SATB2 variants. After 48 h, medium was replaced with phenol red-free DMEM supplemented with 10% foetal bovine serum (both Invitrogen), and cells were moved to a temperature-controlled incubation chamber at 37°C. Fluorescent recordings were acquired using a Zeiss LSM880 and Zen Black Image Software, with an alpha Plan-Apochromat 100x/1.46 Oil DIC M27 objective (Zeiss). FRAP experiments were performed by photobleaching an area of 0.98 µm x 0.98 µm within a single nucleus with 488-nm light at 100% laser power for 15 iterations with a pixel dwell time of 32.97 µs, followed by collection of times series of 150 images with a 2.5 zoom factor and an optical section thickness of 1.4 µm (2.0 Airy units). Individual recovery curves were background subtracted and normalised to the pre-bleach values, and mean recovery curves were calculated using EasyFRAP software[53]. Curve fitting was done with the FrapBot application using direct normalisation and a single-component exponential model, to calculate the half-time and maximum recovery[54].

Luciferase reporter assays

Luciferase reporter assays were performed with a pIL2-luc reporter construct containing the human IL2-promoter region, and a pGL3-basic firefly luciferase reporter plasmid carrying seven repeats of the -TCTTTAATTTCTAATATATTTAGAAttc- MAR sequence identified in an enhancer region 3’ of the immunoglobulin heavy chain (IgH) genes (gift from Dr. Kathleen McGuire and Dr. Sanjeev Galande), as described previously[17-19]. HEK293T/17 cells were transfected with firefly luciferase reporter constructs and a Renilla luciferase (Rluc) normalisation control (pGL4.74; Promega) in a ratio of 50:1, and with pYFP-SATB1 (WT or variant) or empty control vector (pYFP). After 48 h, firefly luciferase and Rluc activity was measured using the Dual-Luciferase Reporter Assay system (Promega) at the Infinite M Plex Microplate reader (Tecan).

BRET saturation assays

BRET assays were performed as previously described[52]. HEK293T/17 cells were transfected in white clear-bottomed 96-well plates with increasing molar ratios of YFP-fusion proteins and constant amounts of Rluc-fusion proteins (donor/acceptor ratios of 1/0.5, 1/1, 1/2, 1/3, 1/6, 1/9). YFP and Rluc fused to a C-terminal nuclear localisation signal were used as control proteins. After 48 h, medium was replaced with phenol red-free DMEM, supplemented with 10% foetal bovine serum (both Invitrogen), containing 60 µM EnduRen Live Cell Substrate (Promega). After incubation for 4 h at 37°C, measurements were taken in live cells with an Infinite M200PRO Microplate reader (Tecan) using the Blue1 and Green1 filters. Corrected BRET ratios were calculated with the following formula: $[\text{Green1}_{(\text{experimental condition})}/\text{Blue1}_{(\text{experimental condition})}] - [\text{Green1}_{(\text{control condition})}/\text{Blue1}_{(\text{control condition})}]$, with only the Rluc control protein expressed in the control condition. YFP fluorescence was measured separately (Ex: 505 nm, Em: 545 nm) to quantify expression of the YFP-fusion proteins. Curve fitting was done with a non-linear regression equation assuming a single binding site using GraphPad Prism Software, after plotting the corrected BRET ratios against the ratio of total luminescence / total YFP fluorescence.

Immunoblotting and gel-shift assays

Whole-cell lysates were collected by treatment with lysis buffer 48 h post-transfection. For immunoblotting, cells were lysed in 1x RIPA buffer (Cell Signalling) with 1% PMSF and protease inhibitor cocktail (Roche). For gel-shift assays[55], cells were lysed in 1x RIPA buffer with 1% PMSF, protease inhibitor cocktail and 50 µM ubiquitin/ubiquitin-like isopeptidases inhibitor PR-619 (Sigma). Samples were incubated for 20 min at 4°C followed by centrifugation for 30 min at 12,000 rpm at 4°C. Proteins were resolved on 4–15% Mini-PROTEAN TGX Precast Gels (Bio-Rad) and transferred onto polyvinylidene fluoride membranes using a TransBlot Turbo Blotting system (Bio-Rad). Membranes were blocked in 5% milk for 1 h at room temperature and then probed with mouse-anti-EGFP (for pYFP constructs; 1:8000; Clontech, 632380) or mouse-anti-V5 tag (1:2000; Genetex, GTX42525). Next, membranes were incubated with HRP-conjugated goat-anti-mouse IgG (1:2000; Bio-Rad) for 1 h at room temperature. Bands were visualised with Novex ECL Chemiluminescent Substrate Reagent (Invitrogen) using a ChemiDoc XRS + System (Bio-Rad). Equal protein loading was confirmed by probing with mouse-anti-β-actin antibody (1:10,000; Sigma, A5441).

Fluorescence-based quantification of protein stability

Cells were transfected in triplicate in clear-bottomed black 96-well plates with YFP-tagged SATB1 variants. After 24 h, MG132 (R&D Systems) was added at a final concentration of 10 µM, and cycloheximide (Sigma) at 50 µg/ml. Cells were incubated at 37°C with 5% CO2 in the Infinite M200PRO microplate reader (Tecan), and the fluorescence intensity of YFP (Ex: 505 nm, Em: 545 nm) was measured over 24 h at 3 h intervals.

Statistical analyses of cell-based functional assays

Statistical analyses for cell-based functional assays were done using a one- or two-way ANOVA followed by a Bonferroni *post-hoc* test, with GraphPad Prism Software. Statistical analyses for FRAP and BRET data were performed on values derived from fitted curves of individual recordings or independent experiments respectively.

Data and Code Availability

Code used in the spatial clustering analysis is available at:

https://github.com/laurensvdwiel/SpatialClustering. Codes of HPO-based clustering analysis and computational facial averaging are available on request.

Acknowledgements

We are extremely grateful to all families participating in this study. In addition, we wish to thank the members of the Genome Technology Center and Cell culture facility, Department of Human Genetics, Radboud university medical centre, Nijmegen, for data processing and cell culture of patient-derived cell lines. This work was financially supported by Aspasia grants of the Dutch Research Council (015.014.036 to TK and 015.014.066 to LV), Netherlands Organization for Health Research and Development (91718310 to TK), the Max Planck Society (JdH, SF), Oxford Brookes University, the Leverhulme Trust and the British Academy (DN), and grants from the Swiss National Science Foundation (31003A_182632 to AR), Lithuanian-Swiss cooperation program to reduce economic and social disparities within the enlarged European Union (AR, VKu) and the Jérôme Lejeune Foundation (AR). We wish to acknowledge ALSPAC, the UK10K consortium, the 100,000 Genomes Project, “TRANSLATE NAMSE” and Genomic Answers for Kids program (see Suppl. Acknowledgements). In addition, the collaborations in this study were facilitated by ERN ITHACA, one of the 24 European Reference Networks (ERNs) approved by the ERN Board of Member States, co-funded by European Commission. The aims of this study contribute to the Solve-RD project (EdB, HB, SB, ASDP, LF, CG, AJ, TK, AV, LV) which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 779257. We wish to thank all the ALSPAC families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (Grant ref:102215/2/13/2) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). Wholegenome sequencing of the ALSPAC samples was performed as part of the UK10K consortium (a full list of investigators who contributed to the generation of the data is available from www.UK10K.org.uk). This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. In addition, in individual 13, 14 and 15, whole-exome sequencing was performed in the framework of the German project “TRANSLATE NAMSE”, an initiative from the National Action League for People with Rare Diseases (Nationales Aktionsbündnis für Menschen mit Seltenen Erkrankungen, NAMSE) facilitating innovative genetic diagnostics for individuals with suggested rare diseases. Part of this work (IT) was performed under the Genomic Answers for Kids program funded by generous donors to the Children’s Mercy Research Institute.

References

1. Wang, Z., Yang, X., Chu, X., Zhang, J., Zhou, H., Shen, Y., and Long, J. (2012). The structural basis for the oligomerization of the N-terminal domain of SATB1. Nucleic Acids Res 40, 4193-4202.
2. Alvarez, J.D., Yasui, D.H., Niida, H., Joh, T., Loh, D.Y., and Kohwi-Shigematsu, T. (2000). The MAR-binding protein SATB1 orchestrates temporal and spatial expression of multiple genes during T-cell development. Genes Dev 14, 521-535.
3. Cai, S., Lee, C.C., and Kohwi-Shigematsu, T. (2006). SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. Nat Genet 38, 1278-1288.
4. Kitagawa, Y., Ohkura, N., Kidani, Y., Vandenbon, A., Hirota, K., Kawakami, R., Yasuda, K., Motooka, D., Nakamura, S., Kondo, M., et al. (2017). Guidance of regulatory T cell development by Satb1-dependent super-enhancer establishment. Nat Immunol 18, 173-183.
5. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell 180, 568-584. e523.
6. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature 586, 757-762.
7. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat 36, 928-930.
8. Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Beroud, C., Gut, I.G., Hansson, M.G., t Hoen, P.B., Patrinos, G.P., Dawkins, H., et al. (2014). RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med 29 Suppl 3, S780-787.
9. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84, 524-533.
10. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434-443.
11. Lelieveld, S.H., Reijnders, M.R., Pfundt, R., Yntema, H.G., Kamsteeg, E.J., de Vries, B.B., Willemsen, M.H., Kleefstra, T., Lohner, K., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci 19, 1194-1196.
12. Lelieveld, S.H., Wiel, L., Venselaar, H., Pfundt, R., Vriend, G., Veltman, J.A., Brunner, H.G., Vissers, L., and Gilissen, C. (2017). Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. Am J Hum Genet 101, 478-484.
13. Kaufman L., R.P.J. (1987). Clustering by means of medoids https://wis.kuleuven.be/stat/robust/papers/publications-1987/kaufmanrousseeuw-clusteringbymedoids-l1norm-1987.pdf.
14. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 47, D1018-d1027.
15. Reijnders, M.R.F., Miller, K.A., Alvi, M., Goos, J.A.C., Lees, M.M., de Burca, A., Henderson, A., Kraus, A., Mikat, B., de Vries, B.B.A., et al. (2018). De Novo and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype-Phenotype Evaluation of a Distinct Neurodevelopmental Disorder. Am J Hum Genet 102, 1195-1203.
16. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82-90.
17. Pavan Kumar, P., Purbey, P.K., Sinha, C.K., Notani, D., Limaye, A., Jayani, R.S., and Galande, S. (2006). Phosphorylation of SATB1, a global gene regulator, acts as a molecular switch regulating its transcriptional activity in vivo. Mol Cell 22, 231-243.
18. Kumar, P.P., Purbey, P.K., Ravi, D.S., Mitra, D., and Galande, S. (2005). Displacement of SATB1-bound histone deacetylase 1 corepressor by the human immunodeficiency virus type 1 transactivator induces expression of interleukin-2 and its receptor in T cells. Mol Cell Biol 25, 1620-1633.
19. Siebenlist, U., Durand, D.B., Bressler, P., Holbrook, N.J., Norris, C.A., Kamoun, M., Kant, J.A.,

and Crabtree, G.R. (1986). Promoter region of interleukin-2 gene undergoes chromatin structure changes and confers inducibility on chloramphenicol acetyltransferase gene during activation of T cells. Mol Cell Biol 6, 3042-3049.

20. Ghosh, R.P., Shi, Q., Yang, L., Reddick, M.P., Nikitina, T., Zhurkin, V.B., Fordyce, P., Stasevich, T.J., Chang, H.Y., Greenleaf, W.J., et al. (2019). Satb1 integrates DNA binding site geometry and torsional stress to differentially target nucleosome-dense regions. Nat Commun 10, 3221.
21. Dickinson, L.A., Dickinson, C.D., and Kohwi-Shigematsu, T. (1997). An atypical homeodomain in SATB1 promotes specific recognition of the key structural element in a matrix attachment region. J Biol Chem 272, 11463-11470.
22. MacDermot, K.D., Bonora, E., Sykes, N., Coupe, A.M., Lai, C.S., Vernes, S.C., Vargha-Khadem, F., McKenzie, F., Smith, R.L., Monaco, A.P., et al. (2005). Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. Am J Hum Genet 76, 1074-1080.
23. Lindeboom, R.G.H., Vermeulen, M., Lehner, B., and Supek, F. (2019). The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. Nat Genet 51, 1645-1651.
24. Tan, J.A., Sun, Y., Song, J., Chen, Y., Krontiris, T.G., and Durrin, L.K. (2008). SUMO conjugation to the matrix attachment region-binding protein, special AT-rich sequence-binding protein-1 (SATB1), targets SATB1 to promyelocytic nuclear bodies where it undergoes caspase cleavage. J Biol Chem 283, 18124-18134.
25. Haijes, H.A., Koster, M.J.E., Rehmann, H., Li, D., Hakonarson, H., Cappuccio, G., Hancarova, M., Lehalle, D., Reardon, W., Schaefer, G.B., et al. (2019). De Novo Heterozygous POLR2A Variants Cause a Neurodevelopmental Syndrome with Profound Infantile-Onset Hypotonia. Am J Hum Genet 105, 283-301.
26. O'Donnell-Luria, A.H., Pais, L.S., Faundes, V., Wood, J.C., Sveden, A., Luria, V., Abou Jamra, R., Accogli, A., Amburgey, K., Anderlid, B.M., et al. (2019). Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental Disorders and Epilepsy. Am J Hum Genet 104, 1210-1222.
27. Reynhout, S., Jansen, S., Haesen, D., van Belle, S., de Munnik, S.A., Bongers, E., Schieving, J.H., Marcelis, C., Amiel, J., Rio, M., et al. (2019). De Novo Mutations Affecting the Catalytic Calpha Subunit of PP2A, PPP2CA, Cause Syndromic Intellectual Disability Resembling Other PP2A-Related Neurodevelopmental Disorders. Am J Hum Genet 104, 139-156.
28. Reijnders, M.R.F., Ansor, N.M., Kousi, M., Yue, W.W., Tan, P.L., Clarkson, K., Clayton-Smith, J., Corning, K., Jones, J.R., Lam, W.W.K., et al. (2017). RAC1 Missense Mutations in Developmental Disorders with Diverse Phenotypes. Am J Hum Genet 101, 466-477.
29. Zarate, Y.A., Bosanko, K.A., Caffrey, A.R., Bernstein, J.A., Martin, D.M., Williams, M.S., Berry-Kravis, E.M., Mark, P.R., Manning, M.A., Bhambhani, V., et al. (2019). Mutation update for the SATB2 gene. Hum Mutat 40, 1013-1029.
30. Lee, J.S., Yoo, Y., Lim, B.C., Kim, K.J., Choi, M., and Chae, J.H. (2016). SATB2-associated syndrome presenting with Rett-like phenotypes. Clin Genet 89, 728-732.
31. Retterer, K., Juusola, J., Cho, M.T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K.G., et al. (2016). Clinical application of whole-exome sequencing across clinical indications. Genet Med 18, 696-704.
32. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367, 1921-1929.
33. DDD-study. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223-228.
34. Gueneau, L., Fish, R.J., Shamseldin, H.E., Voisin, N., Tran Mau-Them, F., Preiksaitiene, E., Monroe, G.R., Lai, A., Putoux, A., Allias, F., et al. (2018). KIAA1109 Variants Are Associated with a Severe Disorder of Brain Development and Arthrogryposis. Am J Hum Genet 102, 116-132.
35. Brunet, T., Radivojkov-Blagojevic, M., Lichtner, P., Kraus, V., Meitinger, T., and Wagner, M. (2020). Biallelic loss-of-function variants in RBL2 in siblings with a neurodevelopmental disorder. Ann Clin Transl Neurol 7, 390-396.
36. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. Jama 312, 1870-1879.
37. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47, D886-d894.
38. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D.,

Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell 176, 535-548.e524.
39. Wiel, L., Baakman, C., Gilissen, D., Veltman, J.A., Vriend, G., and Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. Hum Mutat 40, 1030-1038.
40. Boyd, A., Golding, J., Macleod, J., Lawlor, D.A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., and Davey Smith, G. (2013). Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 42, 111-127.
41. Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., et al. (2013). Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol 42, 97-110.
42. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.
43. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., and Chen, C.F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274-1281.
44. Deng, Y., Gao, L., Wang, B., and Guo, X. (2015). HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PLoS One 10, e0115692.
45. Yamasaki, K., Akiba, T., Yamasaki, T., and Harata, K. (2007). Structural basis for recognition of the matrix attachment region of DNA by transcription factor SATB1. Nucleic Acids Res 35, 5073-5084.
46. Guex, N., and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714-2723.
47. Inoue, K., Hayashi, F., Yokoyama, S., RIKEN Structural Genomics/Proteomics Initiative (RSGI). (2005). Solution structure of the second CUT domain of human SATB2.
48. Izumi, K., Yoshida, M., Hayashi, F., Hatta, R., Yokoyama, S., RIKEN Structural Genomics/ Proteomics Initiative (RSGI). (2004). Solution structure of the homeodomain of KIAA1034 protein.
49. Iyaguchi, D., Yao, M., Watanabe, N., Nishihira, J., and Tanaka, I. (2007). DNA recognition mechanism of the ONECUT homeodomain of transcription factor HNF-6. Structure 15, 75-83.
50. Estruch, S.B., Graham, S.A., Quevedo, M., Vino, A., Dekkers, D.H.W., Deriziotis, P., Sollis, E., Demmers, J., Poot, R.A., and Fisher, S.E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Hum Mol Genet 27, 1212-1227.
51. Estruch, S.B., Graham, S.A., Deriziotis, P., and Fisher, S.E. (2016). The language-related transcription factor FOXP2 is post-translationally modified with small ubiquitin-like modifiers. Sci Rep 6, 20911.
52. Deriziotis, P., Graham, S.A., Estruch, S.B., and Fisher, S.E. (2014). Investigating protein-protein interactions in live cells using bioluminescence resonance energy transfer. J Vis Exp.
53. Koulouras, G., Panagopoulos, A., Rapsomaniki, M.A., Giakoumakis, N.N., Taraviras, S., and Lygerou, Z. (2018). EasyFRAP-web: a web-based tool for the analysis of fluorescence recovery after photobleaching data. Nucleic Acids Res 46 W467-w472.
54. Kohze, R., Dieteren, C.E.J., Koopman, W.J.H., Brock, R., and Schmidt, S. (2017). Frapbot: An open-source application for FRAP data. Cytometry A 91, 810-814.
55. Jakobs, A., Koehnke, J., Himstedt, F., Funk, M., Korn, B., Gaestel, M., and Niedenthal, R. (2007). Ubc9 fusion-directed SUMOylation (UFDS): a method to analyze function of protein SUMOylation. Nat Methods 4, 245-250.

Supplemental information

- Figure S1	Pedigrees of (suspected) mosaic families with *SATB1* variants.
- Figure S2	Amino acid sequence alignments of the CUT1, CUT2 and Homeobox domain of SATB1.
- Figure S3	Heterozygous (partial) gene deletions of the *SATB1* gene.
- Figure S4	Clinical evaluation of individuals with *SATB1* variants.
- Figure S5	Grouped HPO features based on semantic similarity and clustering results per individual.
- Figure S6	Overexpression of SATB1 missense variants as YFP-fusion proteins.
- Figure S7	MetaDome analysis of the SATB1 missense variants.
- Figure S8	Functional characterisation of the SATB1 p.R410* variant.
- Figure S9	Overexpression of SATB1 NMD-escaping PTVs as YFP-fusion proteins.
- Figure S10	SUMOylation of SATB1 protein truncating variants escaping NMD.
- Figure S11	Clinical evaluation of individuals with *SATB1* variants in three subcohorts.
- Figure S12	The SATB2 p.E396Q missense variant has comparable effects on protein functions as the p.E407G and p.E530K/Q SATB1 variants affecting equivalent positions.
- Figure S13	Missense variants identified in individuals with NDD displayed in an amino acid sequence alignment of SATB2 and SATB1.

- Table S1	Splice-AI predictions for missense variants at intron-exon or exon-intron junctions.
- Table S2	Phenotypic information of individuals from the UK10K cohort with rare *SATB1* missense variants.
- Table S3	NMD efficacy predictions for SATB1 truncating variants.
- Table S4	Summary of clinical characteristics associated with (*de novo*) SATB1 PTVs and (partial) gene deletions predicted to result in haploinsufficiency and PTVs in the last exon
- Table S5	Primers for site-directed mutagenesis.
- Table S6	Primers for amplifying and subcloning human UBC9 (NM_194260.2) and SATB1 (NM_001131010.4).
- Table S7	Primers to amplify regions that include the SATB1 NMD escaping truncating variants used for testing for NMD.

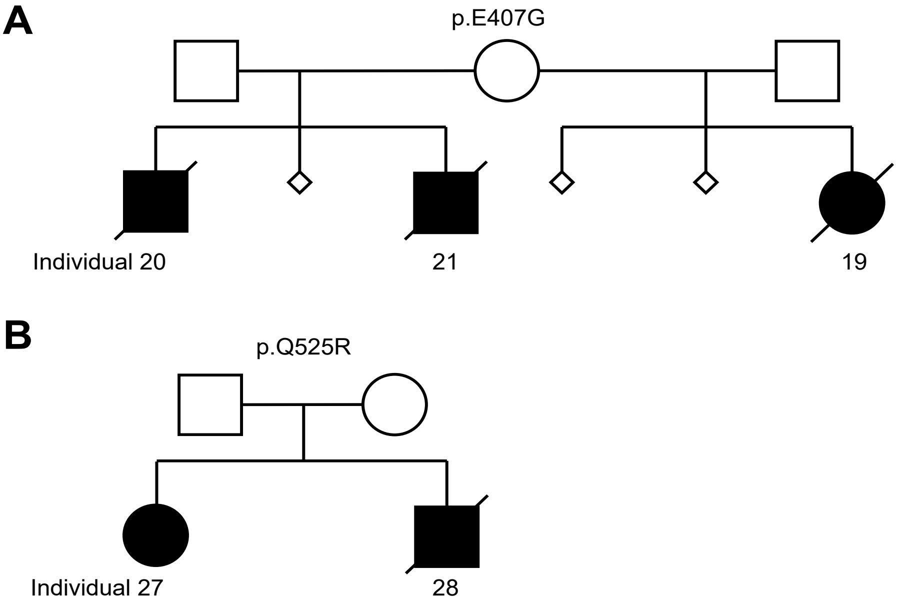
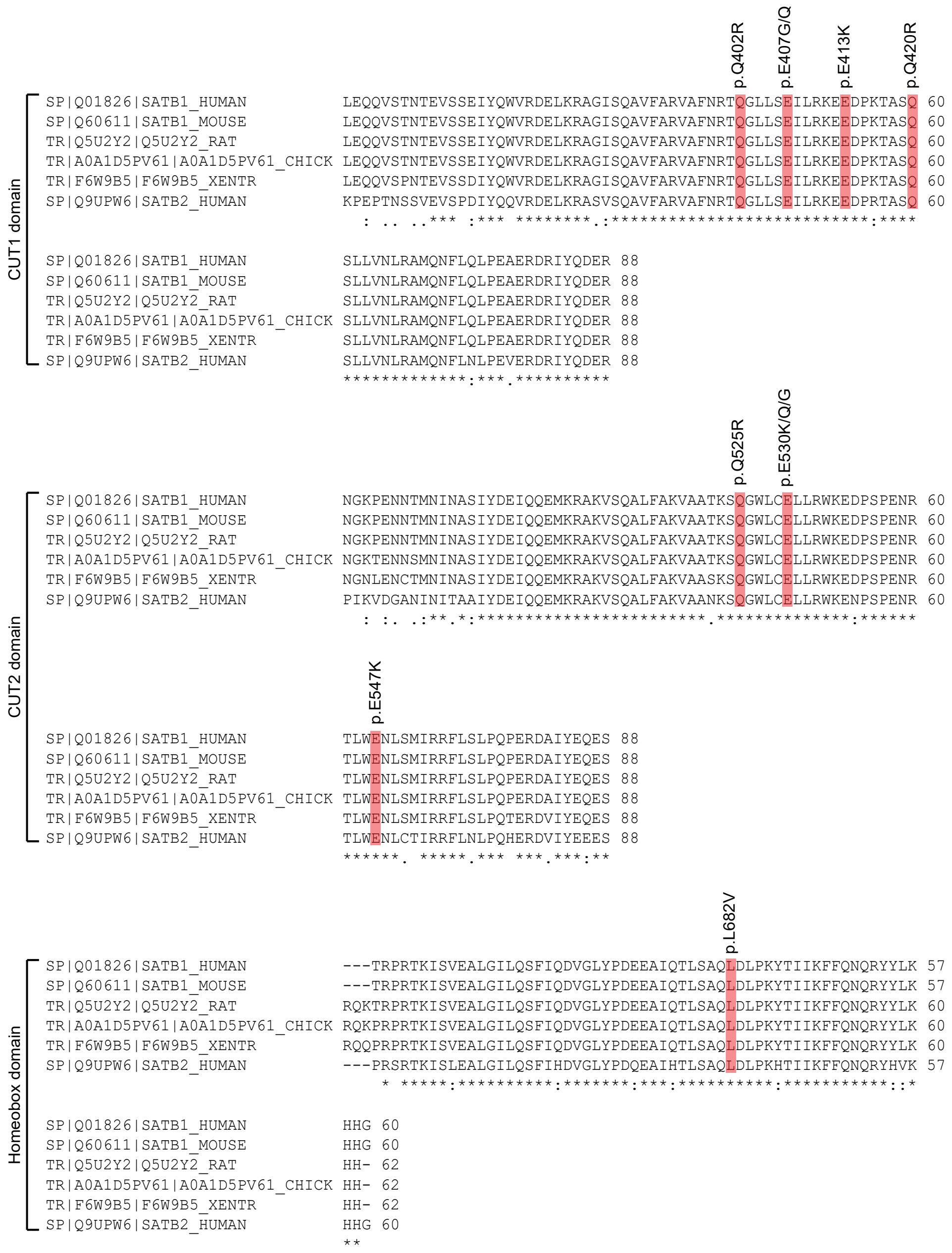


Figure S1. Pedigrees of (suspected) mosaic families with *SATB1* variants. A) Pedigree of family with proband and siblings carrying a heterozygous SATB1 p.E407G variant. The mother presents the variant in 1 of 69 reads in whole exome sequencing data, so the estimated percentage is 1.4% in the peripheral blood. Karyotyping was normal. **B**) Pedigree of family with proband and sibling carrying a heterozygous SATB1 p.Q525R variant. Suspected mosaicism in one of the parents could not be confirmed with Sanger sequencing of DNA derived from peripheral blood. **A-B**) In both families, none of the pregnancies resulted in healthy offspring.



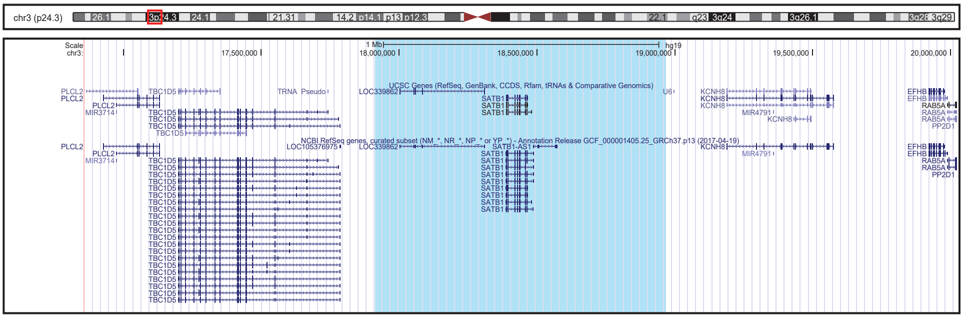
```
SP|Q01826|SATB1_HUMAN          LEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQ 60
SP|Q60611|SATB1_MOUSE          LEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQ 60
TR|Q5U2Y2|Q5U2Y2_RAT           LEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQ 60
TR|A0A1D5PV61|A0A1D5PV61_CHICK LEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQ 60
TR|F6W9B5|F6W9B5_XENTR         LEQQVSPNTEVSSDIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQ 60
SP|Q9UPW6|SATB2_HUMAN          KPEPTNSSVEVSPDIYQQVRDELKRASVSQAVFARVAFNRTQGLLSEILRKEEDPRTASQ 60
                                 : .. ..*** :*** ********.:***************************:****

SP|Q01826|SATB1_HUMAN          SLLVNLRAMQNFLQLPEAERDRIYQDER 88
SP|Q60611|SATB1_MOUSE          SLLVNLRAMQNFLQLPEAERDRIYQDER 88
TR|Q5U2Y2|Q5U2Y2_RAT           SLLVNLRAMQNFLQLPEAERDRIYQDER 88
TR|A0A1D5PV61|A0A1D5PV61_CHICK SLLVNLRAMQNFLQLPEAERDRIYQDER 88
TR|F6W9B5|F6W9B5_XENTR         SLLVNLRAMQNFLQLPEAERDRIYQDER 88
SP|Q9UPW6|SATB2_HUMAN          SLLVNLRAMQNFLNLPEVERDRIYQDER 88
                               *************:***.**********
```

```
SP|Q01826|SATB1_HUMAN          NGKPENNTMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENR 60
SP|Q60611|SATB1_MOUSE          NGKPENNTMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENR 60
TR|Q5U2Y2|Q5U2Y2_RAT           NGKPENNTMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENR 60
TR|A0A1D5PV61|A0A1D5PV61_CHICK NGKTENNSMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENR 60
TR|F6W9B5|F6W9B5_XENTR         NGNLENCTMNINASIYDEIQQEMKRAKVSQALFAKVAASKSQGWLCELLRWKEDPSPENR 60
SP|Q9UPW6|SATB2_HUMAN          PIKVDGANINITAAIYDEIQQEMKRAKVSQALFAKVAANKSQGWLCELLRWKENPSPENR 60
                                 : :. .:**.*:************************.**************:******

SP|Q01826|SATB1_HUMAN          TLWENLSMIRRFLSLPQPERDAIYEQES 88
SP|Q60611|SATB1_MOUSE          TLWENLSMIRRFLSLPQPERDAIYEQES 88
TR|Q5U2Y2|Q5U2Y2_RAT           TLWENLSMIRRFLSLPQPERDAIYEQES 88
TR|A0A1D5PV61|A0A1D5PV61_CHICK TLWENLSMIRRFLSLPQPERDAIYEQES 88
TR|F6W9B5|F6W9B5_XENTR         TLWENLSMIRRFLSLPQTERDVIYEQES 88
SP|Q9UPW6|SATB2_HUMAN          TLWENLCTIRRFLNLPQHERDVIYEEES 88
                               ******. *****.*** ***.***:**
```

```
SP|Q01826|SATB1_HUMAN          ---TRPRTKISVEALGILQSFIQDVGLYPDEEAIQTLSAQLDLPKYTIIKFFQNQRYYLK 57
SP|Q60611|SATB1_MOUSE          ---TRPRTKISVEALGILQSFIQDVGLYPDEEAIQTLSAQLDLPKYTIIKFFQNQRYYLK 57
TR|Q5U2Y2|Q5U2Y2_RAT           RQKTRPRTKISVEALGILQSFIQDVGLYPDEEAIQTLSAQLDLPKYTIIKFFQNQRYYLK 60
TR|A0A1D5PV61|A0A1D5PV61_CHICK RQKPRPRTKISVEALGILQSFIQDVGLYPDEEAIQTLSAQLDLPKYTIIKFFQNQRYYLK 60
TR|F6W9B5|F6W9B5_XENTR         RQQPRPRTKISVEALGILQSFIQDVGLYPDEEAIQTLSAQLDLPKYTIIKFFQNQRYYLK 60
SP|Q9UPW6|SATB2_HUMAN          ---PRSRTKISLEALGILQSFIHDVGLYPDQEAIHTLSAQLDLPKHTIIKFFQNQRYHVK 57
                                   * *****:**********:*******:***:**********:***********::*

SP|Q01826|SATB1_HUMAN          HHG 60
SP|Q60611|SATB1_MOUSE          HHG 60
TR|Q5U2Y2|Q5U2Y2_RAT           HH- 62
TR|A0A1D5PV61|A0A1D5PV61_CHICK HH- 62
TR|F6W9B5|F6W9B5_XENTR         HH- 62
SP|Q9UPW6|SATB2_HUMAN          HHG 60
                               **
```

Figure S2. Amino acid sequence alignments of the CUT1, CUT2 and Homeobox domain of SATB1. Amino acid sequences of the CUT1, CUT2 and Homeobox domain of human SATB1 (Q01826, UniProt) aligned to the mouse (Q60611), rat (Q5U2Y2), chicken (A0A1D5PV61) and *Xenopus tropicalis* (F6W9B5) sequences, and the sequences of the homologue domains in human SATB2 (Q9UPW6). Alignment was performed with Clustal Omega (1.2.4) with default settings using UniProt alignment tool. Missense variants described in this study and identified in these functional domains are shaded in red.

1.05Mb deletion - chr3:17915162-18968823



55Kb deletion - chr3:18376866-18432504

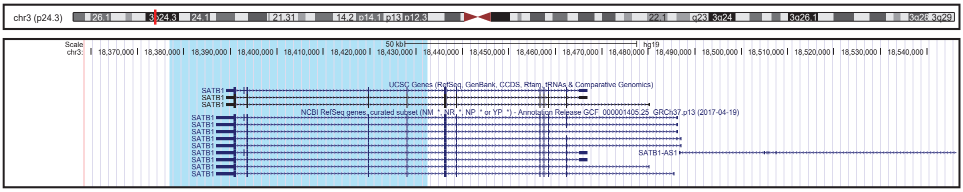
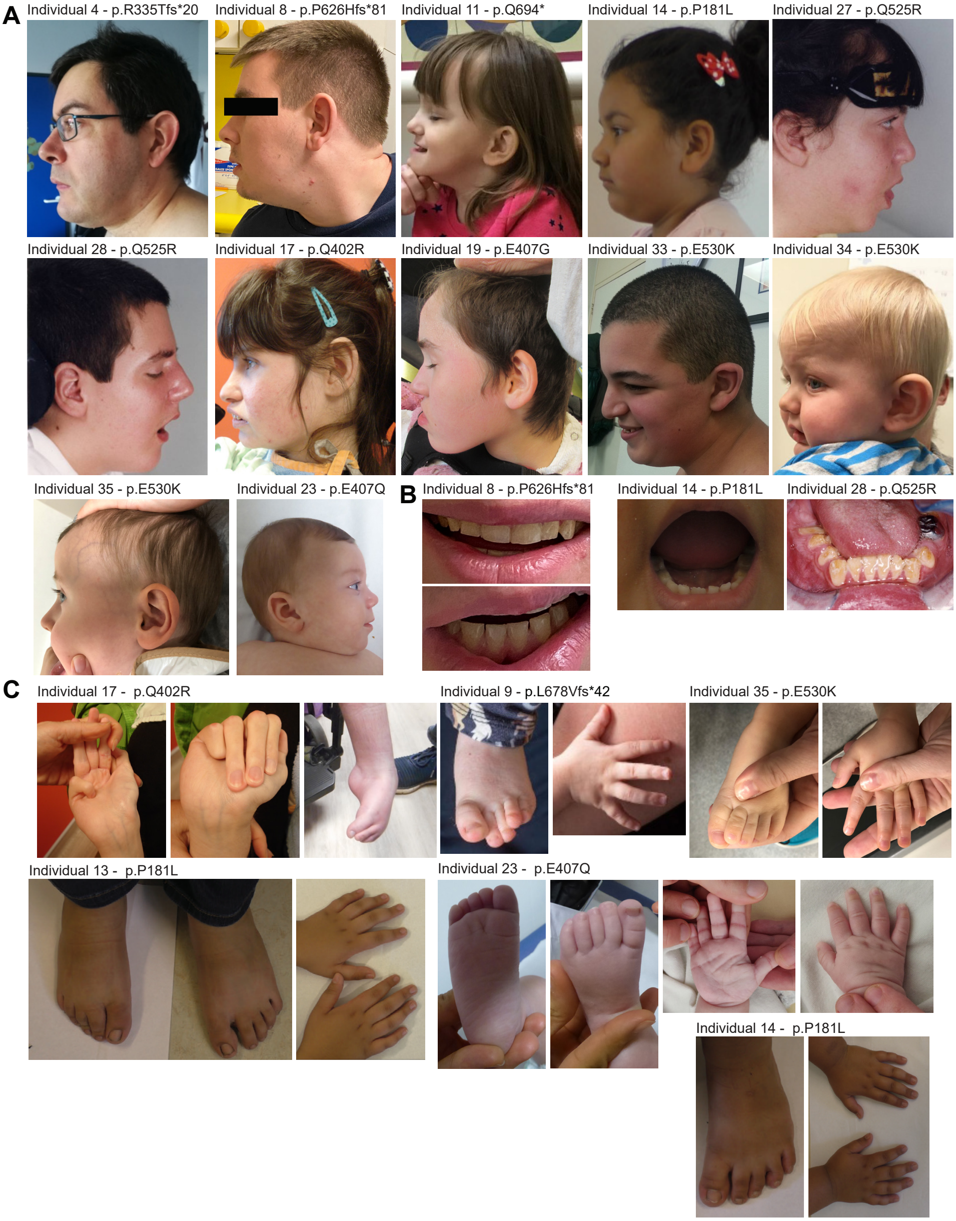


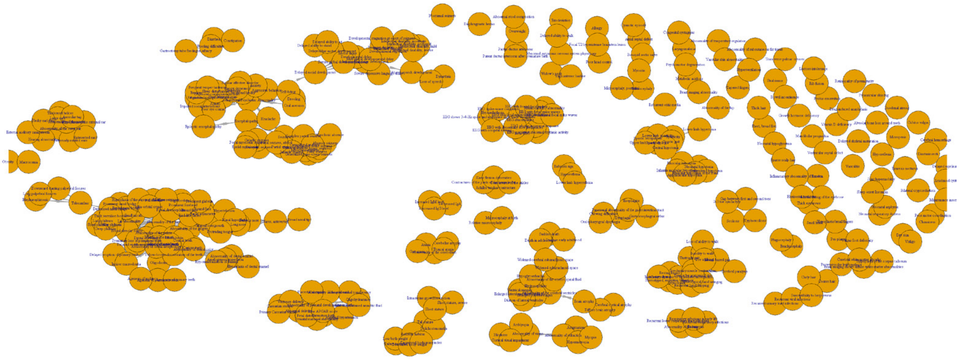
Figure S3. Heterozygous (partial) gene deletions of the *SATB1* gene. Genome overviews of two reported heterozygous deletions that include the *SATB1* gene, generated in the UCSC Genome Browser (assembly Feb. 2009 GRCh37/hg19). The deleted regions are shaded in red in the chromosome ideogram, and in light blue in the genome overview.



[redacted]

Figure S4. Clinical evaluation of individuals with SATB1 variants. A) Side view photographs, depicting prominent ears (individuals 4, 8, 14, 17, 19, 34, 35), with thickened helices (individuals 8, 14, 17, 19, 33, 34, 35), and retrognathia (individuals 8, 14, 17, 19, 27, 34). **B**) Additional photograph of teeth. No evident enamel or dental positioning problems in individuals 8 and 14, although missing molars (individual 8) and malformed teeth (individual 14) are reported. Lower teeth of individual 28: discolouration, malpositioning and teeth decay. **C**) Photographs of hands and feet. Features include contractures resulting from spasticity (individual 17), tapered fingers (individuals 13, 14, 23, 35), short broad fingers (individuals 13, 14, 23), clinodactyly of 5th finger (individual 9), overlapping 2nd toe (individual 35) or 4th toe (individual 9) and broad feet with short toes and small toe nails (individuals 13, 14, 23).

A



B

Identifier	Variant type	2_cluster_pred	2_correct	3_cluster_pred	3_correct
Individual1	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual2	PTV_non_last_exon	PTV	CORRECT	PTV_non_last_exon	CORRECT
Individual3	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual4	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual5	PTV_non_last_exon	Missense	INCORRECT	Missense	INCORRECT
Individual6	PTV_non_last_exon	PTV	CORRECT	PTV_last_exon	INCORRECT
Individual7	PTV_non_last_exon	Missense	INCORRECT	PTV_last_exon	INCORRECT
Individual8	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT
Individual9	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT
Individual10	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT
Individual11	PTV_last_exon	PTV	CORRECT	PTV_non_last_exon	INCORRECT
Individual12	PTV_last_exon	PTV	CORRECT	PTV_last_exon	CORRECT
Individual13	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual14	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual15	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual17	Missense	Missense	CORRECT	Missense	CORRECT
Individual18	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual19	Missense	Missense	CORRECT	Missense	CORRECT
Individual20	Missense	Missense	CORRECT	Missense	CORRECT
Individual21	Missense	Missense	CORRECT	Missense	CORRECT
Individual23	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual24	Missense	Missense	CORRECT	Missense	CORRECT
Individual25	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual26	Missense	Missense	CORRECT	Missense	CORRECT
Individual27	Missense	Missense	CORRECT	Missense	CORRECT
Individual28	Missense	Missense	CORRECT	Missense	CORRECT
Individual29	Missense	Missense	CORRECT	Missense	CORRECT
Individual30	Missense	Missense	CORRECT	Missense	CORRECT
Individual31	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual33	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual34	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual35	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual36	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual37	Missense	Missense	CORRECT	Missense	CORRECT
Individual38	Missense	Missense	CORRECT	Missense	CORRECT
Individual39	Missense	Missense	CORRECT	PTV_non_last_exon	INCORRECT
Individual40	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Individual42	Missense	PTV	INCORRECT	PTV_last_exon	INCORRECT
Correctly predicted individuals:			27		17

Figure S5. Grouped HPO features based on semantic similarity and clustering results per individual. A) The semantic similarity between all the HPO terms used in this cohort (356 features) was calculated using the Wang algorithm in the HPOsim package in R. HPO terms with at least a 0.5 similarity score were grouped and a new feature was created as a replacement, which was the sum of the grouped features. **B**) Individual HPO-based phenotypic clustering results for both analyses with two and three clusters.

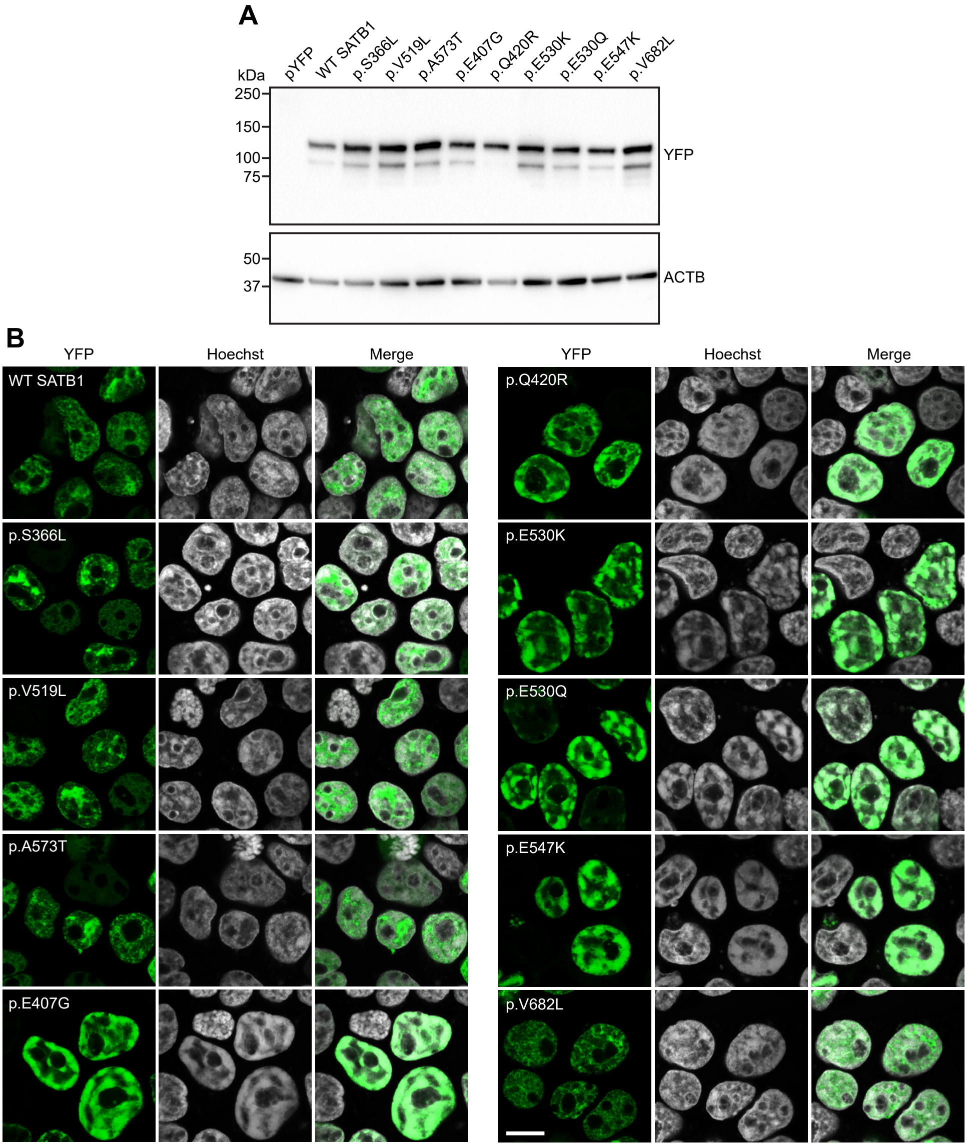
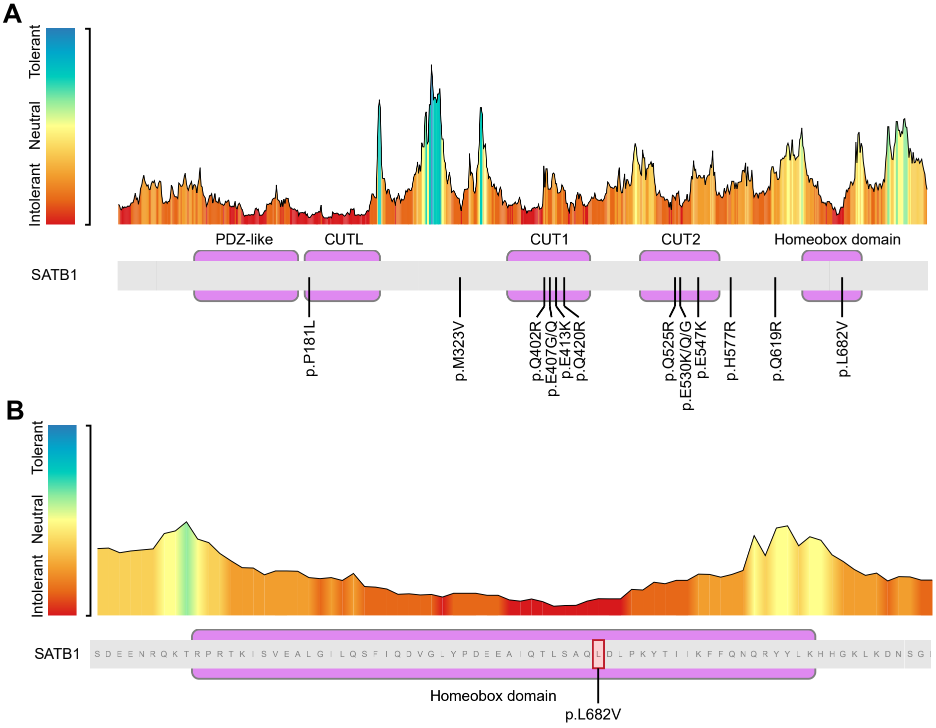


Figure S6. Overexpression of SATB1 missense variants as YFP-fusion proteins. A) Immunoblot of whole-cell lysates expressing YFP-tagged SATB1 variants probed with anti-EGFP antibody. Expected molecular weight for all variants is ~115 kDa. The blot was probed for ACTB to ensure equal protein loading. **B**) Direct fluorescence micrographs of HEK293T/17 cells expressing YFP-SATB1 fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 μm.



C

Gene name	gDNA	cDNA	Residue change	Type	gnomAD allele frequency
HMX3	chr10:124896963C>G	NM_001105574.1:c.790C>G	p.(Leu264Val)	missense	0.000004
HOXC10	chr12:54383114A>G	NM_017409.3:c.913A>G	p.(Ile305Val)	missense	0.000004
HOXC11	chr12:54369087C>G	NM_014212.3:c.805C>G	p.(Leu269Val)	missense	0.000004
ISX	chr22:35478636A>G	NM_001303508.1:c.355A>G	p.(Ile119Val)	missense	0.000004
NANOGNB	chr12:7922891T>G	NM_001145465.1:c.415T>G	p.(Phe139Val)	missense	0.000023
NKX1-2	chr10:126136333G>C	NM_001146340.2:c.598C>G	p.(Leu200Val)	missense	0.000009
NKX2-2	chr20:21492890T>C	NM_002509.3:c.493A>G	p.(Ile165Val)	missense	0.000004
NOBOX	chr7:144097323C>T	NM_001080413.3:c.927G>A	p.(Val309=)	synonymous	0.000004
OTP	chr5:76932672T>C	NM_032109.2:c.421A>G	p.(Ile141Val)	missense	0.00007
PAX3	chr2:223096822G>A	NM_181459.3:c.767C>T	p.(Ala256Val)	missense	0.000004
POU2F2	chr19:42599569G>C	NM_001207025.2:c.1000C>G	p.(Leu334Val)	missense	0.000004
POU6F1	chr12:51584125G>C	NM_001330422.1:c.1741C>G	p.(Leu581Val)	missense	0.000004
ZFHX3	chr16:72828547C>T	NM_006885.3:c.8034G>A	p.(Val2678=)	synonymous	0.000004
ZFHX4	chr8:77767083C>A	NM_024721.4:c.7926C>A	p.(Val2642=)	synonymous	0.000096
ZFHX4	chr8:77767083C>T	NM_024721.4:c.7926C>T	p.(Val2642=)	synonymous	0.000008

Figure S7. MetaDome analysis of the SATB1 missense variants. A) Overview of the SATB1 protein (transcript NM_001131010.2) tolerance landscape. All missense variants identified in affected individuals are indicated. **B**) Detailed overview of the SATB1 homeobox domain tolerance landscape, with the p.L682V variant indicated. **C**) Table listing all residue changes at positions equivalent to the SATB1 p.L682 position in homologue homeobox domain proteins that change to a valine. The gnomAD allele frequency is indicated.

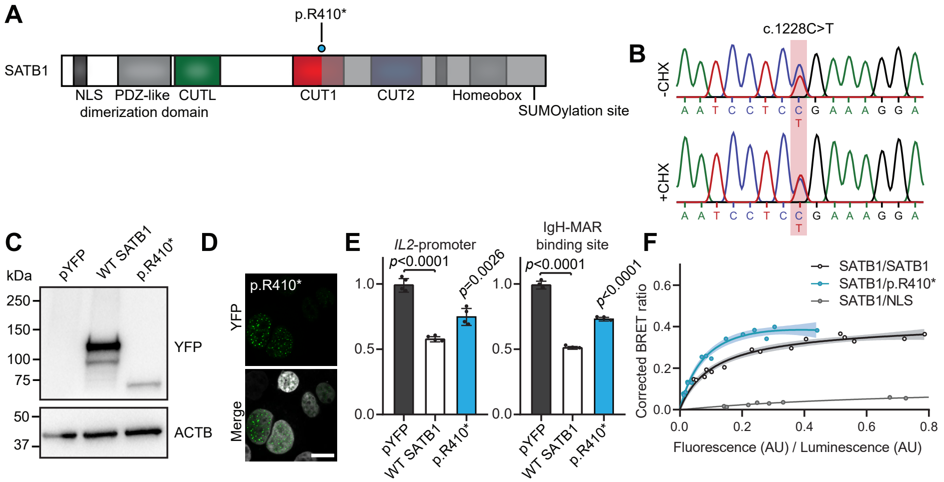


Figure S8. Functional characterisation of the SATB1 p.R410* variant. A) Schematic representation of SATB1 with the p.R410* variant labelled in cyan. **B**) Sanger sequencing traces of patient-derived EBV transformed lymphoblastoid cell lines treated with or without cycloheximide (CHX) to test for NMD. The mutated nucleotides are shaded in red. **C**) Immunoblot of whole-cell lysates expressing YFP-tagged SATB1 and p.R410* probed with anti-EGFP antibody. Expected molecular weight is SATB1: ~115 kDa, p.R410*: ~75kDa. The blot was probed for ACTB to ensure equal protein loading. **D**) Direct fluorescence micrographs of HEK293T/17 cells expressing YFP-SATB1 p.R410* fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 µm. **E**) Luciferase reporter assays using reporter constructs containing the IL2 promoter region and the IgH matrix associated region (MAR) binding site. Values are expressed relative to the control (pYFP; black) and represent the mean ± S.E.M. (*n* = 4 for IL2-promoter, *n* = 3 for IgH-MAR binding site, p-values compared to wildtype (WT) SATB1 (white), one-way ANOVA and *post-hoc* Bonferroni test). **F**) BRET assays for SATB1 dimerisation in live cells. The plot shows the mean BRET saturation curves ± 95% C.I. fitted using a non-linear regression equation assuming a single binding site (y = BRETmax * x / (BRET50 / x); GraphPad). The corrected BRET ratio is plotted against the ratio of fluorescence/luminescence (AU) to correct for expression level differences between conditions (*n* = 3).

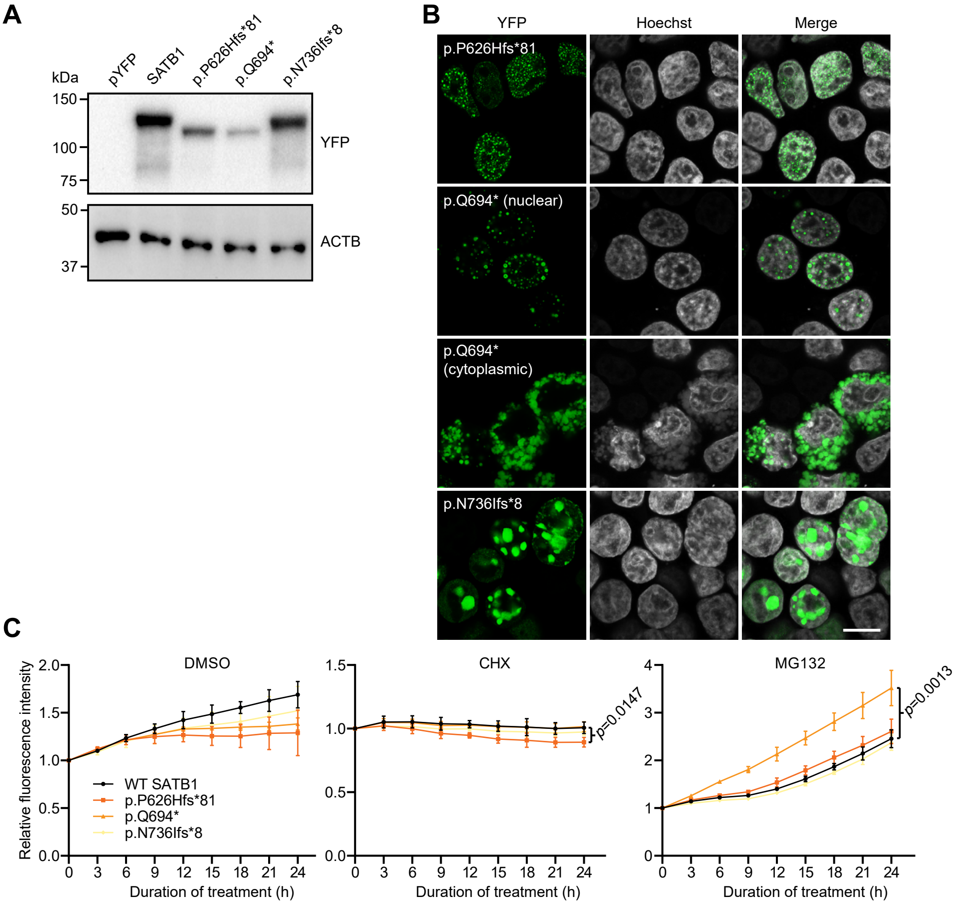
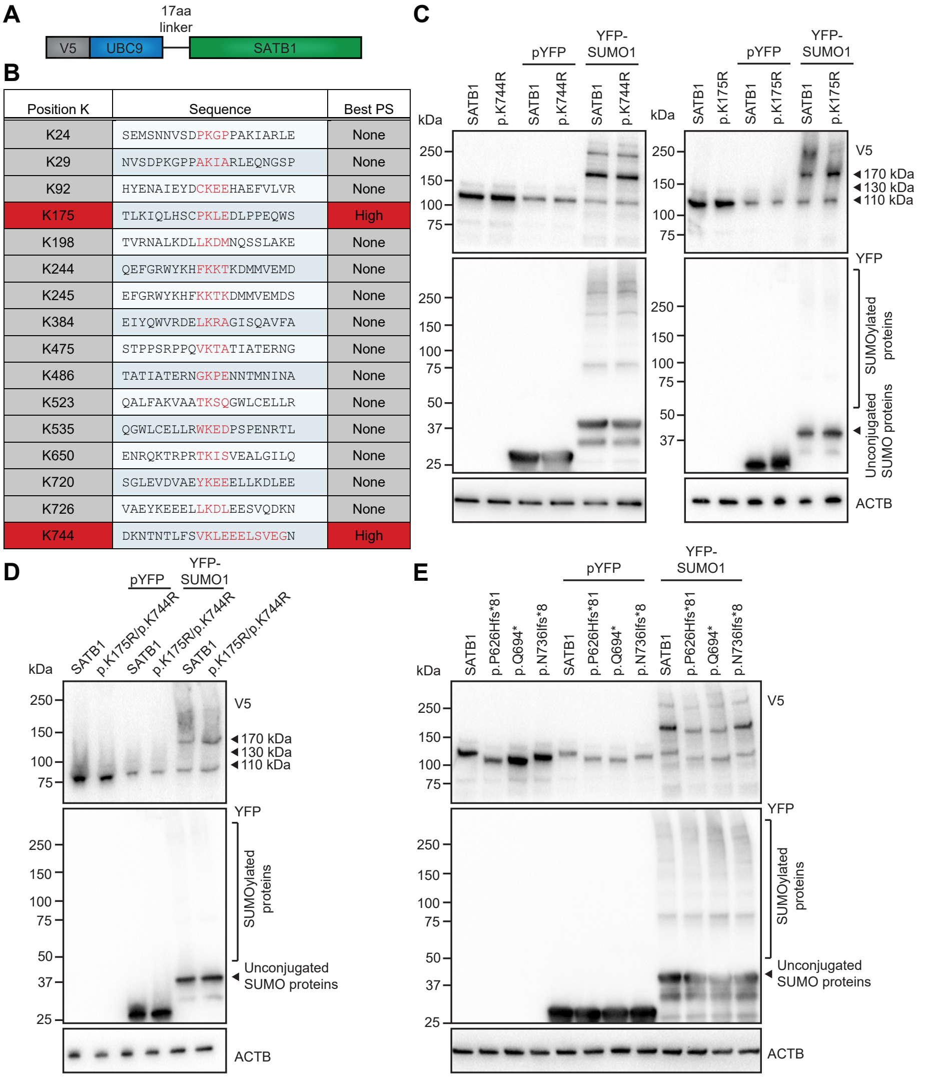
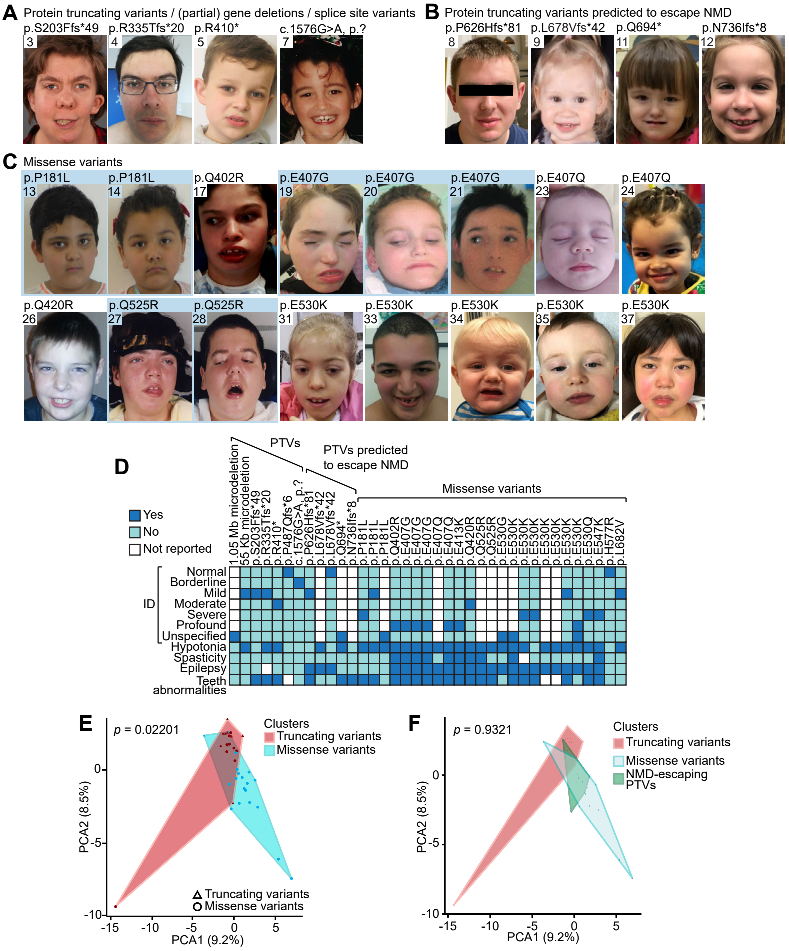


Figure S9. Overexpression of SATB1 NMD-escaping PTVs as YFP-fusion proteins. A) Immunoblot of whole-cell lysates expressing YFP-tagged SATB1 variants probed with anti-EGFP antibody. Expected molecular weight: WT SATB1 = ~115 kDa, p.P626Hfs*81 = ~109 kDa, p.Q694* = ~107 kDa, p.N736Ifs*8 = ~113 kDa. The blot was probed for ACTB to ensure equal protein loading. **B**) Direct fluorescence imaging of HEK293T/17 cells expressing YFP-SATB1 fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 µm. **C**) Results of assay for protein stability of SATB1 NMD-escaping PTVs, using cycloheximide (CHX) to arrest protein synthesis, and MG132 to block protein degradation by the 26S proteasome complex. Values represent the mean protein expression levels of YFP-tagged SATB1 variants ± S.E.M. in live cells as measured by YFP fluorescence and expressed relative to the 0 h time point (n = 3, two-way ANOVA for repeated measures with Geisser-Greenhouse correction, followed by a *post-hoc* Bonferroni test). Although p.P626Hfs*81 showed a slight but significant decrease in relative expression level after treatment with CHX, and p.Q694* showed a significant increase in relative expression level after treatment with MG132 when compared to WT SATB1, none of the variants tested showed both a decrease in levels after CHX treatment and an increase after MG132 treatment, which would be indicative of reduced protein stability.



Position K	Sequence	Best PS
K24	SEMSNNVSDPKGPPAKIARLE	None
K29	NVSDPKGPPAKIARLEQNGSP	None
K92	HYENAIEYDCKEEHAEFVLVR	None
K175	TLKIQLHSCPKLEDLPPEQWS	High
K198	TVRNALKDLLKDMNQSSLAKE	None
K244	QEFGRWYKHFKKTKDMMVEMD	None
K245	EFGRWYKHFKKTKDMMVEMDS	None
K384	EIYQWVRDELKRAGISQAVFA	None
K475	STPPSRPPQVKTATIATERNG	None
K486	TATIATERNGKPENNTMNINA	None
K523	QALFAKVAATKSQGWLCELLR	None
K535	QGWLCELLRWKEDPSPENRTL	None
K650	ENRQKTRPRTKISVEALGILQ	None
K720	SGLEVDVAEYKEEELLKDLEE	None
K726	VAEYKEEELLKDLEESVQDKN	None
K744	DKNTNTLFSVKLEEELSVEGN	High

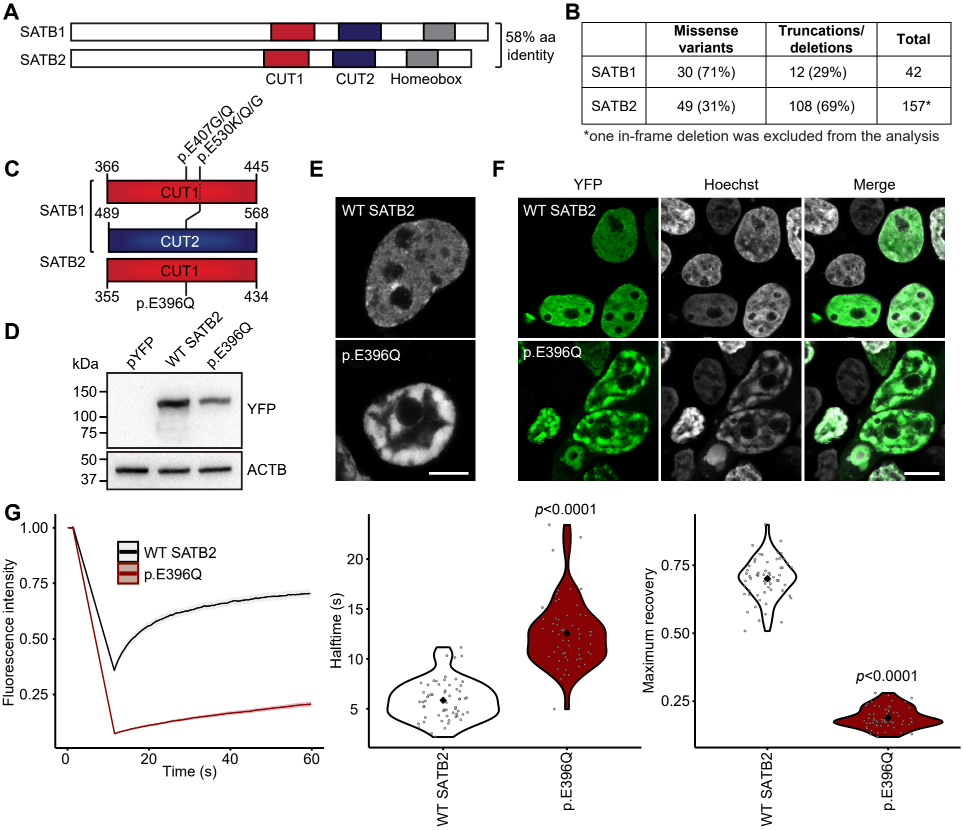
Figure S10. SUMOylation of SATB1 protein truncating variants escaping NMD. A) Schematic representation of the UBC9-SATB1 fusion protein with an N-terminal V5 epitope tag. **B**) Prediction of putative SATB1 (Uniprot Q01826) SUMOylation sites using Joined Advanced SUMOylation Site and SIM Analyser (JASSA, www.jassa.fr/). JASSA uses a scoring system based on a Position Frequency Matrix derived from the alignment of experimental SUMOylation sites. K175 corresponds to a direct consensus site ([Ψ]-[K]-[x]-[α], with Ψ = A,F,I,L,M,P,V or W; α = D or E) with a high prediction score (PS), and K744 to a negatively charged amino acid-dependent SUMOylation site (NDSM, $[\Psi]\text{-}[K]\text{-}[x]\text{-}[\alpha]\text{-}[x]\text{-}[\alpha]_6$ with Ψ = A,F,I,L,M,P,V or W; 2 out of 6 α must be D or E) with a high PS. **C**) Gel shift assay for SATB1 SUMOylation. UBC9-SATB1 and a p.K175R or p.K744R mutant were expressed in HEK293T/17 cells together with a YFP-fusion of SUMO1. Top panel: western blot probed with anti-V5 antibody to detect UBC9-SATB1. The 110 kDa species is unmodified UBC9-SATB1. The 130 kDa species is UBC9-SATB1 modified with endogenous SUMO1. The 170 kDa species is UBC9-SATB1 modified with YFP-SUMO1. Middle panel: western blot probed with anti-YFP antibody, with unconjugated YFP-SUMO1 indicated with an arrow head. Higher molecular weight species are cellular proteins modified with YFP-SUMO1. Bottom panel: western blot probed with anti-ACTB to confirm equal protein loading. **D**) Gel-shift assay for SUMOylation of a SATB1 p.K175R/p.K744R double-mutant. **E**) Gel-shift assay for SUMOylation of SATB1 NMD escaping protein truncating variants.



■ Yes
▨ No
☐ Not reported

		PTVs							PTVs predicted to escape NMD					Missense variants																								
		1.05 Mb microdeletion	55 Kb microdeletion	p.S203Ffs*49	p.R335Tfs*20	p.R410*	p.P487Qfs*6	c.1576G>A, p.?	p.P626Hfs*81	p.L678Vfs*42	p.L678Vfs*42	p.Q694*	p.N736Ifs*8	p.P181L	p.P181L	p.P181L	p.Q402R	p.E407G	p.E407G	p.E407G	p.E407Q	p.E407Q	p.E413K	p.Q420R	p.Q525R	p.Q525R	p.E530G	p.E530K	p.E530K	p.E530K	p.E530K	p.E530K	p.E530K	p.E530K	p.E530Q	p.E547K	p.H577R	p.L682V
ID	Normal	☐	▨	▨	▨	▨	■	▨	▨	☐	■	☐	☐	▨	▨	☐	▨	▨	▨	▨	☐	▨	▨	▨	☐	☐	☐	☐	▨	▨	☐	☐	▨	▨	▨	▨	■	▨
	Borderline	☐	▨	▨	▨	▨	▨	■	▨	☐	▨	☐	☐	▨	▨	☐	▨	▨	▨	▨	☐	▨	▨	▨	☐	☐	☐	☐	▨	▨	☐	☐	▨	▨	▨	▨	▨	▨
	Mild	☐	■	■	■	▨	▨	▨	■	☐	▨	☐	☐	▨	■	☐	▨	▨	▨	▨	☐	▨	▨	▨	☐	☐	☐	☐	▨	▨	☐	☐	■	▨	▨	▨	▨	■
	Moderate	☐	▨	▨	▨	■	▨	▨	▨	☐	▨	☐	☐	▨	▨	☐	▨	▨	▨	▨	☐	▨	▨	■	☐	☐	☐	☐	▨	▨	☐	☐	▨	▨	▨	▨	▨	▨
	Severe	☐	▨	▨	▨	▨	▨	▨	▨	☐	▨	☐	☐	■	▨	☐	▨	▨	▨	▨	☐	▨	▨	▨	☐	☐	☐	☐	■	■	☐	☐	▨	▨	■	■	▨	▨
	Profound	☐	▨	▨	▨	▨	▨	▨	▨	☐	▨	☐	☐	▨	▨	☐	■	■	■	■	☐	■	■	▨	☐	☐	☐	☐	▨	▨	☐	☐	▨	■	▨	▨	▨	▨
	Unspecified	■	▨	▨	▨	▨	▨	▨	▨	☐	▨	■	☐	▨	▨	■	▨	▨	▨	▨	☐	▨	▨	▨	☐	☐	■	■	▨	▨	☐	☐	▨	■	▨	▨	▨	▨
	Hypotonia	▨	■	▨	■	■	▨	▨	▨	■	▨	■	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	■	■	▨	■
	Spasticity	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	■	■	■	▨	■	■	■	■	▨	▨	■	☐	▨	▨	▨	▨	▨	▨	■	▨	▨
	Epilepsy	▨	▨	▨	☐	▨	▨	▨	■	■	■	▨	▨	▨	▨	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	▨
	Teeth abnormalities	▨	▨	■	■	■	☐	▨	■	▨	▨	■	■	■	■	▨	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	☐	☐	■	▨	■	■	▨	▨

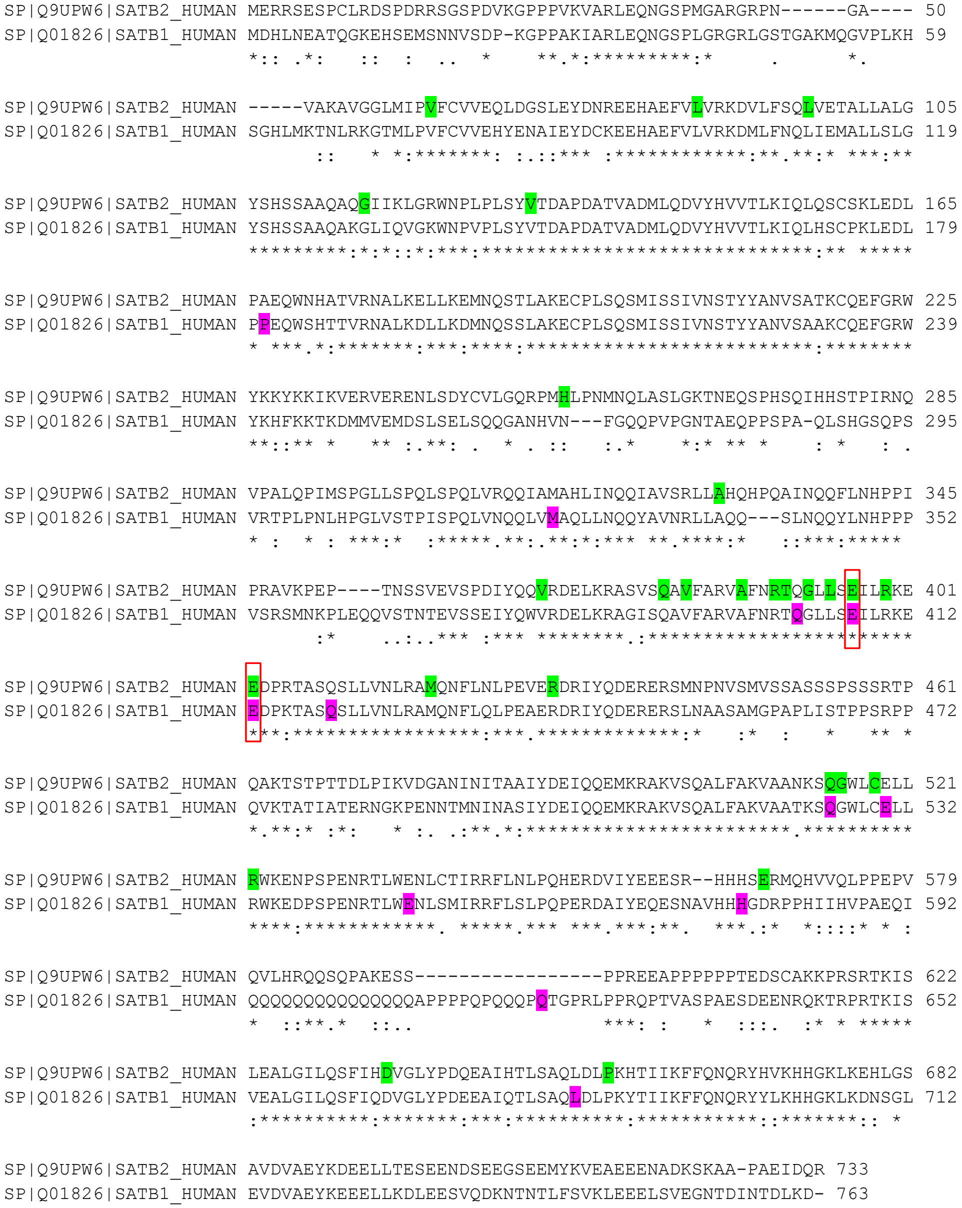
Figure S11. Clinical evaluation of individuals with SATB1 variants in three subcohorts. A-C) Facial photographs of individuals with (partial) gene deletions and truncations predicted to result in haploinsufficiency (A), of individuals with truncations predicted to escape from NMD and resulting in transcriptionally active proteins (B) and of individuals with missense variants (C). All depicted individuals show facial dysmorphisms and although overlapping features are seen, no consistent facial phenotype can be observed for the group as a whole. Overlapping facial dysmorphisms include facial asymmetry, high forehead, prominent ears, straight and/or full eyebrows, puffy eyelids, downslant of palpebral fissures, low nasal bridge, full nasal tip and full nasal alae, full lips with absent cupid's bow, prominent cupid's bow or thin upper lip vermilion. Individuals with missense variants are more alike than individuals in the truncating cohorts, and we observed recognizable overlap between several individuals in the missense cohort (individuals 17, 27, 31, 37, the siblings 19, 20 and 21, and to a lesser extent individuals 24 and 35). A recognizable facial overlap between individuals with the other two variant types could not be observed. Related individuals are marked with a blue box. **D**) Mosaic plot presenting a selection of clinical features. Individuals with no or very limited clinical data were omitted (for details, see Supplemental Materials and Methods). **E**) The Partitioning Around Medoids analysis of clustered HPO-standardised clinical data from 38 individuals with truncating (triangle) and missense variants (circle) shows a significant distinction between the clusters of individuals with missense variants (blue) and individuals with PTVs (red). Applying Bonferroni correction, a *p*-value smaller than 0.025 was considered significant. **F**) Plot of Partitioning Around Medoids clustering analysis on clustered clinical data (HPO) showing no significant distinctions between individuals with missense variants, individuals with truncating variants and deletions, and individuals with NMD-escaping truncating variants.



	Missense variants	Truncations/ deletions	Total
SATB1	30 (71%)	12 (29%)	42
SATB2	49 (31%)	108 (69%)	157*

*one in-frame deletion was excluded from the analysis

Figure S12. The SATB2 p.E396Q missense variant has comparable effects on protein functions as the p.E407G and p.E530K/Q SATB1 variants affecting equivalent positions. A) SATB1 and SATB2 are highly conserved paralogues. **B**) In SATB1 more missense variants (71%) than truncations/deletions (29%) are observed, while for SATB2 the reverse is reported (31% versus 69% respectively). **C**) Schematic representation of SATB1 and SATB2 CUT DNA binding domains, with variants on equivalent positions indicated. **D**) Immunoblot of whole-cell lysates expressing YFP-tagged SATB2 and p.E396Q probed with anti-EGFP antibody. Expected molecular weight is ~112 kDa. The blot was probed for ACTB to ensure equal protein loading. **E**) Direct fluorescence super-resolution imaging of nuclei of HEK293T/17 cells expressing YFP-SATB2 fusion proteins. Scale bar = 5 µm. **F**) Direct fluorescence imaging of HEK293T/17 cells expressing YFP-SATB2 fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 µm. **G**) FRAP experiments to assess the dynamics of SATB2 chromatin binding in live cells. Left, mean recovery curves ± 95% C.I. recorded in HEK293T/17 cells expressing YFP-SATB2 fusion proteins. Right, violin plots with median of the halftime and maximum recovery values based on single-term exponential curve fitting of individual recordings (*n* = 60 nuclei from three independent experiments, *p*-values compared to WT SATB2, unpaired t-test).



```
SP|Q9UPW6|SATB2_HUMAN MERRSESPCLRDSPDRRSGSPDVKGPPPVKVARLEQNGSPMGARGRPN------GA---- 50
SP|Q01826|SATB1_HUMAN MDHLNEATQGKEHSEMSNNVSDP-KGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKH 59
                      *:: .*:   ::  :  ..  *    **.*:*********:*     .      *.

SP|Q9UPW6|SATB2_HUMAN -----VAKAVGGLMIPVFCVVEQLDGSLEYDNREEHAEFVLVRKDVLFSQLVETALLALG 105
SP|Q01826|SATB1_HUMAN SGHLMKTNLRKGTMLPVFCVVEHYENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLG 119
                            ::   * *:*******: :.::*** :************:**.**:* ***:**

SP|Q9UPW6|SATB2_HUMAN YSHSSAAQAQGIIKLGRWNPLPLSYVTDAPDATVADMLQDVYHVVTLKIQLQSCSKLEDL 165
SP|Q01826|SATB1_HUMAN YSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLKIQLHSCPKLEDL 179
                      *********:*:*::*:***:******************************:** *****

SP|Q9UPW6|SATB2_HUMAN PAEQWNHATVRNALKELLKEMNQSTLAKECPLSQSMISSIVNSTYYANVSATKCQEFGRW 225
SP|Q01826|SATB1_HUMAN PPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRW 239
                      * ***.*:*******:***:****:**************************:********

SP|Q9UPW6|SATB2_HUMAN YKKYKKIKVERVERENLSDYCVLGQRPMHLPNMNQLASLGKTNEQSPHSQIHHSTPIRNQ 285
SP|Q01826|SATB1_HUMAN YKHFKKTKDMMVEMDSLSELSQQGANHVN---FGQQPVPGNTAEQPPSPA-QLSHGSQPS 295
                      **::** *   ** :.**: .  * . ::   :.*    *:* ** *    : *   : .

SP|Q9UPW6|SATB2_HUMAN VPALQPIMSPGLLSPQLSPQLVRQQIAMAHLINQQIAVSRLLAHQHPQAINQQFLNHPPI 345
SP|Q01826|SATB1_HUMAN VRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVNRLLAQQ---SLNQQYLNHPPP 352
                      * :  * : ***:*  :*****.**:.**:*:*** **.****:*   ::***:*****

SP|Q9UPW6|SATB2_HUMAN PRAVKPEP----TNSSVEVSPDIYQQVRDELKRASVSQAVFARVAFNRTQGLLSEILRKE 401
SP|Q01826|SATB1_HUMAN VSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKE 412
                            :*    ..:..*** :*** ********.:************************

SP|Q9UPW6|SATB2_HUMAN EDPRTASQSLLVNLRAMQNFLNLPEVERDRIYQDERERSMNPNVSMVSSASSSPSSSRTP 461
SP|Q01826|SATB1_HUMAN EDPKTASQSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPP 472
                      ***:*****************:***.*************:*   :*  :   *   ** *

SP|Q9UPW6|SATB2_HUMAN QAKTSTPTTDLPIKVDGANINITAAIYDEIQQEMKRAKVSQALFAKVAANKSQGWLCELL 521
SP|Q01826|SATB1_HUMAN QVKTATIATERNGKPENNTMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELL 532
                      *.**:* :*:   * :. .:**.*:************************.**********

SP|Q9UPW6|SATB2_HUMAN RWKENPSPENRTLWENLCTIRRFLNLPQHERDVIYEEESR--HHHSERMQHVVQLPPEPV 579
SP|Q01826|SATB1_HUMAN RWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPAEQI 592
                      ****:************. *****.*** ***.***:**.  ***.:*  *::::* * :

SP|Q9UPW6|SATB2_HUMAN QVLHRQQSQPAKESS-----------------PPREEAPPPPPPTEDSCAKKPRSRTKIS 622
SP|Q01826|SATB1_HUMAN QQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKIS 652
                      *  ::**.*  ::..                 ***: :   *  :::.  :* * *****

SP|Q9UPW6|SATB2_HUMAN LEALGILQSFIHDVGLYPDQEAIHTLSAQLDLPKHTIIKFFQNQRYHVKHHGKLKEHLGS 682
SP|Q01826|SATB1_HUMAN VEALGILQSFIQDVGLYPDEEAIQTLSAQLDLPKYTIIKFFQNQRYYLKHHGKLKDNSGL 712
                      :**********:*******:***:**********:***********::*******:: *

SP|Q9UPW6|SATB2_HUMAN AVDVAEYKDEELLTESEENDSEEGSEEMYKVEAEEENADKSKAA-PAEIDQR 733
SP|Q01826|SATB1_HUMAN EVDVAEYKEEELLKDLEESVQDKNTNTLFSVKLEEELSVEGNTDINTDLKD- 763
```

Figure S13. Missense variants identified in individuals with NDD displayed in an amino acid sequence alignment of SATB2 and SATB1. SATB2 (Q9UPW6, UniProt) sequence is aligned to SATB1 (Q01826) sequence. Alignment was performed with Clustal Omega (1.2.4) with default settings using UniProt alignment tool. Previously reported missense variants in SATB2 (PMID: 31021519) are shaded in green, SATB1 missense variants (this study) are shaded in magenta. Only two missense variants occur at equivalent positions (marked with a red box): SATB2 p.E396Q is equivalent to SATB1 p.E407G/Q, and SATB2 p.E402K is equivalent to SATB1 p.E413K. We functionally characterised SATB2 p.E396Q (Figure S12).

Table S1. Splice-AI predictions for missense variants at intron-exon or exon-intron junctions.

g.DNA-position	c.DNA	Protein effect	spliceAI-G delta score§ - acceptor gain (position*)	spliceAI-G delta score§ - acceptor loss (position*)	spliceAI-G delta score§ -donor gain (position*)	spliceAI-G delta score§ -donor loss (position*)
Chr3:g.18435955T>C	c.1205A>G	p.Q402R¥	0 (-1)	0 (45)	0.0099 (32)	0.2482 (-1)
Chr3:g.18419663T>C	c.1574A>G	p.Q525R£	0 (-1)	0 (19)	0 (20)	0 (-1)
Chr3:g.18393687C>T	c.1576G>A	p.G526R#	0.6666 (-2)	0.0937 (0)	0 (-2)	0 (-17)

*a negative nucleotide position represents positions upstream of the variant, a positive nucleotide position represents positions downstream of the variant.
§cut offs for splice-AI delta score: 0.2 (high recall), 0.5 (recommended), and 0.8 (high precision)

¥p.Q402R:

- Although the variant affects the last amino acid of exon 7, none of the Splice-AI delta scores exceeds the recommended cut-off of >0.5, specifically not the scores for loss or gain of splice donor sites.

£p.Q525R

- Although the variant affects the last amino acid of exon 9, none of the Splice-AI delta scores exceeds the recommended cut-off of >0.5, specifically not the scores for loss or gain of splice donor sites.

#p.G526R:

- The variant affects the first amino acid of exon 10. Splice-AI predicts splice acceptor site gain 2 nucleotides upstream of the variant, resulting in a frameshift.

Table S2. Phenotypic information of individuals from the UK10K cohort with rare *SATB1* missense variants. Wechsler Intelligence Scale for Children (WISC) test scores for individuals from the UK10K cohort, carrying rare *SATB1* missense variants. Standard deviation scores (std score) were calculated by comparing individual scores of carriers to the mean test scores from UK10K non-carriers. Test scores that were lower compared to mean non-carrier scores are shaded in red, while test score that were higher compared to mean non-carrier scores are shaded in green. All carrier test scores were within 2.5 standard deviations compared to the mean non-carrier scores, and thus within normal range.

	UK10K non-carriers (n=1732, ±Std)	UK10K carriers (n=9, ±Std)	UK10K carriers								
Variant	-	-	rs148337599	rs148337599	rs148337599	rs148337599	rs148337599	rs760272331	rs760272331	rs185604711	rs185604711
Residue change (SATB1 NM_001131010.4)	-	-	p.S366L	p.S366L	p.S366L	p.S366L	p.S366L	p.V519L	p.V519L	p.A573T	p.A573T
gnomAD v2.1.1 frequency			6.61e-4 (allele 282848)					8.67e-6 (allele 230660)		1.17e-4 (allele 282890)	
WISC - Verbal IQ: F@8	111.92 (±16.57)	116.89 (±16.07)	103	133	139	103	111	133	128	99	103
VIQ std score	-	0.30 (±0.97)	-0.54	1.27	1.63	-0.54	-0.06	1.27	0.97	-0.78	-0.54
WISC - Performance IQ: F@8	103.49 (±16.79)	109.00 (±15.33)	104	125	115	119	109	115	90	80	124
PIQ std score	-	0.33 (±0.91)	0.03	1.28	0.69	0.92	0.33	0.69	-0.80	-1.40	1.22
WISC - Total IQ: F@8	109.12 (±16.00)	114.89 (±14.73)	104	133	132	111	111	130	111	88	114
IQ std score	-	0.36 (±0.92)	-0.32	1.49	1.43	0.12	0.12	1.30	0.12	-1.32	0.31
WISC - Verbal Comprehension Index: F@8	48.47 (±11.10)	51.79 (±12.51)	38	63	72	44	47	58	63	37	44
VCI std score	-	0.30 (±1.13)	-0.94	1.31	2.12	-0.40	-0.13	0.86	1.31	-1.03	-0.40
WISC - Perceptual Organisation Index: F@8	41.51 (±10.50)	43.80 (±8.37)	41	49	50	53	44	50	30	31	47
POI std score	-	0.23 (±0.80)	-0.05	0.71	0.81	1.09	0.24	0.81	-1.10	-1.00	0.52
WISC - Freedom from Distractability Index: F@8	22.17 (±5.96)	24.11 (±5.42)	27	26	22	22	28	35	19	20	18
FDI std score	-	0.33 (±0.91)	0.81	0.64	-0.03	-0.03	0.98	2.15	-0.53	-0.36	-0.70

🟥	≦-1.0 Std score compared to UK10K non-carriers
▨	≦-0.0 Std score compared to UK10K non-carriers
🟩	≧1.0 Std score compared to UK10K non-carriers
▧	≧0.0 Std score compared to UK10K non-carriers

Table S3. NMD efficacy predictions for SATB1 truncating variants.

(Hg19/GRCh37)g.DNA-position	g.DNA-position of introduced (downstream) stopcodon	c.DNA-position (NM_001131010.4)	Protein effect**	NMDetective A¥ (1)	NMDetective A¥ (2)	NMDetective B¥ (1)	NMDetective B¥ (2)	Conclusion based on predictions with NMDetectiveA/B	Prediction based on canonical# and non-canonical§ NMD rules
Chr3:g.18456634_18456635delCT	Chr3:g.18436407	c.607_608delAG	p.S203Ffs*49	0.63	0.45	0.65	0.41	Conflicting; NMDetectiveA/B (1): triggers NMD, NMDetectiveA/B (2): intermediate NMD efficacy.	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18436155_18436156delTC	Chr3:g.18436098	c.1004_1005delGA	p.R335Tfs*20	0.51	0.52	0.41	0.41	Intermediate NMD efficacy	Might escape from NMD. None of canonical NMD rules applicable, non-canonical long-exon rule applicable (exon 7; 454 nucleotides).
Chr3:g.18428082G>A	Chr3:g.18428082	c.1228C>T	p.R410*	0.6		0.65		Triggers NMD	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18419777delG	Chr3:g.18419762	c.1460delC	p.P487Qfs*6	0.62	0.62	0.65	0.65	Triggers NMD	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18393687C>T	Chr3:g.18393611	c.1576G>A	p.(?)	0.57	0.6	0.65	0.65	Triggers NMD	Triggers NMD, none of (non)-canonical NMD rules applicable
Chr3:g.18391077delG	Chr3:g.18390837	c.1877delC	p.P626Hfs*81	0.08	0.26	0	0	Conflicting; NMDetectiveA/B (1) and NMDetectiveB (2): escapes NMD; NMDetectiveA (2): intermediate NMD efficacy,	Escapes NMD based on canonical last exon rule
Chr3:g.18390921_18390922delCA	Chr3:g.18390797	c.2032_2033delCT	p.L678Vfs*42	0.18	0.17	0	0	Escapes NMD	Escapes NMD based on canonical last exon rule
Chr3:g.18390874G>A	Chr3:g.18390874	c.2080C>T	p.Q694*	0.2		0		Escapes NMD	Escapes NMD based on canonical last exon rule
Chr3:g.18390747delT	Chr3:g.18390726	c.2207delA	p.N736Ifs*8	0.16	0.16	0	0	Escapes NMD	Escapes NMD based on canonical last exon rule

**For frameshift mutations, scores for NMDetectiveA and NMDetectiveB were assigned both based on the genomic location of the indel (1) and based on the genomic location of the first downstream stopcodon in the new reading frame (2; first nucleotide of introduced stopcodon) (PMID: 31659324). For splice site mutations, NMDetectiveA and NMDetectiveB were assigned based on the effect predicted by spliceAI (PMID: 30661751).

¥NMDetectiveA and NMDetectiveB cut-off scores (v2):
<0.25 predicted to escape NMD
≥0.25 - ≤0.52 predicted intermediate NDM efficacy
>0.52 predicted to trigger NMD (PMID: 31659324)

#Canonical rules of NMD (PMID: 27618451):
NMD is typically not triggered when the location of the protein truncating variant is
1. less than 50 nucleotides upstream of last exon-exon junction; or
2. in the last exon.

§Non-canonical rules of NMD (PMID: 27618451):
NMD is not triggered when the location of the protein truncating variant is
1. in a very long exon (> ±400 nucleotides); or
2. within 150 nucleotides from the start codon.

$ - predicted amino acid sequences of NMD-escaping truncating variants in SATB1

Amino acid sequence of SATB1 (NM_002971.4/NM_001131010.4) in the normal situation
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTLSAQLDLPKYTIIKFFQNQRYYLKHHGKLKDNSGLEVDVAEYKEEELLKDLEESVQDKNTNTLFSVKLEEELSVEGNTDINT
DLKD

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 5
*Chr3:g.18428082G>A; c.1228C>T; p.R410**
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEIL*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 8
*Chr3:g.18391077del; c.1877del; p.P626Hfs*81*
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPHGNPRWPLQQSQMRKTDRRPGHEQKFQWKPWESSRVSYKTWA
CTLTKRPSRLCLPSSTFPSTPSSSSFRTSGTISSTTAN*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 9 and 10 *Chr3:g.18390921_18390922del; c.2032_2033del; p.L678Vfs*42*
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTVCPARPSQVHHHQVLSEPAVLSQAPRQTEGQFRFRGRCGRI*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 11
*Chr3:g. 18390874G>A; c.2080C>T; p.Q694**
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTLSAQLDLPKYTIIKFF*

Aminoacid sequence of SATB1 (NM_002971.4/NM_001131010.4) in patient 12
*Chr3:g.18390747del; c.2207del; p.N736Ifs*8*
MDHLNEATQGKEHSEMSNNVSDPKGPPAKIARLEQNGSPLGRGRLGSTGAKMQGVPLKHSGHLMKTNLRKGTMLPVFCVVEH
YENAIEYDCKEEHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQDVYHVVTLK
IQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSLAKECPLSQSMISSIVNSTYYANVSAAKCQEFGRWYKHFKKTKDMMV
EMDSLSELSQQGANHVNFGQQPVPGNTAEQPPSPAQLSHGSQPSVRTPLPNLHPGLVSTPISPQLVNQQLVMAQLLNQQYAVN
RLLAQQSLNQQYLNHPPPVSRSMNKPLEQQVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTAS
QSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAASAMGPAPLISTPPSRPPQVKTATIATERNGKPENNTMNINASIYDEIQQE
MKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGDRPPHIIHVPA
EQIQQQQQQQQQQQQQQQAPPPPQPQQQPQTGPRLPPRQPTVASPAESDEENRQKTRPRTKISVEALGILQSFIQDVGLYPDE
EAIQTLSAQLDLPKYTIIKFFQNQRYYLKHHGKLKDNSGLEVDVAEYKEEELLKDLEESVQDKILTPFFQ*

Table S4. Summary of clinical characteristics associated with (*de novo*) *SATB1* PTVs and (partial) gene deletions predicted to result in haploinsufficiency and PTVs in the last exon.

	Individuals with PTVs and (partial) gene deletions predicted to result in haploinsufficiency		Individuals with PTVs in the last exon	
	%	**Present / total assessed**	**%**	**Present / total assessed**
Neurologic				
Intellectual disability	86	6/7	67	2/3
Normal	14	1/7	33	1/3
Borderline	0	0/7	0	0/3
Mild	71	5/7	33	1/3
Moderate	14	1/7	0	0/3
Severe	0	0/7	0	0/3
Profound	0	0/7	0	0/3
Unspecified	0	0/7	33	1/3
Developmental delay	100	7/7	100	5/5
Motor delay	86	6/7	100	5/5
Speech delay	86	6/7	80	4/5
Dysarthria	14	1/7	0	0/4
Epilepsy	0	0/6	40	2/5
EEG abnormalities	0	0/4	67	2/3
Hypotonia	43	3/7	40	2/5
Spasticity	0	0/7	0	0/5
Ataxia	14	1/7	20	1/5
Behavioural disturbances	100	7/7	0	0/5
Sleep disturbances	50	3/6	0	0/5
Abnormal brain imaging	33	1/3	50	2/4
Regression	14	1/7	0	0/5
Growth				
Abnormalities during pregnancy	33	2/6	20	1/5
Abnormalities during delivery	33	2/6	80	4/5
Abnormal term of delivery	0	0/5	20	1/5
Preterm (<37 weeks)	0	0/5	20	1/5
Postterm (>42 weeks)	0	0/5	0	0/5
Abnormal weight at birth	20	1/5	25	1/4
Small for gestational age (<p10)	20	1/5	0	0/4
Large for gestational age (>p90)	0	0/5	25	1/4
Abnormal head circumference at birth	25	1/4	0	0/2
Microcephaly (<p3)	0	0/4	0	0/2
Macrocephaly (>p97)	25	1/4	0	0/2
Abnormal height	14	1/7	0	0/4
Short stature (<p3)	0	0/7	0	0/4
Tall stature (>p97)	14	1/7	0	0/4
Abnormal head circumference	0	0/5	25	1/4
Microcephaly (<p3)	0	0/5	25	1/4
Macrocephaly (>p97)	0	0/5	0	0/4
Abnormal weight	0	0/5	25	1/4
Underweight (<p3)	0	0/5	25	1/4
Overweight (>p97)	0	0/5	0	0/4
Other phenotypic features				
Facial dysmorphisms	67	4/6	60	3/5
Dental/oral abnormalities	50	3/6	60	3/5
Drooling/dysphagia	29	2/7	20	1/5
Hearing abnormalities	17	1/6	20	1/5
Vision abnormalities	67	4/6	80	4/5
Cardiac abnormalities	17	1/6	40	2/5
Skeleton/limb abnormalities	33	2/6	0	0/5
Hypermobility of joints	33	2/6	25	1/4
Gastrointestinal abnormalities	33	2/6	20	1/5
Urogenital abnormalities	0	0/6	0	0/5
Endocrine/metabolic abnormalities	0	0/6	0	0/5
Immunological abnormalities	17	1/6	50	1/2
Skin/hair/nail abnormalities	0	0/6	20	1/5
Neoplasms in medical history	0	0/6	0	0/5

Table S5. Primers for site-directed mutagenesis

SATB1-K175R-F	GGAGGCAAGTCTTCTAGTCGGGGGCAACTGTGTAACTG
SATB1-K175R-R	CAGTTACACAGTTGCCCCCGACTAGAAGACTTGCCTCC
SATB1-S366L-F	TCTGTGTTGGTCAAAACCTGTTGCTCCAAAGGCT
SATB1-S366L-R	AGCCTTTGGAGCAACAGGTTTTGACCAACACAGA
SATB1-E407G-F	CTTCCTTTCGGAGGATTCCTGAAAGCAAGCCCTGA
SATB1-E407G-R	TCAGGGCTTGCTTTCAGGAATCCTCCGAAAGGAAG
SATB1-R410*	GGGGTCCTCTTCCTTTCAGAGGATTTCTGAAAGCA
SATB1-R410*	TGCTTTCAGAAATCCTCTGAAAGGAAGAGGACCCC
SATB1-Q420R-F	GTTTACCAGCAAAGACCGGGATGCAGTCTTGGG
SATB1-Q420R-R	CCCAAGACTGCATCCCGGTCTTTGCTGGTAAAC
SATB1-E530K-F	TCCAGCGTAACAGCTTGCACAACCATCCCTG
SATB1-E530K-R	CAGGGATGGTTGTGCAAGCTGTTACGCTGGA
SATB1-E530Q-F	CCAGCGTAACAGCTGGCACAACCATCCCT
SATB1-E530Q-R	AGGGATGGTTGTGCCAGCTGTTACGCTGG
SATB1-E547K-F	GATCATGGAGAGGTTCTTCCACAGGGTTCTGTTTT
SATB1-E547K-R	AAAACAGAACCCTGTGGAAGAACCTCTCCATGATC
SATB1-V519L-F	GCTTTTGGTTGCTGCAAGCTTTGCAAACAGTGCTT
SATB1-V519L-R	AAGCACTGTTTGCAAAGCTTGCAGCAACCAAAAGC
SATB1-A573T-F	CATGGTGATGCACCGTGTTGCTCTCCTGTTC
SATB1-A573T-R	GAACAGGAGAGCAACACGGTGCATCACCATG
SATB1-P626Hfs*81-F	GTGGGTTGCCGTGGGGGAGCCGAG
SATB1-P626Hfs*81-R	CTCGGCTCCCCCACGGCAACCCAC
SATB1-L682V-F	CTTGGGAAGGTCGACCTGGGCAGACAGAG
SATB1-L682V-R	CTCTGTCTGCCCAGGTCGACCTTCCCAAG
SATB1-Q694*-F	TACCGCTGGTTCTAAAAGAACTTGATGATGGTGTACTTG
SATB1-Q694*-R	CAAGTACACCATCATCAAGTTCTTTTAGAACCAGCGGTA
SATB1-N736I*8-F	AAAAAGGGTGTTAGTATTTTATCTTGGACACTCTCTTCCAAATCCT
SATB1-N736I*8-R	AGGATTTGGAAGAGAGTGTCCAAGATAAAATACTAACACCCTTTTT
SATB1-K744R-F	CACTGACAGCTCTTCTTCTAGTCGCACTGAAAAAAGGGTGTTAGTA
SATB1-K744R-R	TACTAACACCCTTTTTTCAGTGCGACTAGAAGAAGAGCTGTCAGTG
SATB2-E396Q-F	TACGCAGAATCTGAGACAACAATCCCTGTGTGCGG
SATB2-E396Q-R	CCGCACACAGGGATTGTTGTCTCAGATTCTGCGTA

Table S6. Primers for amplifying and subcloning human UBC9 (NM_194260.2) and SATB1 (NM_001131010.4). Sequences of restriction sites are shown in bold, and sequences that were added to extend the linker region between UBC9 and SATB1 are underscored.

UBC9-*BamHI*-F	GAGGGA**GGATCC**TGCTGTCGGGGATCGCCCTCAG
UBC9-*XmaI*-R	TCTAGA**CCCGGG**CAGCGCAAGTGAGGGCGCAAACTTCTTGG
SATB1-*HindIII*-F	CGGTAC**AAGCTT**TTGGCTGTACTGGATCATTTGAACGAGGC
SATB1-*XhoI*-R	CAGTTA**CTCGAG**TCAGTCTTTCAAATCAGTATTAATGTCTG

Table S7. Primers to amplify regions that include the SATB1 NMD-escaping truncating variants used for testing for NMD. The last exon primer set was used for SATB1 p.P626Hfs*81, p.Q694* and p.N736Ifs*8.

SATB1-NMD-R410*-F	CCTGGGCTCGTATCAACACC
SATB1-NMD-R410*-R	CATCCCTGGCTTTTGGTTGC
SATB1-NMD-last_exon-F	GCCATTTATGAACAGGAGAGCA
SATB1-NMD-last exon-R	CAGTATTAATGTCTGTGTTTCCTTCCA

Supplemental notes

3D protein modelling

Method for modelling CUTL variants

PDB entry 4Q2J[1] was used to contextualise the p.P181L variant. PDB entry 2O49[2] was superposed onto PDB entry 4Q2J using Swiss-PdbViewer[3] to highlight the relative orientation of DNA with respect to the SATB1 CUTL domain.

Method for modelling CUT1 variants

The crystal structure of the N-terminal CUT Domain of SATB1 Bound to Matrix Attachment Region DNA (PDB entry 2O4A[2]), and the ONECUT homeodomain of transcription factor HNF-6[4] were used to contextualise the various mutations with respect to DNA, using Swiss-PdbViewer[3].

Method for modelling CUT2 variants

The first NMR model of the PDB entry 2CSF [DOI:10.2210/pdb2CSF/pdb] was used as a template to align residues T491 to H577 of the SATB1 human protein (uniprot entry Q01826), and build a model using Swiss-PdbViewer[3]. The resulting model has been superposed onto the CUT1 domain of pdb entry 2O4A[2] using the “magic fit” option of Swiss-PdbViewer to highlight the position of the variants with respect to DNA.

Method for modelling homeobox domain variants

The Solution structure of the homeodomain of human SATB2 (second NMR model of the PDB entry 1WI3 [DOI:10.2210/pdb1wi3/pdb]) was used as a template to align residues P647 to G704 of the SATB1 human protein (uniprot entry Q01826), and build a model using Swiss-PdbViewer[3]. Chains A, C and D of the crystal structure of HNF-6alpha DNA-binding domain in complex with the TTR promoter (PDB entry 2D5V[4]), which has a DNA binding domain similar to the CUT2 domain of SATB1 and a second DNA binding domain similar to the homeobox of SATB1, was used as a template to superpose the model of the SATB1 homeobox domain onto the HNF-6alpha structure using the “magic fit” option of Swiss-PdbViewer.

Modelling

p.P181L

The variant P181L variant sits in a linker region between the ubiquitin-like domain (ULD; grey) and a CUT repeat-like (CUTL) domain (dim green). P181 is preceded by another proline, which confers some rigidity and restricts the range of possible relative orientation of the CUTL domain with respect to the UBL domain. There is a third proline in the linker (Pro174), which is preceded by Cys173 and makes a cis peptide bond (highlighted in yellow in Figure 1). Cis-peptide bonds are quite rare (about 0.3% of peptide bonds, although they occur in about 6% of residues followed by a Proline[5], which shows the importance of the conformation of the linker region. Furthermore, Lys175 and Ser185 (in pink) can be respectively acetylated and phosphorylated and influence the DNA binding capability of SATB11. Sidechains of Glu 182 (from the linker bottom left) and Arg 238 (from the CUTL domain bottom right), positioned just below Pro181 further lock the linker region and the CUTL domain through electrostatic interaction. The relative orientation of these domains cannot be maintained with the

P181L mutation, because a leucine sidechain at this position would severely clash into the CUTL domain (backbone of residues Gly237 and Arg238), forcing the linker to adopt a different conformation (Figure 2), which may also potentially affect the ability of K175 to be acetylated.

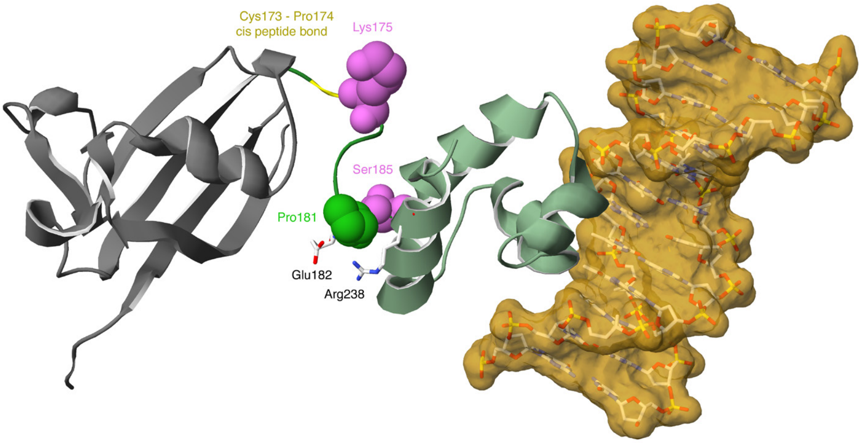


Figure 1. Highlight of the P181 position (green spacefill) with respect to the ubiquitin-like domain (ULD; grey) and the CUT repeat-like (CUTL) domain (dim green). The position of the C173-P174 cis peptide bond is highlighted in yellow. K175 and S185 which can be respectively acetylated and phosphorylated are shown in pink spacefill (top and bottom, respectively).

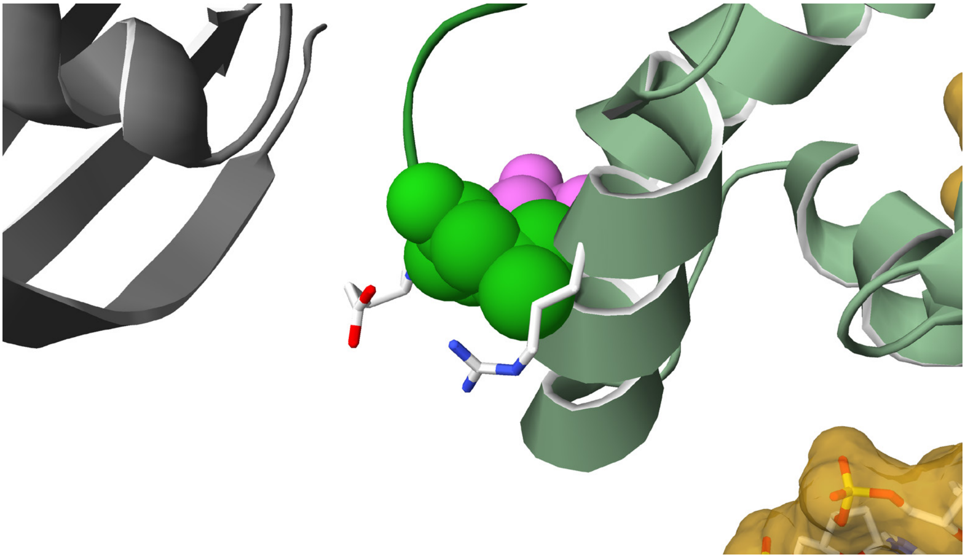


Figure 2. P181L sidechain (green spacefill) clashes into an alpha-helix (A230-K241) of the CUTL domain (dim green), in particular the backbone of residues G237 and R238, as well as in the sidechain of the latter.

p.Q402R

Q402 is located in the CUT1 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT2 domain Q525. Since its sidechain makes direct contact with a nucleotide, a mutation to an arginine, which has a longer sidechain, would need to adopt a conformation less favourable to DNA binding to avoid colliding into the DNA, hence affecting the DNA binding affinity at the cognate sites (Figure 3).

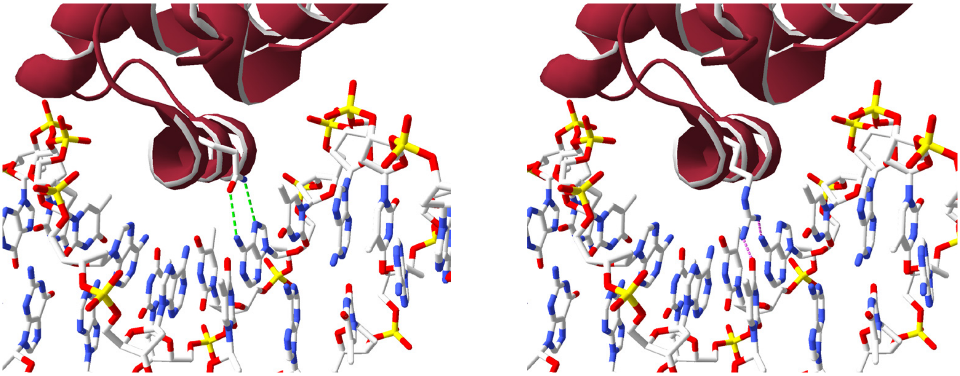


Figure 3. Closeup of the Q402 – DNA interaction (pdb structure 2O4A) highlighting the native residue (Gln, left panel) which makes nice hydrogen bonds to the base (green dotted lines), whereas the longer Arg sidechain (right panel) might collide into the DNA (purple dotted lines) and be forced to adopt a conformation less favourable with respect to binding its cognate DNA.

p.E407G

E407 is located in the middle of the CUT1 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT2 domain E530. Since its sidechain help maintain the sidechain of Arg410 in place via hydrogen bonds and that both residue make direct contact with the nucleotides, a mutation to a glycine, which bears no sidechain and is not favoured in alpha-helices will likely disrupt the local conformation and alter the DNA binding affinity at the cognate sites (Figure 4).

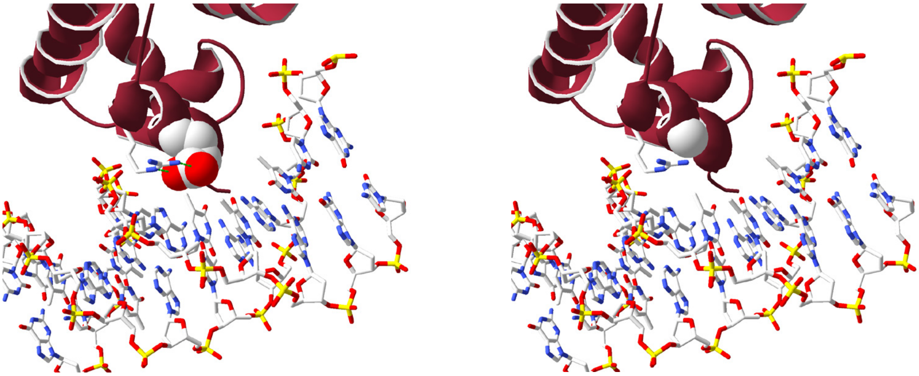


Figure 4. Closeup of the E407 - DNA binding interaction (pdb structure 2O4A) highlighting the native residue (Glu, spacefilled, left panel), which locks in place the sidechain of Arg410 through hydrogen bonds (green dotted lines) and the hole left by the mutation (Gly, spacefilled right panel).

p.E413K

E413 is located in a loop right after the end of the CUT1 domain alpha-helix that binds the major groove of the DNA. Although it does not directly bind to DNA, it is in relatively close proximity (within 10 angstroms) to the negatively charged DNA backbone, and in an extended conformation along Lys411. The mutation E413K would replace a negatively charged residue by a positively charged one and may potentially affect the DNA binding affinity of the CUT1 domain through long range electrostatic interactions (Figure 5).

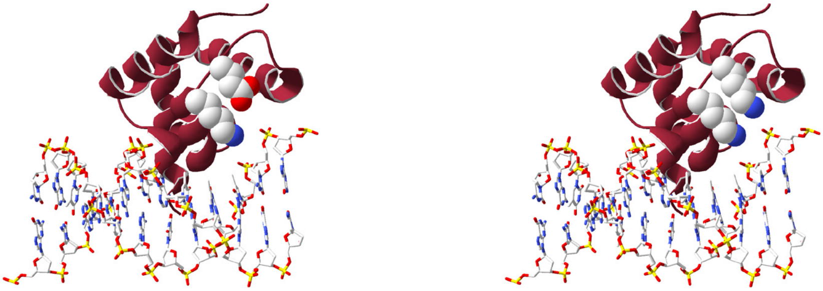


Figure 5. Closeup of E413, solvent exposed in a loop, along Lys411 (left panel). E413 does not make direct DNA contact, and there is enough space to accommodate the E413K mutation (right panel).

p.Q420R

Q420 is located at the surface of the CUT1 domain, not in direct contact with DNA. An arginine at this position could easily be accommodated, but since it is bulkier and positively charged, it may affect the binding of CUT1 to other domains. Of note, the superposition of the CUT1 domain onto the DNA binding domain of rat HNF6 alpha bound to the TTR promoter (pdb entry 2D5V, chain A) reveals that Q420R would be roughly in the same position as HNF6alpha K53, which points in the minor groove of the DNA and makes indirect contact to the DNA backbone via structural water molecules (Figure 6). This mutation may likely affect the overall affinity of the structural complex.

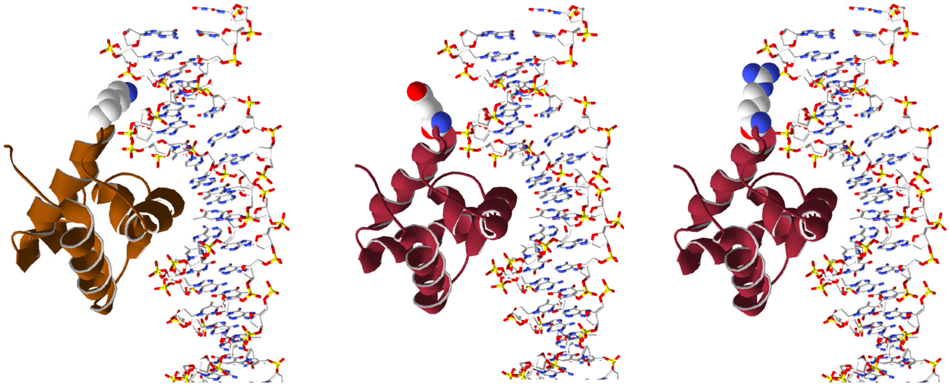


Figure 6. Highlight of the Q420R mutation after superposition of the SATB1 CUT1 domain (pdb entry 2O4A) onto the HNF6alpha DNA binding domain bound to DNA (pdb entry 2D5V) showing its close proximity to DNA backbone. Left: HNFa, middle: SATB1 WT, right: SATB1 mutant.

p.Q525R

Q525 is located in the CUT2 domain alpha-helix that binds the major groove of the DNA, and is the equivalent of CUT1 domain Q402. Since its sidechain makes direct contact with a nucleotide, a mutation to an arginine, which has a longer sidechain, would need to adopt a conformation less favourable to DNA binding to avoid colliding into the DNA, hence affecting the DNA binding affinity at the cognate sites (Figure 7).

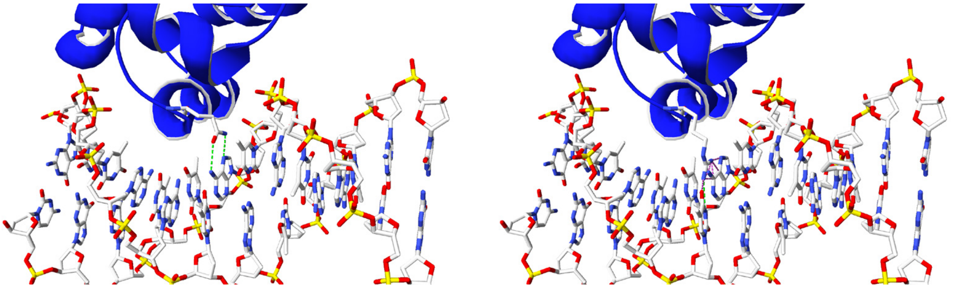


Figure 7. Closeup of the Q525 – DNA interaction highlighting the native residue (Gln, left panel) which could make hydrogen bonds to the base (green dotted lines), whereas the longer Arg sidechain (right panel) might collide into the DNA (purple dotted lines) and be forced to adopt a conformation less favourable with respect to binding its cognate DNA.

p.E530G

E530 is located in the middle of the CUT2 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT1 domain E407. Since its sidechain helps to keep the sidechain of Arg533 in place via hydrogen bonds and both residues make direct contact with the nucleotides, a mutation to a glycine, which bears no sidechain and is not favoured in alpha-helices will likely disrupt the local conformation and alter the DNA binding affinity at the cognate sites (Figure 8).

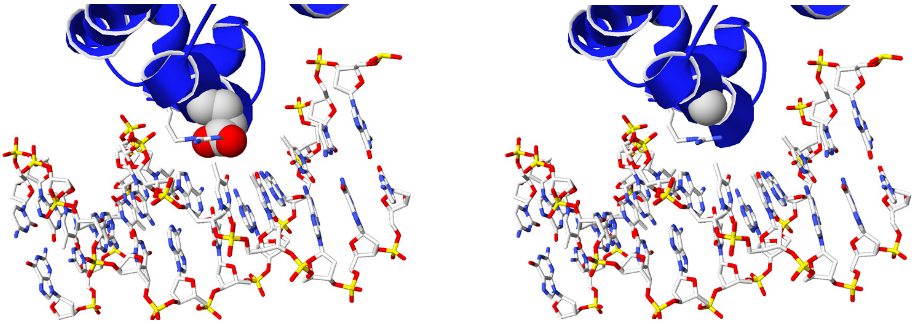


Figure 8. Closeup of the E530 - DNA binding interaction (pdb structure 2O4A) highlighting the native residue (Glu, spacefilled, left panel), which locks in place the sidechain of Arg533 through hydrogen bonds (green dotted lines) and the hole left by the mutation (Gly, spacefilled right panel).

p.E530K

E530 is located in the middle of the CUT2 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT1 domain E407. Since its sidechain help maintain the sidechain of Arg533 in place via hydrogen bonds and that both residues make direct contact with the nucleotides. A mutation to a Lysine, which is very flexible and can be accommodated from a steric point of view will likely induce a rearrangement of these two positively charged sidechains, both in close proximity to

DNA bases, and result in a change of affinity at the cognate sites (Figure 9).

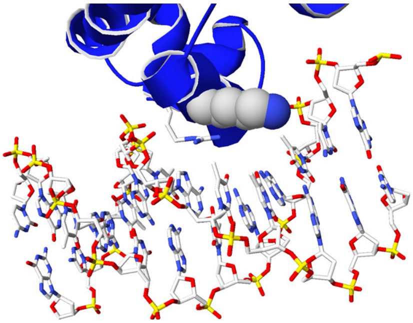


Figure 9. Closeup of the E530 – a conformation that could be adopted by a lysine at this position.

p.E530Q

E530 is located in the middle of the CUT2 domain alpha-helix that binds the major groove of the DNA and is the equivalent of CUT1 domain E407. Since its sidechain help maintain the sidechain of Arg533 in place via hydrogen bonds and that both residues make direct contact with the nucleotides. A mutation to a Glutamine can probably be accommodated from a steric point of view but will induce a rearrangement of these two residues, both in close proximity to DNA bases, and probably result in a change of affinity at the cognate sites (Figure 10).

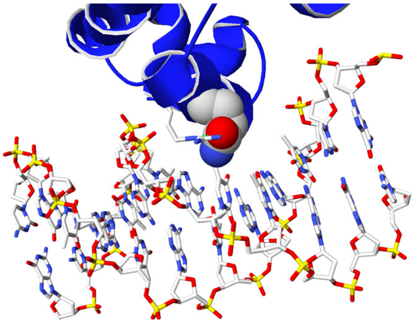


Figure 10. Closeup of the E530 – a conformation that could be adopted by a glutamine at this position.

p.E547K

E547 is located at the surface of the CUT2 domain, not in direct contact with DNA. A lysine at this position could easily be accommodated, but since it substitutes a negative charge with a positive one, it may affect the binding of CUT2 to other domains. Of note, the superposition of the CUT2 domain onto the DNA binding domain of rat HNF6 alpha bound to the TTR promoter (pdb entry 2D5V, chain A4) reveals that E547K would be roughly in the same position as HNF6alpha E57, which is solvent exposed. Interestingly, it is also in a position close to the CUT1 domain variant Q420R, just one turn of alpha-helix away. This mutation will likely affect the overall binding affinity of other domains to the CUT2 domain (Figure 11).

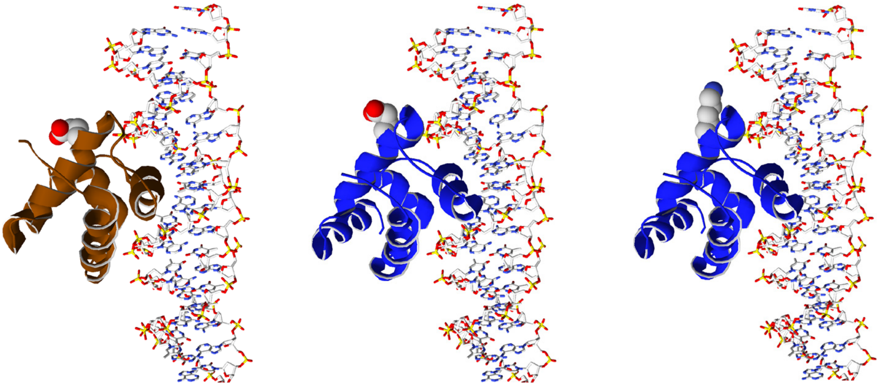


Figure 11. Highlight of the E547K mutation after superposition of the SATB1 CUT2 domain model onto the HNF6alpha DNA binding domain bound to DNA (pdb entry 2D5V) showing its close proximity to DNA backbone. Left: HNFa, middle: SATB1 WT, right: SATB1 mutant.

p.L682V

L682 is not proximal to DNA. It is located at the end of the alpha-helix E672-L682, just before a loop, neither of which are either in contact with the DNA. It is buried and probably contributes to maintain the homeobox domain fold. The valine mutant will have a less optimal packing of this region, and its branched sidechain is predicted to moderately clash with Ala 655 and Leu 684 sidechains and is expected to induce a small conformational change in this region. This in turn might subtly affect the binding affinity of other protein domains of the whole complex (Figure 12).

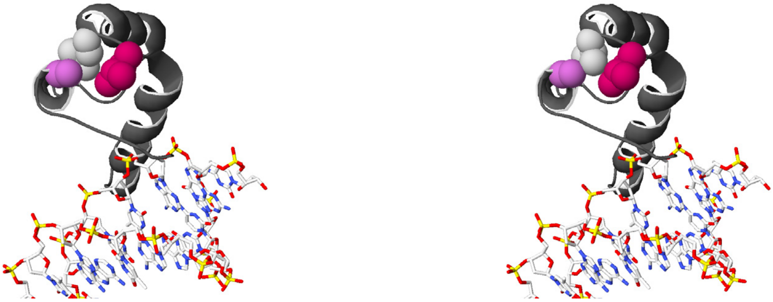


Figure 12. Closeup of the L682V mutation. Left: L682 sidechain (white) is tightly packed with A655 (pink) and L684 (strawberry). Right: V682 sidechain slightly bumps into A655 and L684.

References supplemental notes

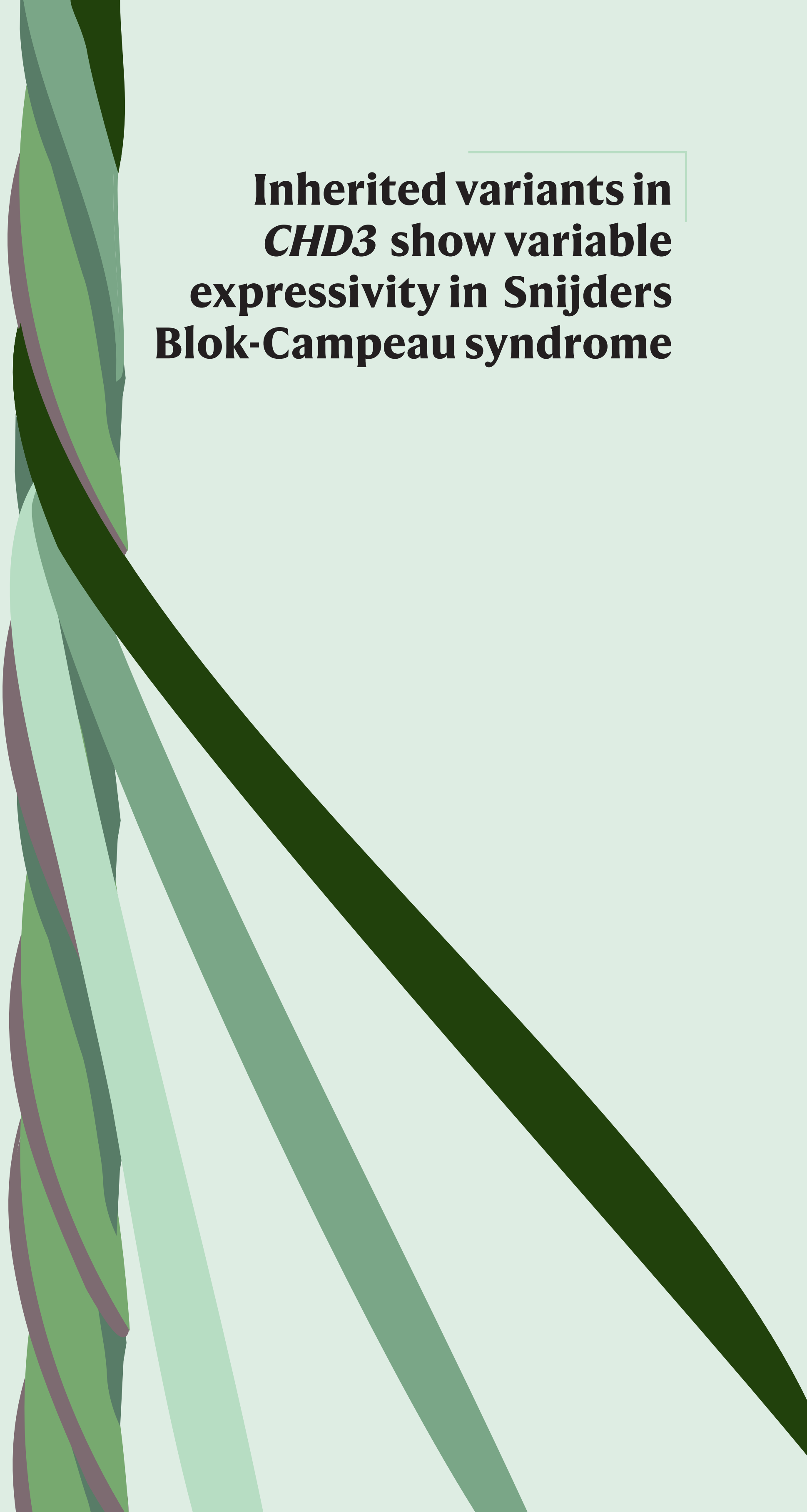
1. Wang, Z., Yang, X., Guo, S., Yang, Y., Su, X.C., Shen, Y., and Long, J. (2014). Crystal structure of the ubiquitin-like domain-CUT repeat-like tandem of special AT-rich sequence binding protein 1 (SATB1) reveals a coordinating DNA-binding mechanism. The Journal of biological chemistry 289, 27376-27385.
2. Yamasaki, K., Akiba, T., Yamasaki, T., and Harata, K. (2007). Structural basis for recognition of the matrix attachment region of DNA by transcription factor SATB1. Nucleic acids research 35, 5073-5084.
3. Johansson, M.U., Zoete, V., Michielin, O., and Guex, N. (2012). Defining and searching for structural motifs using DeepView/Swiss-PdbViewer. BMC bioinformatics 13, 173.
4. Iyaguchi, D., Yao, M., Watanabe, N., Nishihira, J., and Tanaka, I. (2007). DNA recognition mechanism of the ONECUT homeodomain of transcription factor HNF-6. Structure (London, England : 1993) 15, 75-83.
5. Weiss, M.S., Jabs, A., and Hilgenfeld, R. (1998). Peptide bonds revisited. Nature Structural Biology 5, 676-676.

4

This chapter has been published as:
van der Spek, J.*, den Hoed, J.*, Snijders Blok, L., Dingemans, A. J., Schijven, D., Nellaker, C., ... & Kleefstra, T. (2021). Inherited variants in CHD3 demonstrate variable expressivity in Snijders Blok-Campeau syndrome. Genetics in Medicine 24(6), 1283-1296.

* authors contributed equally

Inherited variants in *CHD3* show variable expressivity in Snijders Blok-Campeau syndrome



Abstract

Purpose
Common diagnostic next-generation sequencing strategies are not optimised to identify inherited variants in genes associated with dominant neurodevelopmental disorders (NDDs) as causal when the transmitting parent is clinically unaffected, leaving a significant number of cases with NDDs undiagnosed.

Methods
We characterised 21 families with inherited heterozygous missense or protein-truncating variants (PTVs) in *CHD3*, a gene in which *de novo* variants cause Snijders Blok-Campeau syndrome (SNIBCPS).

Results
Computational facial and human phenotype ontology-based comparisons demonstrated that the phenotype of probands with inherited *CHD3* variants overlaps with the phenotype previously associated with *de novo CHD3* variants, while heterozygote parents are mildly or not affected, suggesting variable expressivity. Additionally, similarly reduced expression levels of *CHD3* protein in cells of an affected proband and of healthy family members with a *CHD3* PTV, suggested that compensation of expression from the wildtype allele is unlikely to be an underlying mechanism. Notably, the majority of the inherited *CHD3* variants were maternally transmitted.

Conclusion
Our results point to a significant role of inherited variation in SNIBCPS, a finding that is critical for correct variant interpretation and genetic counselling and warrants further investigation towards understanding the broader contributions of such variation to the landscape of human disease.

Introduction

The availability of exome sequencing in clinical practice has greatly improved the yield of genetic diagnostics for individuals with neurodevelopmental disorders (NDDs). In particular, sequencing of proband-parent trios, followed by filtering for *de novo*[1] or bi-allelic variants[2], has proven a powerful tool to identify causal variants in individuals with sporadic dominant and recessive NDDs. However, while *de novo* and bi-allelic variants explain a substantial proportion of cases with NDDs[1; 2], the majority remains undiagnosed[3]. Various factors may explain the difficulties to diagnose these individuals, including variation in genes not yet associated to disease, polygenic inheritance or variation in non-coding regions[4]. Also coding variants associated with reduced penetrance and variable expressivity may underlie unexplained NDD cases. Common diagnostic strategies to analyse next-generation sequencing data are not optimised to identify the contributions of these factors to disease. While penetrance indicates the proportion of individuals with a particular variant with a phenotype, expressivity describes the variability in severity of the phenotype between individuals with this variant[4]. Variable expressivity can cause highly variable symptoms, even in severe disorders that are caused by variants with a large effect[4; 5].

In the present study we show variable expressivity for variation in *CHD3*. CHD3 is an ATP-dependent chromatin remodelling protein that serves as core member of the Nucleosome Remodelling Deacetylase complex[6]. Heterozygous variants in *CHD3* have recently been shown to cause a neurodevelopmental syndrome with a variable phenotype, ranging from mildly to more severely affected cases (MIM #618205, Snijders Blok-Campeau syndrome: SNIBCPS)[6; 7]. *CHD3* is extremely intolerant for both loss-of-function (LoF) and missense variation (pLI = 1, o/e = 0.09 (0.05 - 0.15); Z = 6.15, o/e = 0.5 (0.46 – 0.53)), suggesting haploinsufficiency as a possible disease mechanism. However, the large majority of cases diagnosed with SNIBCPS carry confirmed *de novo* missense variants or single amino acid in-frame deletion variants (51/55, 93% of cases)[6; 7], clustering in the ATPase-Helicase domain of the encoded protein, and affecting its ATPase activity and/or chromatin remodelling functions, which could be consistent with a dominant-negative mechanism[6].

We assembled a cohort of 21 families with inherited *CHD3* variants and used a combination of objectified in-depth clinical analyses, cell-based expression studies and large population cohort analyses to confirm the association of inherited *CHD3* variants with SNIBCPS and to show that heterozygote parents, who were predominantly females, often have (very) mild phenotypes, demonstrating variable expressivity.

Results

Phenotypic features in probands with inherited *CHD3* variants overlap with the SNIBCPS-phenotype

We identified 21 families with SNIBCPS, each initially identified through a proband diagnosed with a syndromic NDD carrying a rare inherited *CHD3* missense variant (*n*=13) or PTV (*n*=8) (NM_001005273.2/ENST00000330494.7; Figure 1). Based on clinical observations, all probands had phenotypes overlapping with the SNIBCPS-phenotype associated with *de novo* variants in *CHD3* (Figure 2; Supplemental Notes

1; Table 1). Computational facial analysis also confirmed the presence of a SNIBCPS facial gestalt in probands (Figure S3), and composite images showed similarities in facial features between probands with *de novo* and inherited *CHD3* variants (squared face, deep set eyes, pointed chin; Figure 2B).

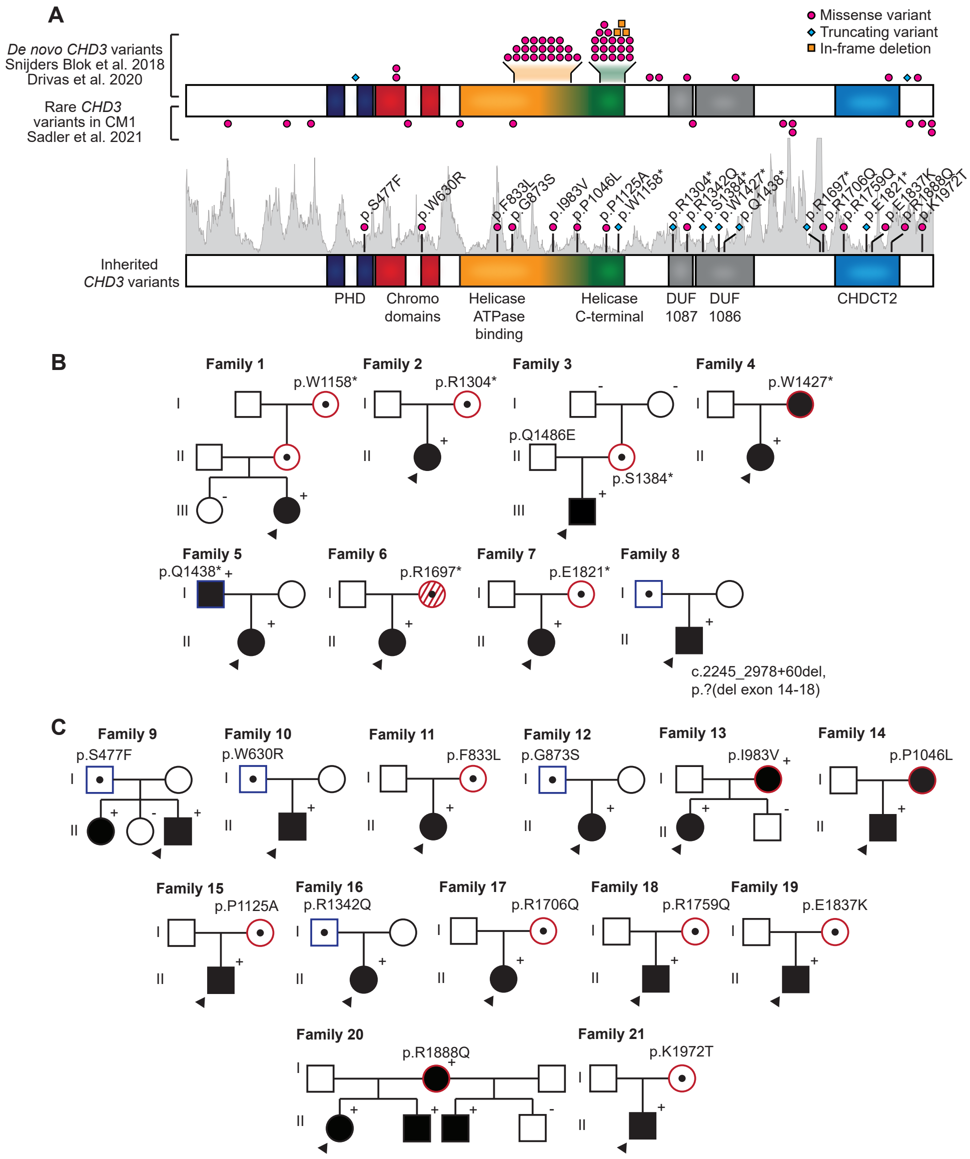


Figure 1. Twenty-one families with inherited CHD3 variants. A) Schematic representation of the CHD3 protein (NM_001005273.2/NP_001005273.2), including functional domains, with PTVs labelled as cyan diamonds, in-frame deletions as orange squares, and missense variants as magenta circles. The intolerance landscape visualised using MetaDome[43] and computed based on single-nucleotide variants in the GnomAD database showing per amino acid position the missense over synonymous ratio, is shaded in grey. The top schematic shows the cases with *de novo CHD3* variants identified in NDD reported in Ref. 6

legend continues on next page

In depth phenotypic analysis of probands with inherited *CHD3* variants and their heterozygote parents

For seventeen heterozygote parents (17/21, 81%) phenotypic information minimally regarding development and dysmorphisms was available. All parents had at least one feature of SNIBCPS. In five parents (5/17, 29%) this was limited to only one (family 1 and 10) or two phenotypic features (family 9, 16 and 21). Whereas the majority of heterozygote parents (16/17, 94%) presented with a single (e.g. prominent forehead or deep-set eyes) or several facial features known in SNIBCPS and 50% had macrocephaly (8/16) (Table 1), the parents had either mild/borderline intellectual disability (4/19, 21%) or no history of intellectual disability (15/19, 79%) (Table 1; Supplemental Notes 1). Taken together, these observations suggest a combination of both variable expressivity and reduced penetrance for these rare genetic variations in CHD3.

We more objectively compared the phenotypes of probands with *de novo* and inherited CHD3 variants based on Human Phenotype Ontology (HPO) terminology[8], using a Partitioning Around Medoids clustering algorithm. While this computational analysis did not identify a phenotypic difference between probands with *de novo* and inherited variants (31/55 individuals clustered correctly, $p=0.44771$; Figure 2C and Figure S4), it confirmed a phenotypic difference between probands with inherited *CHD3* variants and their heterozygote parents (33/40 individuals clustered correctly, $p<0.00001$; Figure 2C and Figure S4).

Maternal transmission of inherited *CHD3* variants is predominant

We noticed that the majority of variants in our cohort were maternally inherited (15/21, 71%, $p=0.0392$; Figure 1B-C and Figure S5A). For single nucleotide variants with a LoF effect, 6/7 (86%, $p=0.0625$) variants were maternally inherited (Figure 1B and Figure S5A). Notably, the only father transmitting a LoF single nucleotide variant was affected (mild intellectual disability). This observation could hint at a female-protective effect for genetic variation in *CHD3*. However, we did not observe a sex-bias in the affected probands (12/21 female, $p>0.9999$), or more severe intellectual disability in male compared to female *de novo* or inherited cases[6; 7]. To further explore the hypothesis of a female protective effect at population level, we analysed all *CHD3* LoF variants in GnomAD (15/198,800 individuals) and found that a significantly higher number of these variants were present in females than in males (12/15, $p=0.0173$; Figure S5B).

---

and Ref. 7 and rare variants associated with Chiari I malformations (CM1) reported in Ref. 34. The bottom schematic presents cases with inherited *CHD3* variants described in this study. **B-C**) Pedigrees of families identified with inherited *CHD3* variants, with in (B) families with predicted LoF variants and in (C) families with missense variants. The arrow head indicates the proband, filled symbols represent affected individuals (defined as individuals with developmental delay and/or intellectual disability), open symbols with a central dot represent confirmed heterozygotes without developmental delay/intellectual disability, '+' is used for a confirmed familial *CHD3* variant and '-' for individuals confirmed not to carry the variant. Symbols with red contours represent female heterozygotes, symbols with blue contours represent male heterozygotes. Dashed symbol for family 6 represents mosaic state of the variant in the mother. In pedigrees, only genetically tested siblings of the proband are shown.

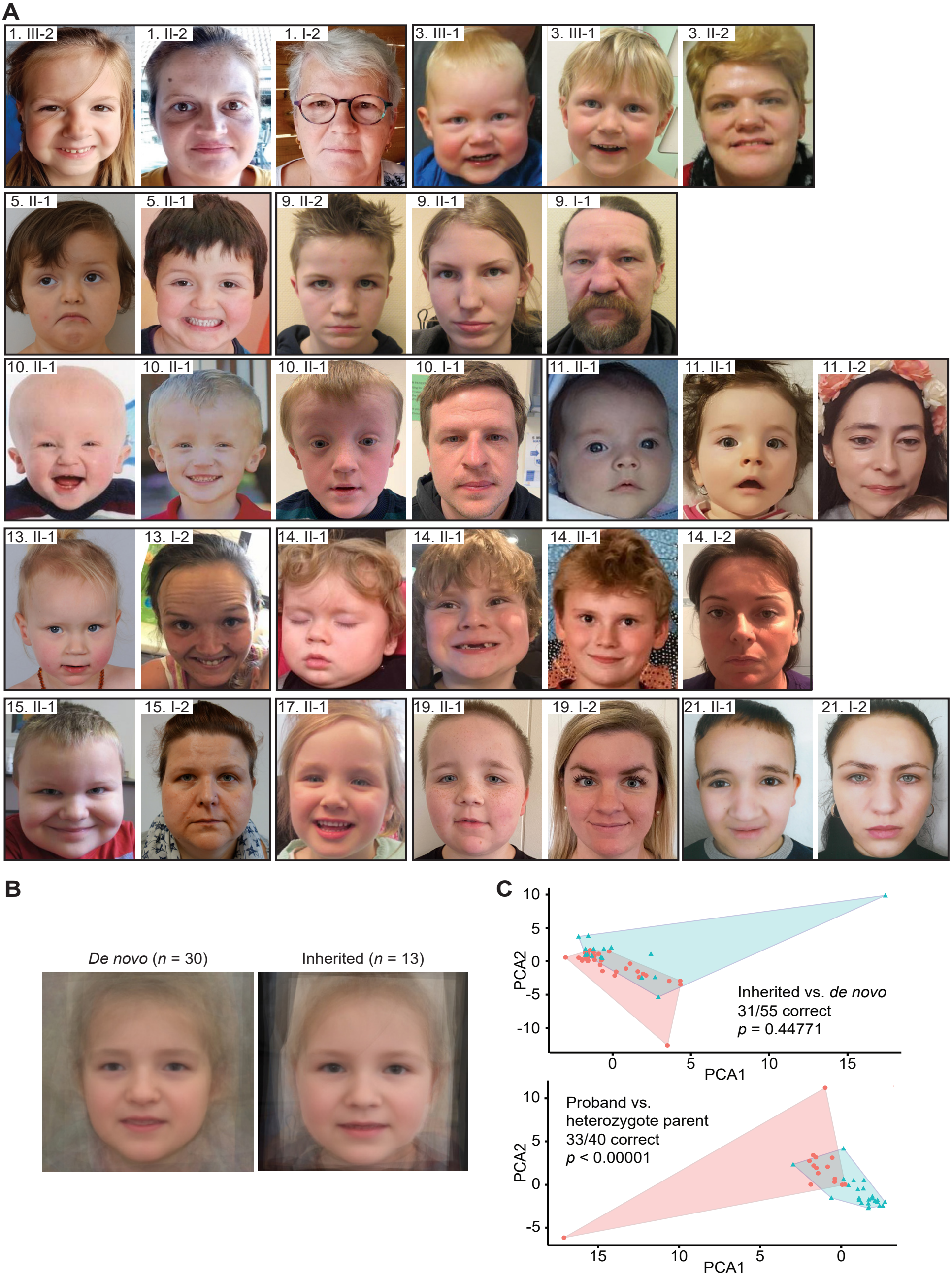


Figure 2. Facial features and clinical evaluation of individuals with inherited *CHD3* variants. A) Facial photographs of individuals with inherited *CHD3* variants. Individuals demonstrate features also observed in individuals with de *novo CHD3* variants, including a squared appearance of the face, prominent forehead, widely spaced eyes, thin upper lip, pointed chin and deep-set eyes. These characteristics are also present in heterozygote parents. As observed previously, facial gestalt changes with age[7]. For example, a prominent nose is especially seen in adult individuals. For childhood pictures of heterozygote

legend continues on next page

Table 1. Summary of phenotypes seen in individuals with *CHD3* variants

	Probands with *de novo* variant[1]	Probands with inherited variant	Heterozygote parents
Development			
Developmental delay	100% (55/55)	100% (21/21)	17% (3/18)
Intellectual disability	98% (46/47)	79% (11/14)	21% (4/19)
borderline/borderline-mild	6% (3/47)	14% (2/14)	11% (2/18)
mild/mild-moderate	30% (14/47)	29% (4/14)	6% (1/18)
moderate/moderate-severe	36% (17/47)	14% (2/14)	0% (0/18)
severe	23% (11/47)	0% (0/14)	0% (0/18)
level unknown	2% (1/47)	21% (3/14)	6% (1/18)
Speech delay/disorder	100% (53/53)	100% (20/20)	24% (4/17)
Autism or autism-like features	35% (18/51)	53% (10/19)	18% (3/17)
Neurology			
Hypotonia	81% (39/48)	89% (17/19)	17% (2/12)
Macrocephaly	53% (28/53)	47% (9/19)	50% (8/16)
CNS abnormalities	50% (24/48)	62% (8/13)	25% (1/4)
Neonatal feeding problems	31% (10/32)	21% (4/19)	6% (1/16)
Facial dysmorphisms			
High/broad/prominent forehead	85% (28/33)	85% (17/20)	53% (9/17)
Thin upper lip	79% (15/19)	55% (11/20)	47% (8/17)
Widely spaced eyes	69% (35/51)	70% (14/20)	24% (4/17)
Broad nasal bridge	75% (15/20)	80% (16/20)	24% (4/17)
Full cheeks	58% (11/19)	70% (14/20)	13% (2/16)
Pointed chin	60% (12/20)	53% (10/19)	41% (7/17)
Deep-set eyes	55% (11/20)	47% (9/19)	50% (8/16)
Other			
Joint laxity (generalized and/or local)	36% (18/50)	40% (8/20)	29% (4/14)
Vision problems	72% (38/53)	25% (5/20)	53% (8/15)
Male genital abnormalities	32% (8/25)	22% (2/9)	0% (0/4)
Hernia (umbilical, inguinal, hiatal)	13% (6/48)	10% (2/20)	0% (0/14)

[1] Combined cases from Snijders Blok et al. 2018 and Drivas et al. 2020 (confirmed *de novo* only)

CNS: central nervous system

---

parents, see Figure S2. **B**) Computational average of facial photographs of 30 individuals with *de novo CHD3* variants (left) and 13 probands with inherited *CHD3* variants (right). **C**) Partitioning Around Medoids analyses of clustered HPO-standardised clinical data from 35 individuals with *de novo CHD3* variants, 20 affected probands with an inherited variant, and 20 heterozygote parents. The analyses do not show a significant distinction between the clusters of probands with *de novo* and probands with inherited variants (upper graph; p=0.44771). There is, however, a significant difference between the clusters of affected probands with inherited variants and heterozygote parents (bottom graph; p<0.00001).

Effects of an inherited CHD3 PTV on transcript and protein expression levels

Few cases with SNIBCPS have been described with confirmed *de novo CHD3* PTVs (4/55, 7.3% of cases)[6; 7] including one which is predicted to escape nonsense-mediated decay (NMD; NP_001005273.1:p.(Phe1935GlufsTer108)). However, in our study we identified seven families with inherited single nucleotide PTVs and one with an intragenic deletion with a predicted LoF effect (8/20, 40%; Figure 1A). None of the inherited PTVs were predicted to escape NMD. We functionally confirmed this in family 1 (Figure 3A), for which we treated lymphoblastoid cell lines from the proband (individual III-2), heterozygote mother (II-2) and grandmother (I-2) and the healthy sibling of the proband that did not carry the variant (III-1) with cycloheximide to inhibit NMD, followed by direct amplification and Sanger sequencing of the *CHD3* transcript. We found that treatment with cycloheximide increased the expression of mutant allele, showing that the NM_001005273.2:c.3473G>A variant was targeted by NMD in all samples, as expected (Figure 3B).

An explanation for variable expressivity of PTVs could be compensation of expression by the wildtype allele to maintain normal expression levels. To test if such compensation plays a role in variable expressivity of *CHD3* PTVs, we evaluated the expression of the *CHD3* variant in family 1 (c.3473G>A, p.(W1158*)) on a transcript and protein level. We found that this variant resulted in lower levels of *CHD3* transcript and CHD3 protein in lymphoblastoid cells from individuals I-2, II-2 and III-2 compared to the levels observed in cells from the healthy sibling who did not carry the variant (individual III-1; Figure 3C-D). These findings confirm the LoF effect of the stop-gain variant in this family, and make it unlikely that compensation by the wildtype allele is an underlying mechanism for the milder phenotype in the heterozygote mother and grandmother.

In silico and functional analyses of inherited *CHD3* missense variants

In addition to the seven inherited *CHD3* single nucleotide PTVs and the intragenic deletion, we identified 13 families with inherited missense variants. One of the identified inherited missense variants, also present in an unaffected heterozygote parent, was identical to a variant previously reported as a *de novo* variant in an individual with SNIBCPS (p.(R1342Q); individual 32 in Ref. 6). Based on the phenotypes observed in the probands with inherited CHD3 missense variants, the conservation of affected positions (Figure S1), and *in silico* predictions of pathogenicity (Figure S6), we considered these inherited *CHD3* missense variants as likely pathogenic with variable expressivity in the parents. Clinically, probands carrying a *CHD3* missense variant did not seem to be more severely affected than individuals with PTVs (Table S1). We followed up on the inherited missense variants using cell-based functional assays to test for chromatin binding (for p.(S477F)) and GATAD2B-binding (for p.(R1342Q), p.(E1837K) and p.(Q1888R)), but did not find evidence that these protein functions were affected (Figure S7).

Rare *CHD3* variants in a large population cohort

The presence of rare, likely pathogenic CHD3 variants in healthy individuals prompted us to study possible effects of variation in this gene at a population level, using data from the UK Biobank resource[9-14]. For a detailed description of these analyses, see Supplemental Notes 3. These analyses were limited to white-British ancestry. First, we found no associations between rare missense variation at variation-intolerant

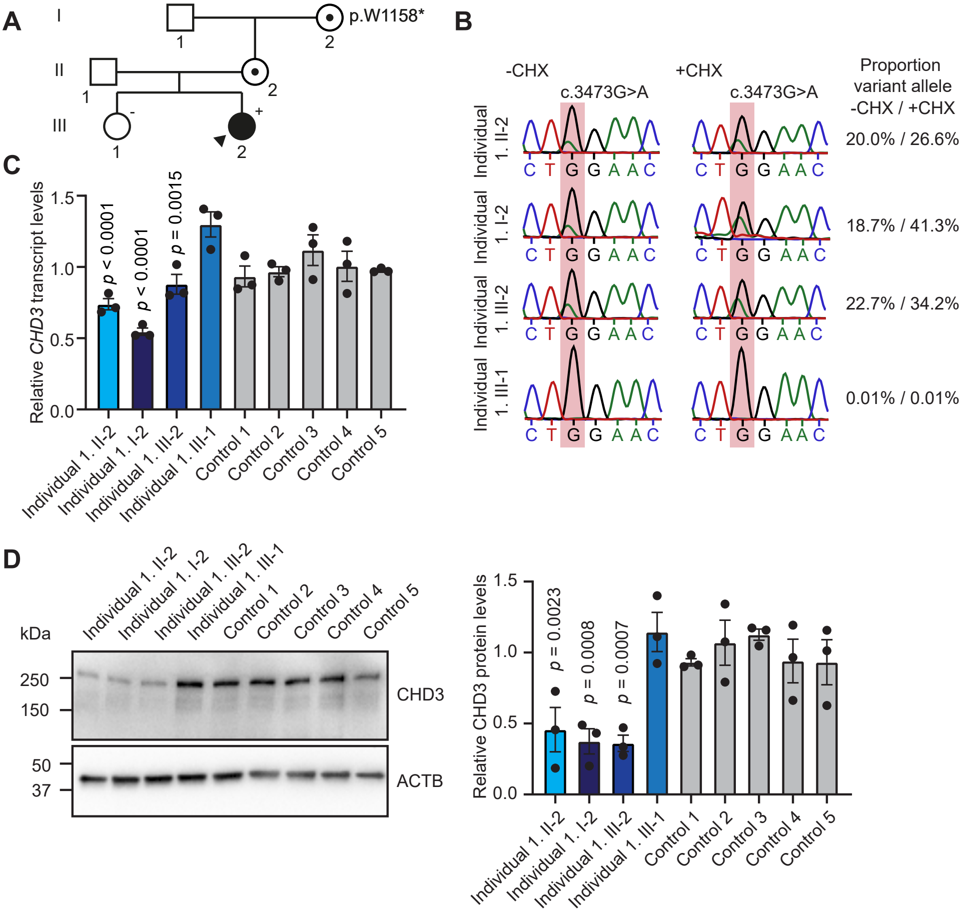


Figure 3. Functional consequences of the CHD3 p.(W1158*) PTV in subject-derived lymphoblastoid cell lines. A) Pedigree of family identified with an inherited *CHD3* c.3473G>A, p.(W1158*) variant **B**) Sanger sequencing chromatographs of EBV-immortalised lymphoblastoid cell lines derived from members of family 1. Individuals I-2, II-2 and III-2 carried the c.3473G>A, p.(W1158*) variant and individual III-1 was a healthy sibling that did not carry the variant. Cells were treated with (+CHX) or without cycloheximide (-CHX) to test for NMD. The mutated position is shaded in red. The transcript carrying the variant allele is present at lower levels than the wild-type allele, and increases after CHX treatment (proportion variant allele calculated as: peak area variant allele / (peak area variant + wildtype allele), showing that this variant is targeted by NMD. **C**) qPCR of EBV-immortalised lymphoblastoid cell lines of family 1 (shades of blue) and five unrelated controls (grey) for *CHD3* transcript levels (NM_001005273.2). Values are normalised to expression of *PPIA* and *TBP* and shown relative to unrelated controls. Bars represent the mean ± S.E.M. with individual data points plotted (*n*=3; *p*-values compared to individual III-1 (healthy sibling that did not carry the variant), one-way ANOVA and *post-hoc* Bonferroni test). **D**) Left, a representative immunoblot of protein lysates prepared from lymphoblastoid cell lines for CHD3 (expected molecular weight: ~227 kDa). The blot was probed for ACTB, to ensure equal protein loading. Right, a graph showing the quantification of immunoblots with bars presenting the mean ± S.E.M. and individual data points plotted (*n*=3; *p*-values compared to individual III-1 (healthy sibling that did not carry the variant), one-way ANOVA and *post-hoc* Bonferroni test). Controls are shaded in grey and samples from family 1 are shaded in blue. **C-D**) The cell lines carrying c.3473G>A, p.(W1158*) show lower *CHD3* transcript/protein levels compared to the control samples.

locations in *CHD3* (minor allele frequency ≤ 1%, located in functional domains, damaging in PolyPhen or SIFT and with a CADD-PHRED score > 25) and fluid intelligence (N=77,998), educational qualification (N=120,596) or intracranial volume (N=18,254). We then tested for group differences for these three phenotypes between individuals with and without rare putative *CHD3* LoF variants. At nominal significance, we observed a larger intracranial volume in individuals with rare CHD3 putative LoF variants (*n*=4, t=2.37, *p*=0.018). We note that this result does not remain significant after a conservative Bonferroni correction for testing of three different phenotypes (adjusted *p*=0.054). However, in light of the observed macrocephaly in 47-53% of probands with a (likely) pathogenic *CHD3* variant and in 50% of heterozygote parents (Table 1), and the link of rare CHD variants with abnormal brain growth[15], the potential convergence of findings in the four individuals with *CHD3* LoF variants in this independent population cohort is intriguing. To also test possible relationships between *CHD3* common genetic variation and head circumference and/or intracranial volume, we performed gene-level analyses using previously published SNP-wise association summary statistics for these traits[16; 17], but none of the results survived multiple testing correction (Supplemental Notes 3).

Discussion

In the present study we used inherited variation to show variable expressivity for SNIBCPS. The phenotypic spectrum of individuals with an inherited *CHD3* variant ranged from moderate intellectual disability combined with multiple other features, to only a single facial feature or macrocephaly. Additional evidence for variable expressivity for *CHD3* variation is provided by the recently identified association of 19 rare *CHD3* missense variants with Chiari I malformations in individuals without features of SNIBCPS (Figure 1A)[15]. It is important to note that, although younger generations seem more severely affected than previous generations this may be due to ascertainment bias[18].

The female predominance we observed among the heterozygote parents in our cohort and for individuals with *CHD3* LoF variants in GnomAD could indicate a female protective effect for *CHD3* variation. Previous studies have repeatedly demonstrated a male bias in NDDs, a higher pathogenic variant burden in females and a maternal transmission bias in rare inherited variants[3; 19; 20], suggesting that female sex protects against genetic variation in disease. This phenomenon might contribute to the variable expressivity observed for the inherited *CHD3* variants.

Using transcript and protein expression studies we found significantly lower CHD3 expression levels in three family members carrying a *CHD3* PTV, independent of whether or not these individuals were affected with intellectual disability. Hence, we found no evidence for compensatory expression by the wild type allele in blood derived cells. However, it remains, to be determined, whether such LoF variance can have a tissue-specific, temporal expression specific, and/or transcript specific effect. It is unclear whether results from blood-derived cells can be extrapolated to neuronal cell types, which would be more relevant considering the NDD phenotypes in our cohort, especially given that neuron-specific alternative splicing has previously been described for *CHD3*[21].

Other explanations for the clinical variable expressivity of inherited *CHD3* variants include the presence of a second-hit on the other allele by either rare or common variation, a genome wide higher mutational burden of high-penetrant variants, or common variants in promoter/enhancers regions or in other genes, inherited from the parent that did not transmit the inherited *CHD3* variant[4; 22]. Such a compound inheritance mechanism, has, for example, been described for thrombocytopaenia with absent radii (TAR) syndrome, where the inheritance of a rare null allele together with one of two low-frequency SNPs in regulatory regions causes disease[23]. In four probands with an inherited *CHD3* variant a copy number variant (CNV) was also reported, including one 22q11.2 duplication, which has been described with highly variable features (MIM #608363) and three CNVs of unknown significance. Proband 7 had other (likely) pathogenic variants contributing to the phenotype (Supplemental Notes 1). A comparison with the prevalence of additional genetics finding in individuals with *de novo CHD3* variants could not be made due to lack of reporting on additional genetic findings[6,7].

The individuals with *de novo* missense variants published to date were mostly (although not entirely) localised to the ATPase-Helicase domain[6; 7]. No clustering to the ATPase-Helicase domain or elsewhere was observed among the inherited missense variants of our cohort (Figure 1A). It has been speculated that the *de novo* missense variants clustering in the ATPase-Helicase domain are unlikely to lead to a sole LoF effect[6], and may potentially act in a dominant-negative way. The identification of eight families with an inherited LoF variant and the lack of clustering of the inherited missense variants may suggest a LoF effect as the main mechanism for inherited cases, which may underlie the variable expressivity. However, our cell-based analyses did not find evidence of LoF for the protein functions that we tested (Figure S7). This does not exclude that these variants have an effect on other biological functions of CHD3. Based on 3D-protein modelling, the prior published *de novo* missense variants within the ATPase-Helicase domain localise more closely to the ATP-binding site than the inherited missense variants of our cohort (Supplemental Notes 2). Interestingly, the p.(I983V) (family 13) variant was found to be closer to published *de novo* variants (Supplemental Notes 2) and the heterozygote parent with this missense variant did have a neurodevelopmental phenotype which was more pronounced than in other heterozygote parents (Figure 1C and Figure 2; Supplemental Notes 1).

With the identification and characterisation of inherited *CHD3* variants with variable expressivity in 21 families, we showed that, in addition to highly penetrant *de novo* variants, rare predicted likely pathogenic inherited variants in *CHD3* should be considered as possibly pathogenic depending on variant characteristics in cases with phenotypic concordance to SNIBCPS. Interestingly, variable penetrance and expressivity has been noted in numerous families with another dominant NDD, KBG syndrome, caused by LoF variants in *ANKRD11* (MIM #148050)[24]. So this phenomenon is likely more common for dominant NDDs, with important implications for clinical genetic counselling, in the context of recurrence risk, prenatal diagnostics, prognosis and variant interpretation.

Clinically, we recommend that it can be helpful to evaluate the parents of children with *CHD3* variants for subtle SNIBCPS features. In particular macrocephaly and facial dysmorphisms including a prominent forehead and pointed chin could be recognised in a substantial number of heterozygote parents (50% and 94% respectively; Figure 2A). Taken together, our results illustrate the continuum of causality for NDDs with genetic origins[18; 25] and significantly underline the hypothesis that variable expressivity and reduced penetrance likely explain a large portion of as yet unexplained NDD cases. Overall, we show that even for genes already known to be implicated in a NDD, inherited variation and variable expressivity can play a major role, and are thus important to consider in genetic counselling.

Methods

Individuals and consent

The cohort presented in this study was assembled from hospitals and laboratories across the Netherlands, Germany, United States of America, Slovenia, Australia and Canada. Informed consent for the use and publication of medical data and biological material was obtained from all patients or their legal representative by the involved clinician. Consent for publication of photographs was obtained separately.

Next-generation-sequencing

CHD3 variants in all probands were identified using exome sequencing or genome sequencing (family 4 and 12). According to the American College of Medical Genetics (ACMG) guidelines, all *CHD3* variants were classified as variants of unknown significance (class III)[26], with inheritance from seemingly healthy/mildly affected parents combined with previously unreported reduced penetrance as important criteria. Inheritance of variants was confirmed either as part of trio exome sequencing or using targeted Sanger sequencing after identification in singleton exon analysis. Similarly, if applicable, other family members were tested using targeted sequencing.

Pathogenicity of missense variants was further evaluated using CADD-PHRED v1.6[27], PolyPhen-2[28] and SIFT[29] scores. Allele frequencies of all variants in GnomAD were based on ENST00000330494.7[30].

Facial analysis

We established a 2D hybrid facial model which combines the analysis of the ‘Clinical Face Phenotype Space’ pipeline with the facial recognition system of the ‘OpenFace’ pipeline[31; 32]. First, we generated a 468-dimensional feature vector of the facial features of 30 individuals with *de novo CHD3* variants. After extraction of the hybrid features for each of the individuals, we calculated whether the individuals with *de novo CHD3* variants cluster together when compared to a group of matched controls based on the nearest neighbour principle (Euclidean distance) – these matched controls were individuals with ID and are age-, ethnicity- and sex matched. The Mann-Withney U test was used to determine whether the clustering of individuals with *de novo CHD3* variants was significantly higher than expected based on random chance. A *p*-value smaller than 0.05 was considered significant.

Furthermore, a classifier was built using a logistic regression model trained on the 468-dimensional feature vector of the 30 individuals. The performance was evaluated performing leave-one-out cross validation and the classifier was shown to have a sensitivity of 0.91, a specificity of 0.83 and an overall area under the ROC-curve of 0.91. Finally, using the trained classifier, we determined for each inherited case whether that individual clusters within the *de novo CHD3* group or the control group (Figure S3).

Construction of composite face

For 13 individuals with an inherited *CHD3* variant and 30 with a *de novo CHD3* variant, facial 2D-photographs were available for generating a composite face. As previously described, average faces were generated while allowing for asymmetry preservation and equal representation by individuals[33].

Human Phenotype Ontology (HPO)-based phenotype clustering analysis

We performed HPO-based clustering analysis using 35 individuals with *de novo CHD3* variants[6], 20 of 21 probands with an inherited *CHD3* variant, and 20 of 21 heterozygote parents in the analysis: the proband and heterozygote mother of family 6 were excluded because no clinical data were available, and the mother is mosaic for the *CHD3* variant (~37%). The Wang score (a measure of semantic similarity) between all terms was calculated using the HPO Sim package[34; 35]. The terms were divided in groups, based on the similarity score: a new feature – the sum of the terms in the group - was created as a replacement for the terms in that specific group (Figure S4A). HPO terms that could not be added to a group feature were added as a separate term. To quantify and visualise possible differences in our cohort, we used Partitioning Around Medoids (PAM) clustering on these grouped features. We compared probands with a *de novo* and inherited variant and probands with inherited variants and their heterozygote parents in a second analysis. To assess statistical significance, a permutations test (*n*=100,000) was used with relabelling based on variant types, while keeping the original distribution of variant types into account.

Three-dimensional protein modelling

We modeled the protein structure of the ATPase-Helicase domain of CHD3 in interaction with the DNA using the homology modelling script in the WHAT IF[36] & YASARA[37] Twinset with standard parameters. As a template, we used PDB file 6RYR which contains the human Nucleosome-CHD4 complex structure of a single copy of CHD4[38]. The PHD2 variant (p.(S477F)) was modeled in the PHD2 domain of CHD4 (PDB 2L75, 89% sequence identity with CHD3)[39].

DNA expression constructs and site-directed mutagenesis

The cloning of *CHD3* (NM_001005273.2/ENST00000330494.7) has been described previously[6]. The coding DNA sequence of *GATAD2B* (NM_020699.3/ENST00000368655.4) and a C-terminal region of CHD3-encoding residues 1246-1944 (NM_001005273.2) were amplified using primers listed in Table S2. Variants in full-length CHD3 or the C-terminal CHD3 construct were generated using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent). The primers used for site-directed mutagenesis are listed in Table S3. cDNAs were subcloned using BamHI/HpaI (full-length CHD3), BamHI/XbaI (GATAD2B) or HindIII/BamHI (C-terminal CHD3 construct) into pYFP, pHisV5 and pRluc, created by modification of the pEGFP-C2 vector (Clontech). All constructs were verified by Sanger sequencing.

Cell culture

Lymphoblastoid cell lines were established by Epstein-Barr virus transformation of peripheral lymphocytes from blood samples collected in heparin tubes, and maintained in RPMI medium (Sigma) supplemented with 15% foetal bovine serum and 5% HEPES (both Invitrogen). HEK293T/17 cells (CRL-11268, ATCC) were grown in DMEM supplemented with 10% foetal bovine serum and 1x penicillin-streptomycin (all Invitrogen) at 37°C with 5% CO2. Transfections were performed using GeneJuice (Millipore) following the manufacturer's protocol.

Testing for nonsense mediated decay of truncating variants

Lymphoblastoid cell lines of members of family 1 and controls were grown overnight with 100 µg/ml cycloheximide (Sigma) to block NMD. After treatment, cell pellets were collected, and RNA and protein were extracted using the RNeasy Mini Kit (Qiagen) or with 1x RIPA buffer supplemented with 1% PMSF and 1x PIC, respectively. RT-PCR was performed using SuperScript III Reverse Transcriptase (ThermoFisher) with random primers, and regions of interest were amplified from cDNA using primers listed in Table S4. Sanger trace peak sizes of the wildtype and variant allele were measured using the 'Area' option in ImageJ and proportion of the variant allele was calculated: peak area variant allele / (peak area variant + wildtype allele). qPCR was performed with cDNA using iQ SYBR Green supermix (BioRad) and primers listed in Table S5 on the CFX Real-Time PCR Detection System (BioRad).

Direct fluorescent imaging

HEK293T/17 cells were grown on coverslips coated with poly-D-lysine (Sigma). Forty-eight hours after transfection with the YFP-tagged C-terminal CHD3 construct and HisV5-tagged GATAD2B, cells were fixed with 4% paraformaldehyde (PFA, Electron Microscopy Sciences). Nuclei were stained with Hoechst 33342 (Invitrogen). Fluorescence images were acquired with a Zeiss LSM880 confocal microscope and Airyscan unit using ZEN Image Software (Zeiss).

FRAP assays

HEK293T/17 cells were transfected in clear-bottomed black 96-well plates with YFP-tagged full-length CHD3 or p.(S477F). After 48 h, medium was replaced with phenol red-free DMEM supplemented with 10% foetal bovine serum (both Invitrogen), and cells were moved to a temperature-controlled incubation chamber at 37°C. Fluorescent recordings were acquired using a Zeiss LSM880 and Zen Black Image Software, with an alpha Plan-Apochromat 100x/1.46 Oil DIC M27 objective (Zeiss). FRAP experiments were performed by photobleaching an area of 0.98 µm x 0.98 µm within a single nucleus with 488-nm light at 100% laser power for three iterations with a pixel dwell time of 32.97 µs, followed by collection of times series of 150 images with a 2.5 zoom factor and an optical section thickness of 1.4 µm (2.0 Airy units). Individual recovery curves were background subtracted and normalised to the pre-bleach values, and mean recovery curves were calculated using EasyFRAP software[40]. Curve fitting was done with the FrapBot application using direct normalisation and a single component exponential model, to calculate the half-time and maximum recovery[41].

Immunoblotting

Whole-cell lysates were collected in 1x RIPA buffer (ThermoFisher) supplemented with 1x PIC (Roche) and 1% PMSF (Sigma). Cells were lysed for 20 min at 4 °C followed by centrifugation for 20 min at 12,000 rpm. Samples were loaded on 4–15% Mini-PROTEAN TGX Precast Gels (Bio-Rad) and transferred onto polyvinylidene fluoride membranes. Membranes were blocked in 5% milk for 1 h at room temperature and then probed with rabbit-anti-CHD3 antibody (1:1000; Abcam, ab109195) or mouse-anti-GFP (1:8000; Clontech, 632380) overnight at 4°C. Next, membranes were incubated with HRP-conjugated goat-anti-rabbit or goat-anti-mouse antibody (1:10,000; Jackson ImmunoResearch) for 1.5 h at room temperature. Bands were visualised with the SuperSignal West Femto Maximum Sensitivity Substrate Reagent Kit (CHD3; ThermoFisher) or the Novex ECL Chemiluminescent Substrate Reagent Kit (YFP-fusion proteins; Invitrogen) using a ChemiDoc XRS + System (Bio-Rad).

Co-immunoprecipitation

HEK293T/17 cells were transfected with the YFP-tagged C-terminal region of CHD3 and Rluc-tagged GATAD2B. After 48h, whole-cell lysates were collected in Pierce IP Lysis Buffer (25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% NP-40 and 5% glycerol; ThermoFisher) supplemented with 1x PIC (Roche) and 1% PMSF (Sigma). Cells were lysed for 20 min at 4°C followed by centrifugation for 20 min at 12,000 rpm. YFP-fusion proteins were immobilised on GFP-trap magnetic agarose beads (Chromotek) overnight at 4°C. Deactivated beads (Chromotek) were used as a negative control. The elutions and 5% of the input were resolved on 4–15% Mini-PROTEAN TGX Precast Gels (Bio-Rad) and transferred onto polyvinylidene fluoride membranes. Membranes were blocked in 5% milk for 1 h at room temperature and then probed with rabbit-anti-Rluc antibody (1:2000; GeneTex) overnight at 4°C. Next, membranes were incubated with HRP-conjugated goat-anti-rabbit antibody (1:10,000; Jackson ImmunoResearch) for 1.5 h at room temperature. Bands were visualised with the SuperSignal West Femto Maximum Sensitivity Substrate Reagent Kit (ThermoFisher) using a ChemiDoc XRS + System (Bio-Rad).

Population-based analysis of the association of *CHD3* variation with intelligence, educational qualification and intracranial volume/head circumference

Using exome sequencing data of 200,000 individuals from the UKB Exome Sequencing Consortium (Ref. 10,11, and https://www.ukbiobank.ac.uk/media/cfulxh52/uk-biobank-exome-release-faq_v9-december-2020.pdf) we studied the association of *CHD3* missense and putative LoF variants with 'Fluid intelligence score' (data field ID 3533), 'Qualifications' (data field ID 6138) and 'Volume of EstimatedTotalIntraCranial' (data field ID 7054). Additionally, we used genome-wide association meta-analysis summary statistics of head circumference ($N \leq 18{,}881$), and head circumference combined with intracranial volume ($N \leq 45{,}458$) in child- and adulthood[16], and infant head circumference ($N \leq 10{,}768$)[17] to calculate gene-level *p*-values reflecting the common variant associations of *CHD3* with these traits using MAGMA[42]. For detailed description of the methods, see Supplemental Notes 3.

Data availability

All datasets generated and analysed during the current study are available from the corresponding author on request.

Acknowledgements

We are extremely grateful to all families participating in this study. In addition, we wish to thank the members of the Cell culture facility, Department of Human Genetics, RadboudUMC, Nijmegen for culture of cell lines. This work was financially supported by the Dutch Research Council grant to TK (015.014.036 and 1160.18.320) and LELMV (015014066) and Netherlands Organization for Health Research and Development to TK (91718310) and LELMV (843002608, 846002003), by Donders Junior Researcher Grant 2019 to TK and LELMV and the Max Planck Society (JdH, DS, CF, SEF). Authors of this publication are members of the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability ERN-ITHACA [EU Framework Partnership Agreement ID: 3HP-HP-FPA ERN-01-2016/739516]. The aims of this study contribute to the Solve-RD project (CG, HGB, LELMV and TK) which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257. For families 10,13, 14 and 20 sequencing and analysis were provided by the Broad Institute of MIT and Harvard Center for Mendelian Genomics (Broad CMG) and were funded by the National Human Genome Research Institute, the National Eye Institute, and the National Heart, Lung and Blood Institute grant UM1 HG008900 and in part by National Human Genome Research Institute grant R01 HG009141. For family 16 exome sequencing was performed in the framework of the German project "TRANSLATE NAMSE", an initiative from the National Action League for People with Rare Diseases (Nationales Aktionsbündnis für Menschen mit Seltenen Erkrankungen, NAMSE) facilitating genetic diagnostics for individuals with suggested rare diseases. Part of this research has been conducted using the UK Biobank Resource under application number 16066, with Clyde Francks as the principal applicant. Our study made use of imaging-derived phenotypes generated by an image-processing pipeline developed and run on behalf of UK Biobank.

Ethics declaration

All study proceedings involving humans were in compliance with the principles set out in the Declaration or Helsinki. This study was approved by the institutional review board 'Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen' under number 2011/188. Written informed consent for the use and publication of medical data and biological material was obtained from all individuals or their legal representative by the involved clinician. Written informed consent for publication of photographs was obtained specifically and separately. This study includes data from the UK Biobank Study (http://www.ukbiobank.ac.uk)[9]. UK Biobank received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382) and had obtained informed consent from all participants. The present analyses were conducted under UK Biobank data application number 16066.

References

1. Deciphering Developmental Disorders, S., McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433.
2. Martin, H.C., Jones, W.D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson,

J.D., Jones, C.P., Handsaker, J., Gallone, G., Bruntraeger, M., et al. (2018). Quantifying the contribution of recessive coding variation to developmental disorders. Science (New York, NY) 362, 1161-1164.

3. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature 586, 757-762.
4. Castel, S.E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., and Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat Genet 50, 1327-1334.
5. Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., Zhou, H., Tian, L., Prakash, O., Lemire, M., et al. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat Biotechnol 34, 531-538.
6. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nature Communications 9, 4619.
7. Drivas, T.G., Li, D., Nair, D., Alaimo, J.T., Alders, M., Altmüller, J., Barakat, T.S., Bebin, E.M., Bertsch, N.L., Blackburn, P.R., et al. (2020). A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome. European journal of human genetics : EJHG 28, 1422-1431.
8. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 47, D1018-d1027.
9. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12, e1001779.
10. Szustakowski, J.D., Balasubramanian, S., Sasson, A., Khalid, S., Bronson, P.G., Kvikstad, E., Wong, E., Liu, D., Davis, J.W., Haefliger, C., et al. (2020). Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank. medRxiv, 2020.2011.2002.20222232.
11. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature.
12. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203-209.
13. Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400-424.
14. Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19, 1523-1536.
15. Sadler, B., Wilborn, J., Antunes, L., Kuensting, T., Hale, A.T., Gannon, S.R., McCall, K., Cruchaga, C., Harms, M., Voisin, N., et al. (2021). Rare and de novo coding variants in chromodomain genes in Chiari I malformation. Am J Hum Genet 108, 100-114.
16. Haworth, S., Shapland, C.Y., Hayward, C., Prins, B.P., Felix, J.F., Medina-Gomez, C., Rivadeneira, F., Wang, C., Ahluwalia, T.S., Vrijheid, M., et al. (2019). Low-frequency variation in TP53 has large effects on head circumference and intracranial volume. Nature communications 10, 357.
17. Taal, H.R., Pourcain, B.S., Thiering, E., Das, S., Mook-Kanamori, D.O., Warrington, N.M., Kaakinen, M., Kreiner-Møller, E., Bradfield, J.P., Freathy, R.M., et al. (2012). Common variants at 12q15 and 12q24 are associated with infant head circumference. Nat Genet 44, 532-538.
18. Wright, C.F., West, B., Tuke, M., Jones, S.E., Patel, K., Laver, T.W., Beaumont, R.N., Tyrrell, J., Wood, A.R., Frayling, T.M., et al. (2019). Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. Am J Hum Genet 104, 275-286.
19. Jacquemont, S., Coe, B.P., Hersch, M., Duyzend, M.H., Krumm, N., Bergmann, S., Beckmann, J.S., Rosenfeld, J.A., and Eichler, E.E. (2014). A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. Am J Hum Genet 94, 415-425.

20. Duyzend, M.H., Nuttle, X., Coe, B.P., Baker, C., Nickerson, D.A., Bernier, R., and Eichler, E.E. (2016). Maternal Modifiers and Parent-of-Origin Bias of the Autism-Associated 16p11.2 CNV. Am J Hum Genet 98, 45-57.
21. Porter, R.S., Jaamour, F., and Iwase, S. (2018). Neuron-specific alternative splicing of transcriptional machineries: Implications for neurodevelopmental disorders. Mol Cell Neurosci 87, 35-45.
22. Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R.A., McConnell, J.S., Angle, B., Meschino, W.S., et al. (2012). Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med 367, 1321-1331.
23. Albers, C.A., Paul, D.S., Schulze, H., Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet 44, 435-439, s431-432.
24. Low, K., Ashraf, T., Canham, N., Clayton-Smith, J., Deshpande, C., Donaldson, A., Fisher, R., Flinter, F., Foulds, N., Fryer, A., et al. (2016). Clinical and genetic aspects of KBG syndrome. Am J Med Genet A 170, 2835-2846.
25. Katsanis, N. (2016). The continuum of causality in human genetic disorders. Genome Biol 17, 233.
26. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405-424.
27. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47, D886-d894.
28. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature methods 7, 248-249.
29. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nat Protoc 11, 1-9.
30. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434-443.
31. Dingemans, A.J.M., Stremmelaar, D.E., van der Donk, R., Vissers, L., Koolen, D.A., Rump, P., Hehir-Kwa, J.Y., and de Vries, B.B.A. (2021). Quantitative facial phenotyping for Koolen-de Vries and 22q11.2 deletion syndrome. Eur J Hum Genet.
32. van der Donk, R., Jansen, S., Schuurs-Hoeijmakers, J.H.M., Koolen, D.A., Goltstein, L., Hoischen, A., Brunner, H.G., Kemmeren, P., Nellåker, C., Vissers, L., et al. (2019). Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. Genet Med 21, 1719-1725.
33. Reijnders, M.R.F., Miller, K.A., Alvi, M., Goos, J.A.C., Lees, M.M., de Burca, A., Henderson, A., Kraus, A., Mikat, B., de Vries, B.B.A., et al. (2018). De Novo and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype-Phenotype Evaluation of a Distinct Neurodevelopmental Disorder. Am J Hum Genet 102, 1195-1203.
34. Deng, Y., Gao, L., Wang, B., and Guo, X. (2015). HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PLoS One 10, e0115692.
35. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., and Chen, C.F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274-1281.
36. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J Mol Graph 8, 52-56, 29.
37. Krieger, E., Koraimann, G., and Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins 47, 393-402.
38. Farnung, L., Ochmann, M., and Cramer, P. (2020). Nucleosome-CHD4 chromatin remodeler structure maps human disease mutations. Elife 9.
39. Mansfield, R.E., Musselman, C.A., Kwan, A.H., Oliver, S.S., Garske, A.L., Davrazou, F., Denu, J.M., Kutateladze, T.G., and Mackay, J.P. (2011). Plant homeodomain (PHD) fingers of CHD4 are histone H3-binding modules with preference for unmodified H3K4 and methylated H3K9. J Biol Chem 286, 11779-11791.
40. Koulouras, G., Panagopoulos, A., Rapsomaniki, M.A., Giakoumakis, N.N., Taraviras, S., and

Lygerou, Z. (2018). EasyFRAP-web: a web-based tool for the analysis of fluorescence recovery after photobleaching data. Nucleic Acids Res 46, W467-w472.

41. Kohze, R., Dieteren, C.E.J., Koopman, W.J.H., Brock, R., and Schmidt, S. (2017). Frapbot: An open-source application for FRAP data. Cytometry A 91, 810-814.

42. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 11, e1004219.

43. Wiel, L., Baakman, C., Gilissen, D., Veltman, J.A., Vriend, G., and Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. Hum Mutat 40, 1030-1038.

Supplemental information

- Figure S1	Amino acid sequence alignments of the CHD3 protein.
- Figure S2	Childhood images of heterozygote parents with a *CHD3* variant.
- Figure S3	t-SNE plot of the distribution of individuals with *de novo* and inherited *CHD3* variants versus controls with intellectual disability.
- Figure S4	Grouped HPO features based on semantic similarity and clustering results per individual.
- Figure S5	Parental inheritance of inherited *CHD3* variants and sex distribution of healthy individuals with *CHD3* loss-of-function variants.
- Figure S6	*In silico* prediction of pathogenicity scores for *CHD3* variants in NDD and CM1.
- Figure S7	Functional characterisation of inherited *CHD3* missense variants.

- Table S1	Summary of phenotypes seen in individuals with CHD3 variants, including separate characteristics for individuals with inherited LoF and missense variants.
- Table S2	Primers to clone the CHD3 C-terminal construct and GATAD2B.
- Table S3	Primers used for site-directed mutagenesis.
- Table S4	Primers to amplify the region that carries the *CHD3* p.W1158* stop-gain variant to test for NMD.
- Table S5	Primers used for qPCR.

p.S477F

```
TR|F6PUX9|F6PUX9_XENTR CCDACVSSYHIHCLNPPLPDIPHGEWLCPRCTCPQLKGKVQKILHWRWGVPPEGVPLPQP 550
SP|Q12873|CHD3_HUMAN   CCDACISSYHIHCLNPPLPDIPNGEWLCPRCTCPVLKGRVQKILHWRWGEPPVAVPAPQQ 529
TR|B1AR17|B1AR17_MOUSE CCDACISSYHIHCLNPPLPDIPNGEWLCPRCTCPVLKGRVQKILHWRWGEPPVAVPAPQQ 581
TR|F1LPP8|F1LPP8_RAT   CCDACISSYHIHCLNPPLPDIPNGEWLCPRCTCPVLKGRVQKILHWRWGEPPVAVPAPQQ 361
                       *****:****************:*********** ***:********** ** .** **
```

pW630R

```
TR|F6PUX9|F6PUX9_XENTR PPPFDYGSGEDEGKSEK--SKDAEYSDLEERFYRYGIKPEWMSIQRIINHSLDKRGNYHY 668
SP|Q12873|CHD3_HUMAN   PPPLDYGSGEDDGKSDKRKVKDPHYAEMEEKYYRFGIKPEWMTVHRIINHSVDKKGNYHY 649
TR|B1AR17|B1AR17_MOUSE PPPLDYGSGEDDGKSDKRKVKDPHYAEMEEKYYRFGIKPEWMTVHRIINHSMDKKGNYHY 701
TR|F1LPP8|F1LPP8_RAT   PPPLDYGSGEDDGKSDKRKVKDPHYAEMEEKYYRFGIKPEWMTVHRIINHSMDKKGNYHY 481
                       ***:*******:***:*   ** .*:::**::**:*******:::******:**:*****
```

p.F833L p.G873S

```
TR|F6PUX9|F6PUX9_XENTR EFSYQDNVMKGGKKAFKMKAQAQVKFHVLLTSYELVTIDQAALGSIRWACLVVDEAHRLK 908
SP|Q12873|CHD3_HUMAN   EFSFEDNAIKGGKKAFKMKREAQVKFHVLLTSYELITIDQAALGSIRWACLVVDEAHRLK 889
TR|B1AR17|B1AR17_MOUSE EFSFEDNAIKGGKKAFKMKREAQVKFHVLLTSYELITIDQAALGSIRWACLVVDEAHRLK 941
TR|F1LPP8|F1LPP8_RAT   EFSFEDNAIKGGKKAFKMKREAQVKFHVLLTSYELITIDQAALGSIRWACLVVDEAHRLK 721
                       ***::**.:********** :**************:************************
```

p.I983V

```
TR|F6PUX9|F6PUX9_XENTR EDQIKKLHDLLGPHLLRRMKADVFKNMPAKTELIVRVELSPMQKKYYKFILTRNFEALNS 1028
SP|Q12873|CHD3_HUMAN   EDQIKKLHDLLGPHMLRRLKADVFKNMPAKTELIVRVELSPMQKKYYKYILTRNFEALNS 1009
TR|B1AR17|B1AR17_MOUSE EDQIKKLHDLLGPHMLRRLKADVFKNMPAKTELIVRVELSPMQKKYYKYILTRNFEALNS 1061
TR|F1LPP8|F1LPP8_RAT   EDQIKKLHDLLGPHMLRRLKADVFKNMPAKTELIVRVELSPMQKKYYKYILTRNFEALNS 841
                       **************:***:*****************************:***********
```

p.P1046L

```
TR|F6PUX9|F6PUX9_XENTR RGGGNQVSLLNIMMDLKKCCNHPYLFPAASLESPKLPSGAYEGGSLVKASGKLLLLHKML 1088
SP|Q12873|CHD3_HUMAN   RGGGNQVSLLNIMMDLKKCCNHPYLFPVAAMESPKLPSGAYEGGALIKSSGKLMLLQKML 1069
TR|B1AR17|B1AR17_MOUSE RGGGNQVSLLNIMMDLKKCCNHPYLFPVAAMESPKLPSGAYEGGALIKSSGKLLLLQKML 1121
TR|F1LPP8|F1LPP8_RAT   RGGGNQVSLLNIMMDLKKCCNHPYLFPVAAMESPKLPSGAYEGGALIKSSGKLLLLQKML 901
                       ***************************.*::*************:*:*:****:**:***
```

p.P1125A

```
TR|F6PUX9|F6PUX9_XENTR RKLNEQGHRVLIFSQMTKMLDILEDFLDFEGYKYERIDGGITGALRQEAIDRFNAPGAQQ 1148
SP|Q12873|CHD3_HUMAN   RKLKEQGHRVLIFSQMTKMLDLLEDFLDYEGYKYERIDGGITGALRQEAIDRFNAPGAQQ 1129
TR|B1AR17|B1AR17_MOUSE RKLKEQGHRVLIFSQMTKMLDLLEDFLDYEGYKYERIDGGITGALRQEAIDRFNAPGAQQ 1181
TR|F1LPP8|F1LPP8_RAT   RKLKEQGHRVLIFSQMTKMLDLLEDFLDYEGYKYERIDGGITGALRQEAIDRFNAPGAQQ 961
                       ***:*****************:******:*******************************
```

p.R1342Q

```
TR|F6PUX9|F6PUX9_XENTR QEENVDPDYWEKLLRHHYEQQQEDLARNLGKGKRVRKQVNYNDAAQEDQDNQSEYSVGSE 1388
SP|Q12873|CHD3_HUMAN   QEENVDPDYWEKLLRHHYEQQQEDLARNLGKGKRVRKQVNYNDAAQEDQDNQSEYSVGSE 1368
TR|B1AR17|B1AR17_MOUSE QEENVDPDYWEKLLRHHYEQQQEDLARNLGKGKRVRKQVNYNDAAQEDQDNQSEYSVGSE 1420
TR|F1LPP8|F1LPP8_RAT   QEENVDPDYWEKLLRHHYEQQQEDLARNLGKGKRVRKQVNYNDAAQEDQDNQSEYSVGSE 1200
                       ************************************************************
```

p.R1706Q

```
TR|F6PUX9|F6PUX9_XENTR PRFMFNIADGGFTELHTLWQNEERAAICSGRLNDIWHRRHDYWLLAGIVVHGYARWQDIQ 1807
SP|Q12873|CHD3_HUMAN   PRFMFNIADGGFTELHTLWQNEERAAISSGKLNEIWHRRHDYWLLAGIVLHGYARWQDIQ 1764
TR|B1AR17|B1AR17_MOUSE PRFMFNIADGGFTELHTLWQNEERAAISSGKLNEIWHRRHDYWLLAGIVLHGYARWQDIQ 1817
TR|F1LPP8|F1LPP8_RAT   PRFMFNIADGGFTELHTLWQNEERAAISSGKLNEIWHRRHDYWLLAGIVLHGYARWQDIQ 1597
                       ***************************.**:**:***************:**********
```

p.R1759Q

```
TR|F6PUX9|F6PUX9_XENTR PRFMFNIADGGFTELHTLWQNEERAAICSGRLNDIWHRRHDYWLLAGIVVHGYARWQDIQ 1807
SP|Q12873|CHD3_HUMAN   PRFMFNIADGGFTELHTLWQNEERAAISSGKLNEIWHRRHDYWLLAGIVLHGYARWQDIQ 1764
TR|B1AR17|B1AR17_MOUSE PRFMFNIADGGFTELHTLWQNEERAAISSGKLNEIWHRRHDYWLLAGIVLHGYARWQDIQ 1817
TR|F1LPP8|F1LPP8_RAT   PRFMFNIADGGFTELHTLWQNEERAAISSGKLNEIWHRRHDYWLLAGIVLHGYARWQDIQ 1597
                       ***************************.**:**:***************:**********
```

p.E1837K

```
TR|F6PUX9|F6PUX9_XENTR PAMALNARFSEVECLAESHQHLSKESLAGNKPANAVLHKVLNQLEELLSDMKADVTRLPA 1927
SP|Q12873|CHD3_HUMAN   PAMALHARFAEAECLAESHQHLSKESLAGNKPANAVLHKVLNQLEELLSDMKADVTRLPA 1884
TR|B1AR17|B1AR17_MOUSE PAMALHARFAEAECLAESHQHLSKESLAGNKPANAVLHKVLNQLEELLSDMKADVTRLPA 1937
TR|F1LPP8|F1LPP8_RAT   PAMALHARFAEAECLAESHQHLSKESLAGNKPANAVLHKVLNQLEELLSDMKADVTRLPA 1717
                       *****:***:*.************************************************
```

p.R1888Q

```
TR|F6PUX9|F6PUX9_XENTR SLARIPPIAARLQMSERSILSRLASKGSESTAPPVFPPGPYTTPPMFGATFPSNPQSA-A 1986
SP|Q12873|CHD3_HUMAN   TLSRIPPIAARLQMSERSILSRLASKGTEPHPTPAYPPGPYATPPGYGAAFSAAPVGALA 1944
TR|B1AR17|B1AR17_MOUSE TLSRIPPIAARLQMSERSILSRLASKGTEPHPTPAFPPGPYATPPGYGAAFSAAPVGALA 1997
TR|F1LPP8|F1LPP8_RAT   TLSRIPPIAARLQMSERSILSRLASKGTEPHPTPAFPPGPYATPPGYGAAFSAAPVGALA 1777
                       :*:************************:*    *.:*****:*** :**:* : * .* *
```

p.K1972T

```
TR|F6PUX9|F6PUX9_XENTR LTGTNYNQMPLGSFLSA-SNGPPVLVKKEREMDL--------LDRKESRGGEVICIDD 2035
SP|Q12873|CHD3_HUMAN   AAGANYSQMPAGSFITAATNGPPVLVKKEKEMVGALVSDG--LDRKEPRAGEVICIDD 2000
TR|B1AR17|B1AR17_MOUSE AAGANYSQMPAGSFITATTNGPPVLVKKEKEMVGALVSDGLGLDRKEPRAGEVICIDD 2055
TR|F1LPP8|F1LPP8_RAT   AAGANYSQMPAGSFITATTNGPPVLVKKEKEMVGALVSDGLGLDRKEPRAGEVICIDD 1835
                        :*:**.*** ***::* :**********:**          ***** *.********
```

Figure S1. Amino acid sequence alignments of the CHD3 protein. Amino acid sequences of the human CHD3 protein (Q01826, UniProt) aligned to the *Xenopus tropicalis* (F6PUX9), mouse (B1AR17) and rat (F1LPP8) sequences. Alignment was performed with Clustal Omega (1.2.4) with default settings using EMBL-EBI alignment tool. Missense variants described in this study are shaded in red.

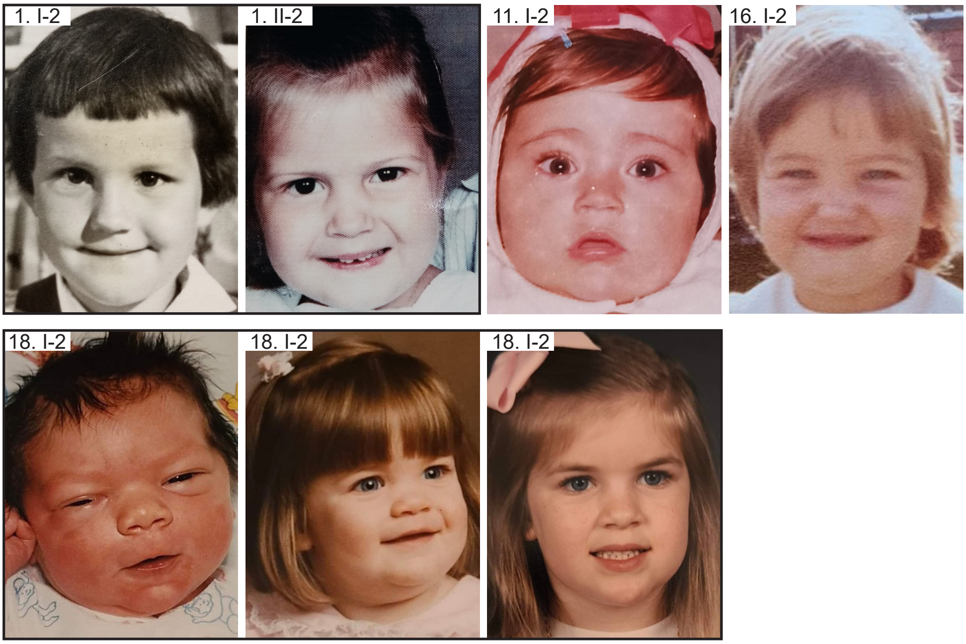


Figure S2. Childhood images of heterozygote parents with a *CHD3* variant. Facial photographs taken during childhood of heterozygote parents with a CHD3 variant. Individuals 1. I-2, 1. II-2 and 16. I-2 are four years old in the picture, individual 11. I-2 is one year old. Individual 18. I-2, from left to right, as infant, toddler and child. Images show several features also seen in probands with *CHD3* variants including a squared face, pointed chin, deep-set eyes and a broad nasal bridge. These features become less prominent when these individuals get older (compare to Figure 2A in the main text).

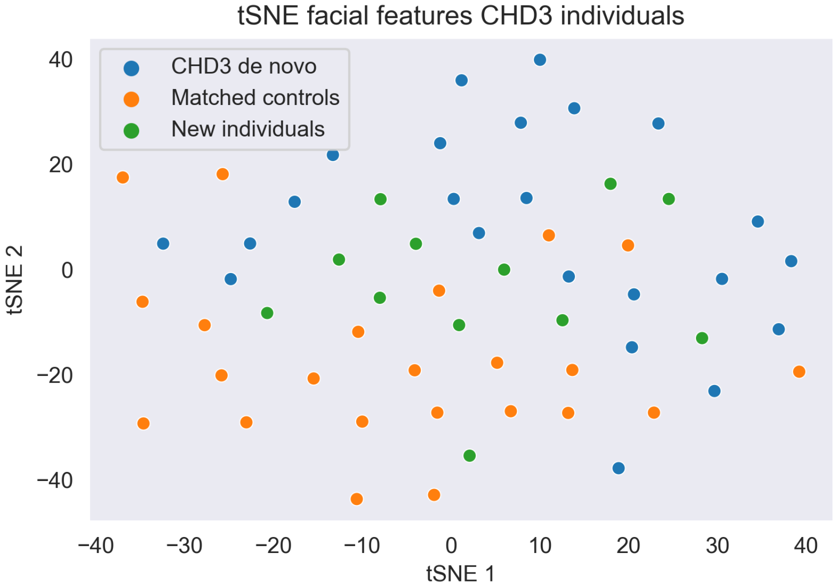
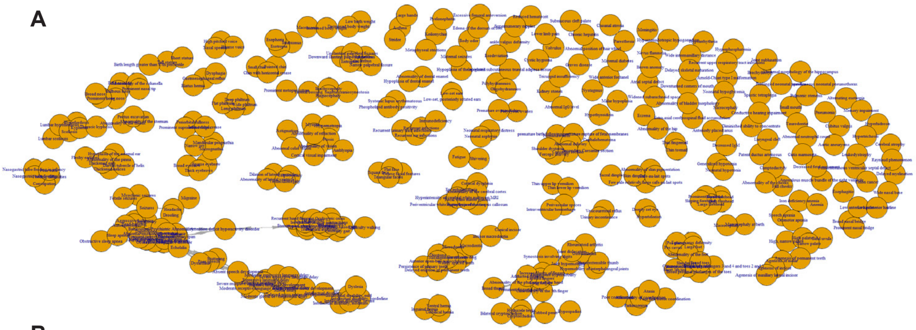


Figure S3. t-SNE plot of the distribution of individuals with *de novo* and inherited *CHD3* variants versus controls with intellectual disability. t-distributed stochastic neighbour embeddings (t-SNE) plot that visualises the distribution of the hybrid feature vectors for individuals with *de novo* ('CHD3 de novo'; blue) and inherited *CHD3* variants ('New individuals'; green), and intellectual disability (ID) controls ('Matched controls'; orange). Relative clustering of individuals with a *de novo* variant in *CHD3* suggests that this group of individuals with intellectual disability shows more similarity in facial features than is expected by chance (AROC: 0.91). The control group was matched to the *de novo CHD3* group for sex, ethnicity, and age. Individuals with an inherited *CHD3* variant were tested for clustering with either the *de novo CHD3* group, or the ID controls group, based on the nearest-neigbour principle.



B

Cluster	Labels	Index	Correct classification
1	0	CHD3_dn_001	FALSE
0	0	CHD3_dn_002	TRUE
1	0	CHD3_dn_003	FALSE
0	0	CHD3_dn_004	TRUE
0	0	CHD3_dn_005	TRUE
1	0	CHD3_dn_006	FALSE
0	0	CHD3_dn_007	TRUE
0	0	CHD3_dn_008	TRUE
0	0	CHD3_dn_009	TRUE
0	0	CHD3_dn_010	TRUE
0	0	CHD3_dn_011	TRUE
1	0	CHD3_dn_012	FALSE
0	0	CHD3_dn_013	TRUE
0	0	CHD3_dn_014	TRUE
0	0	CHD3_dn_015	TRUE
0	0	CHD3_dn_016	TRUE
0	0	CHD3_dn_017	TRUE
0	0	CHD3_dn_018	TRUE
1	0	CHD3_dn_019	FALSE
0	0	CHD3_dn_020	TRUE
0	0	CHD3_dn_021	TRUE
1	0	CHD3_dn_022	FALSE
0	0	CHD3_dn_023	TRUE
0	0	CHD3_dn_024	TRUE
0	0	CHD3_dn_025	TRUE
0	0	CHD3_dn_026	TRUE
0	0	CHD3_dn_027	TRUE
1	0	CHD3_dn_028	FALSE
1	0	CHD3_dn_029	FALSE
0	0	CHD3_dn_030	TRUE
1	0	CHD3_dn_031	FALSE
0	0	CHD3_dn_032	TRUE
1	0	CHD3_dn_033	FALSE
0	0	CHD3_dn_034	TRUE
0	0	CHD3_dn_035	TRUE
0	1	CHD3_inh_001	FALSE
1	1	CHD3_inh_002	TRUE
0	1	CHD3_inh_003	FALSE
0	1	CHD3_inh_004	FALSE
0	1	CHD3_inh_005	FALSE
0	1	CHD3_inh_007	FALSE
0	1	CHD3_inh_008	FALSE
0	1	CHD3_inh_009	FALSE
0	1	CHD3_inh_010	FALSE
0	1	CHD3_inh_011	FALSE
0	1	CHD3_inh_012	FALSE
1	1	CHD3_inh_013	TRUE
1	1	CHD3_inh_014	TRUE
0	1	CHD3_inh_015	FALSE
1	1	CHD3_inh_016	TRUE
1	1	CHD3_inh_017	TRUE
1	1	CHD3_inh_018	TRUE
0	1	CHD3_inh_019	FALSE
0	1	CHD3_inh_020	FALSE
0	1	CHD3_inh_021	FALSE

Cluster	Labels	Index	Correct classification
0	0	CHD3_inh_001	TRUE
1	1	CHD3_inh_001_p	TRUE
0	0	CHD3_inh_002	TRUE
1	1	CHD3_inh_002_p	TRUE
0	0	CHD3_inh_003	TRUE
1	1	CHD3_inh_003_p	TRUE
1	0	CHD3_inh_004	FALSE
1	1	CHD3_inh_004_p	TRUE
0	0	CHD3_inh_005	TRUE
1	1	CHD3_inh_005_p	TRUE
0	0	CHD3_inh_007	TRUE
1	1	CHD3_inh_007_p	TRUE
1	0	CHD3_inh_008	FALSE
1	1	CHD3_inh_008_p	TRUE
1	0	CHD3_inh_009	FALSE
1	1	CHD3_inh_009_p	TRUE
0	0	CHD3_inh_010	TRUE
1	1	CHD3_inh_010_p	TRUE
0	0	CHD3_inh_011	TRUE
1	1	CHD3_inh_011_p	TRUE
1	0	CHD3_inh_012	FALSE
1	1	CHD3_inh_012_p	TRUE
0	0	CHD3_inh_013	TRUE
0	1	CHD3_inh_013_p	FALSE
0	0	CHD3_inh_014	TRUE
1	1	CHD3_inh_014_p	TRUE
1	0	CHD3_inh_015	FALSE
1	1	CHD3_inh_015_p	TRUE
0	0	CHD3_inh_016	TRUE
1	1	CHD3_inh_016_p	TRUE
0	0	CHD3_inh_017	TRUE
1	1	CHD3_inh_017_p	TRUE
0	0	CHD3_inh_018	TRUE
1	1	CHD3_inh_018_p	TRUE
1	0	CHD3_inh_019	FALSE
1	1	CHD3_inh_019_p	TRUE
0	0	CHD3_inh_020	TRUE
1	1	CHD3_inh_020_p	TRUE
0	0	CHD3_inh_021	TRUE
1	1	CHD3_inh_021_p	TRUE

Figure S4 Grouped HPO features based on semantic similarity and clustering results per individual. **A**) The semantic similarity between all the HPO terms used in this cohort was calculated using the Wang algorithm in the HPOsim package in R. HPO terms with at least a 0.5 similarity score were grouped and a new feature was created as a replacement, which was the sum of the grouped features. **B**) Individual HPO-based phenotypic clustering results for analyses comparing individuals with *de novo* (dn under 'Index') and inherited variants (inh under 'Index'; left), and probands and heterozygote parents (indicated with a 'p' under 'Index'; right). Index numbers match the family numbering of the cohort.

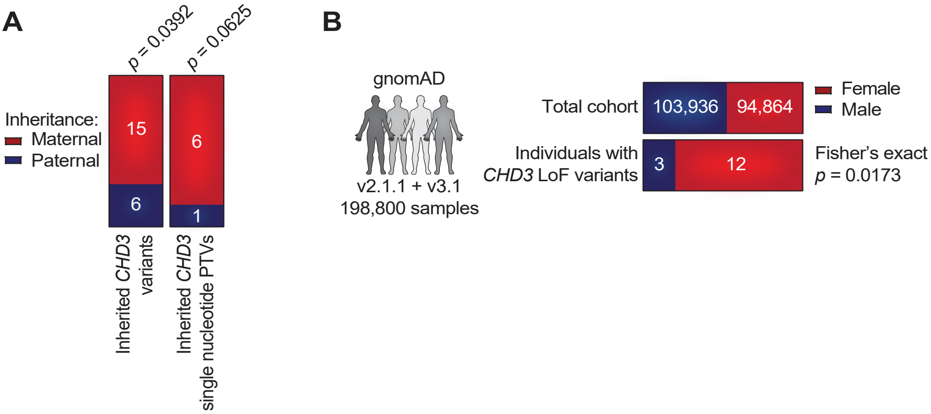
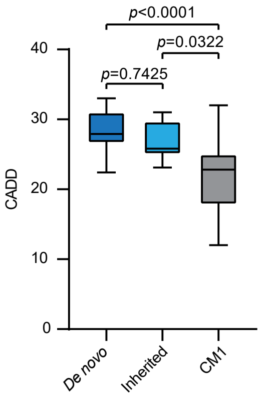


Figure S5. Parental inheritance of inherited *CHD3* variants and sex distribution of healthy individuals with *CHD3* loss-of-function variants. A) Inheritance of all inherited *CHD3* variants (left), and of only inherited *CHD3* single nucleotide PTVs (right), in families with a proband with NDD (one-sided binomial test with expected ratios of 0.5 for paternal inheritance and 0.5 for maternal inheritance. The *p*-value is the probability of the found number of maternally inherited cases or more). **B**) Sex distribution of the GnomAD cohort (top) and of individuals with *CHD3* LoF variants in the GnomAD cohort (bottom; *p*-value based on a two-sided Fisher's exact test). Only individuals with stop-gain and frameshift variants were included. Individuals with first and last exon variants were excluded, as well as individuals with pLoF-flagged variants and mosaic cases.



De novo CHD3 missense variants	CADD v1.6
p.Gln569Arg	33
p.His886Arg	25.9
p.Leu915Phe	23.7
p.Asn917Tyr	28.6
p.Glu921Lys	29.8
p.Phe944Tyr	26.6
p.Ser948Pro	27.3
p.Gly961Glu	27.9
p.Arg966Trp	30
p.Arg966Pro	31
p.Lys969Glu	27.3
p.Arg985Trp	29.8
p.Arg985Gln	32
p.Arg1121Pro	31
p.Thr1136Ile	26.6
p.Trp1158Arg	27.9
p.Asn1159Lys	22.8
p.His1161Arg	24.9
p.Arg1169Trp	26.7
p.His1171Arg	26.6
p.Arg1172Gln	32
p.Leu1080His	28.5
p.Arg1187Pro	27.2
p.Leu1236Pro	28.5
p.Arg1262Trp	26.8
p.Arg1342Gln	31
p.Arg1415Cys	32
p.Arg1881Leu	31
p.Ala1955Ser	22.4
Mean	28.23
Standard deviation	2.81

Inherited CHD3 missense variants	CADD v1.6
p.Ser477Phe	25.8
p.Trp630Arg	29.5
p.Phe833Leu	24.9
p.Gly873Ser	23.1
p.Ile983Val	25.1
p.Pro1046Leu	31
p.Pro1125Ala	25.6
p.Arg1342Gln	31
p.Arg1706Gln	25.3
p.Arg1759Gln	29.9
p.Glu1837Lys	29.5
p.Arg1888Gln	27.6
p.Lys1972Thr	23.1
Mean	27.03
Standard deviation	2.86

Rare CHD3 missense variants associated with CM1	CADD v1.6
p.Cys17Tyr*	18.81
p.Arg24Trp*	22.8
p.Ala25Val*	16.31
p.Lys114Thr	19.58
p.Arg271Gln	25.7
p.Arg337Trp	25
p.Tyr595Cys	23.6
p.Gly733Arg	32
p.Arg876Cys	23.5
p.Asp1358Glu	16.97
p.Arg1600Gln	22.8
p.Val1624Leu	12.02
p.Ala1937Thr	17.83
p.Lys1972Glu	23.9
p.Cys1997Tyr	26.8
Mean	21.84
Standard deviation	4.97

Figure S6. *In silico* prediction of pathogenicity scores for *CHD3* variants in NDD and CM1. CADD scores (v1.6) for *de novo CHD3* variants in NDD, described in Snijders Blok et al. 2018 and Drivas et al. 2020, for inherited *CHD3* variants identified in the current study, and for rare missense variants associated with Chiari I malformation (CM1) described in Saddler et al. 2021 were obtained, and plotted as box plots, showing the median CADD scores for each group. CADD scores for *de novo* and inherited *CHD3* missense variants in NDD are significantly higher compared to CADD scores of missense variants associated with CM1 without NDD, while *de novo* and inherited *CHD3* missense variants in NDD do not have significantly different CADD scores (Kruskal-Wallis test followed by a Dunn's multiple comparisons test, NS: not significant). In the table, positions are with reference to sequence ENST00000380358.4 (NM_001005273.2), variants marked with an asterisk are with reference to sequence ENST00000481999.1.

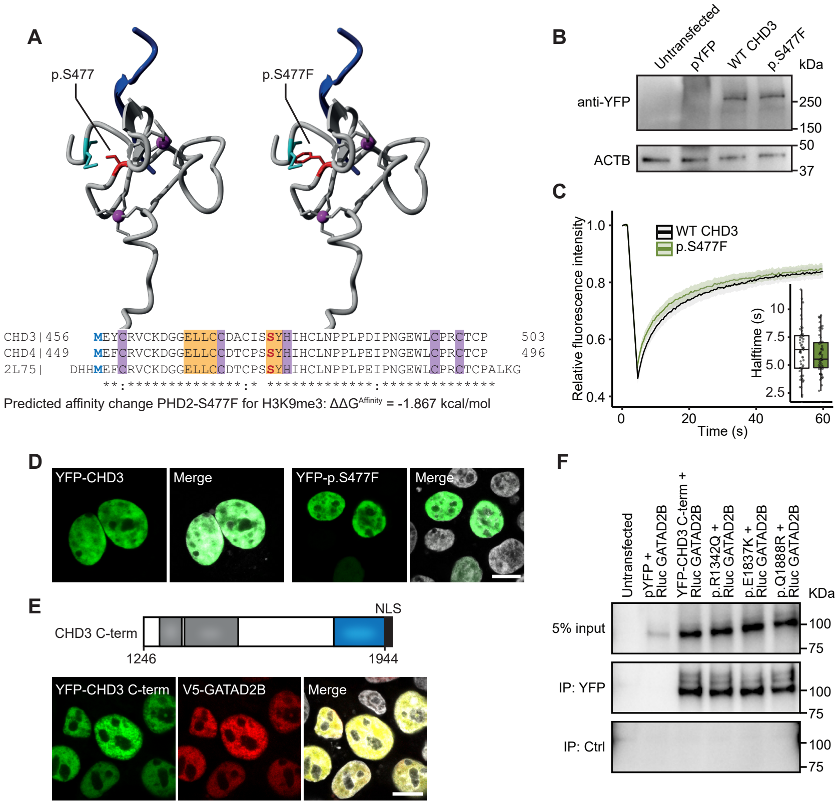


Figure S7. Functional characterisation of inherited CHD3 missense variants. A) Three-dimensional modelling of the CHD3 p.S477F variant in the PHD2 domain of CHD4 (PDB 2L75, 89% sequence identity with CHD3). In both the protein model and the alignment, the affected residue is depicted in red, and p.M456, that forms a hydrogen bond with p.S477, is shown in cyan. Zinc metals, and zinc binding residues are purple, and H3K9me3 is shown in dark blue. Residues that form a beta-sheet are shaded in yellow in the alignment. The effect of p.S477F on binding of PHD2 to H3K9me3 was assessed using the mCSM-PPI2 machine learning tool, with $\Delta\Delta G^{\text{Affinity}} = \Delta\Delta G_{\text{wt-mt}} = \Delta G_{\text{wildtype}} - \Delta G_{\text{mutant}}$. The p.S477F variant is predicted to impact CHD3 histone binding. **B**) Immunoblot of whole-cell lysates expressing YFP-tagged CHD3 and the p.S477F variant probed with anti-EGFP antibody, showing bands at the expected molecular weight of ~260 kDa. The blot was probed for ACTB to ensure equal protein loading. **C**) FRAP experiments to assess the dynamics of CHD3 chromatin binding in live cells. Graph shows the mean recovery curves ± 95% C.I. recorded in HEK293T/17 cells expressing YFP-CHD3 fusion proteins. Right corner, box plot of the halftime based on single-term exponential curve fitting of individual recordings (n = 55 nuclei from three independent experiments, no significant difference, Student's t-test). The protein mobility of the p.S477F variant in the nucleus is not significantly different from WT protein, suggesting no loss of chromatin binding. **D**) Direct fluorescence micrographs of HEK293T/17 cells expressing YFP-CHD3 fusion proteins (green). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 μm. The p.S477F variant localised to the nucleus similar to WT protein. **E**) Top: Schematic of the cCHD3-NLS construct used for protein-interaction assays, encoding residues 1246-1944 including the CHDCT2 domain (blue), and an SV40-NLS appended at the C-terminus (black). Bottom: Direct fluorescence micrographs of HEK293T/17 cells co-expressing YFP-CHD3 (green) and V5-GATAD2B (red). Nuclei were stained with Hoechst 33342 (white). Scale bar = 10 μm. **F**) Co-immunoprecipitation assay to study the CHD3-GATAD2B interaction. YFP-tagged C-terminal truncations of CHD3 (WT, p.R1342Q, p.E1837K, p.R1888Q) were coexpressed with Rluc-GATAD2B in HEK293T/17 cells. YFP-fusion proteins were immobilised on αGFP affinity beads and used as baits to pull down the coexpressed Rluc-GATAD2B. Lysates of untransfected cells and cells co-transfected with pYFP and Rluc-GATAD2B were used as negative controls. As binding control, lysates were incubated with deactivated beads. All tested CHD3 C-terminal variants were still able to pull-down GATAD2B similar to WT protein, suggesting that they do not disrupt the CHD3-GATAD2B interaction.

Table S1. Summary of phenotypes seen in individuals with *CHD3* variants, including separate characteristics for individuals with inherited LoF and missense variants

	Probands with de novo variant[1]	Probands with inherited variant	Heterozygote parents	Proband with inherited LoF variants	Proband with inherited missense variants
Development					
Developmental delay	100% (55/55)	100% (21/21)	17% (3/18)	100% (8/8)	100% (13/13)
Intellectual disability	98% (46/47)	79% (11/14)	21% (4/19)	80% (4/5)	78% (7/9)
borderline/borderline-mild	6% (3/47)	14% (2/14)	11% (2/18)	0% (0/5)	22% (2/9)
mild/mild-moderate	30% (14/47)	29% (4/14)	6% (1/18)	40% (2/5)	22% (2/9)
moderate/moderate-severe	36% (17/47)	14% (2/14)	0% (0/18)	0% (0/5)	22% (2/9)
severe	23% (11/47)	0% (0/14)	0% (0/18)	0% (0/5)	0% (0/9)
level unknown	2% (1/47)	21% (3/14)	6% (1/18)	40% (2/5)	11% (1/9)
Speech delay/disorder	100% (53/53)	100% (20/20)	24% (4/17)	100% (8/8)	100% (12/12)
Autism or autism-like features	35% (18/51)	53% (10/19)	18% (3/17)	43% (3/7)	58% (7/12)
Neurology					
Hypotonia	81% (39/48)	89% (17/19)	17% (2/12)	86% (6/7)	92% (11/12)
Macrocephaly	53% (28/53)	47% (9/19)	50% (8/16)	33% (2/6)	54% (7/13)
CNS abnormalities	50% (24/48)	62% (8/13)	25% (1/4)	33% (1/3)	70% (7/10)
Neonatal feeding problems	31% (10/32)	21% (4/19)	6% (1/16)	14% (1/7)	25% (3/12)
Facial dysmorphisms					
High/broad/prominent forehead	85% (28/33)	85% (17/20)	53% (9/17)	86% (6/7)	85% (11/13)
Thin upper lip	79% (15/19)	55% (11/20)	47% (8/17)	57% (4/7)	54% (7/13)
Widely spaced eyes	69% (35/51)	70% (14/20)	24% (4/17)	71% (5/7)	69% (9/13)
Broad nasal bridge	75% (15/20)	80% (16/20)	24% (4/17)	86% (6/7)	77% (10/13)
Full cheeks	58% (11/19)	70% (14/20)	13% (2/16)	57% (4/7)	77% (10/13)
Pointed chin	60% (12/20)	53% (10/19)	41% (7/17)	67% (4/6)	46% (6/13)
Deep-set eyes	55% (11/20)	47% (9/19)	50% (8/16)	50% (3/6)	46% (6/13)
Other					
Joint laxity (generalized and/or local)	36% (18/50)	40% (8/20)	29% (4/14)	43% (3/7)	38% (5/13)
Vision problems	72% (38/53)	25% (5/20)	53% (8/15)	14% (1/7)	31% (4/13)
Male genital abnormalities	32% (8/25)	22% (2/9)	0% (0/4)	50% (1/2)	14% (1/7)
Hernia (umbilical, inguinal, hiatal)	13% (6/48)	10% (2/20)	0% (0/14)	0% (0/7)	15% (2/13)

[1] Combined cases from Snijders Blok et al. 2018 and Drivas et al. 2020 (confirmed de novo only)

CNS: central nervous system. LoF: loss of function.

Table S2. Primers to clone the CHD3 C-terminal construct and GATAD2B. Underscored italic sequence marks the restriction sites (*HindIII*/*BamHI*) and the bold sequence is an SV40 NLS sequence that was added to ensure nuclear localisation of the encoded protein.

CHD3-Cterm-cloning-F1	5'-GAGGGG*AAGCTT*AAAGGAGGAGGACAGCAGTGT-3'
CHD3-Cterm-cloning-R1+NLS	5'-TGCAGG*GGATCC*TCA**GACCTTGCGCTTCTTCTT**CGGGGCCAGGGCCCCTACGGGTG-3'
GATAD2B-cloning-F1	5'-*GGATCC*TGGATAGAATGACAGAAGATGC-3'
GATAD2B-cloning-R1	5'-*TCTAGA*TTATTTCTGTCCACTGATGG-3'

Table S3. Primers used for site-directed mutagenesis.

CHD3 p.S477F F	5'-GACAATGAATGTGGTAGAAGGAGATGCACGCGTCA-3'
CHD3 p.S477F R	5'-TGACGCGTGCATCTCCTTCTACCACATTCATTGTC-3'
CHD3 p.R1342Q F	5'-CTTGCTTGCGAACCTGCTTGCCCTTGCCT-3'
CHD3 p.R1342Q R	5'-AGGCAAGGGCAAGCAGGTTCGCAAGCAAG-3'
CHD3 p.E1837K F	5'-GGCCAGGCACTTGGCCTCGGCGA-3'
CHD3 p.E1837K R	5'-TCGCCGAGGCCAAGTGCCTGGCC-3'
CHD3 p.R1888Q F	5'-GATGGGGGGTATTTGGGACAGCGTGGC-3'
CHD3 p.R1888Q R	5'-GCCACGCTGTCCCAAATACCCCCCATC-3'

Table S4. Primers to amplify the region that carries the *CHD3* p.W1158* stop-gain variant to test for NMD.

CHD3-NMD-W1158*-F	5'-TCGGTTTAATGCTCCTGGGG-3'
CHD3-NMD-W1158*-R	5'-ACAACCAGGTGTGTCAGCAT-3'

Table S5. Primers used for qPCR.

CHD3-F	5'-AGGAAGACCAAGACAACCAGTCAG-3'
CHD3-R	5'-TGACTGTCTACGCCCTTCAGGA-3'
TBP-F	5'-GGGCACCACTCCACTGTATC-3'
TBP-R	5'-CGAAGTGCAATGGTCTTTAGG-3'
PPIA-F	5'-TATCTGCACTGCCAAGACTGAGTG-3'
PPIA-R	5'-CTTCTTGCTGGTCTTGCCATTCC-3'

Supplemental notes 1

Cohort

We identified inherited variants in *CHD3* likely explaining the phenotype in 21 probands (12 females/9 males) who were evaluated for syndromic developmental delay using next generation sequencing. The mean age of the probands at diagnosis was 5 years and 7 months (range 12 months – 17 years). The probands presented with developmental delay (21/21, 100%), speech delay (20/20, 100%), autism spectrum disorder or autism-like features (10/19, 53%), hypotonia (17/19, 89%), and facial dysmorphisms including a broad/prominent forehead (17/20, 85%) (Figure 2; Table 1, S1 and S2). These features and frequencies were consistent with those previously observed in individuals with *de novo* CHD3 variants with Snijders Blok-Campeau syndrome (SNIBCPS) (Table 1)[1; 2]. However, in the present cohort of individuals with inherited *CHD3* variants we observed fewer probands with intellectual disability (98% in the *de novo* cohorts versus 79% in the inherited cohort). Based on the available data intellectual disability seemed more severe in probands with *de novo CHD3* variants. However, this comparison is complicated by lack of data on intelligence level, for example due to too young age to perform formal testing. The cohort with individuals with inherited *CHD3* variants also contained less individuals with vision problems (72% versus 25%). Notably, 53% of the heterozygote parents (8/15) had vision problems.

Whereas the large majority (91%) of previously published *de novo* pathogenic *CHD3* variants were missense variants or in frame deletions[1; 2], we identified seven families with inherited single nucleotide protein truncating variants (PTVs) and one family with an intragenic deletion of exon 14-18 (family 8), adding up to a total of eight families with a predicted loss of function effect (38%). One of the probands with a single nucleotide PTV (p.W1427*, proband 4) was recently published alongside a cohort of individuals with *de novo CHD3* variants (patient 18 in 1). Inherited missense variants were found in thirteen families. The recently published cohort contained one additional individual with an inherited *CHD3* variant, patient 23[1]; this patient was excluded in the present cohort since the patient had a multigenic duplication of 17p13.1 (7,339,633-7,902,885; hg18), including *CHD3* and 6 other genes.

In 15/21 families (71%, $p = 0.0392$) the *CHD3* variant was maternally inherited. One single nucleotide PTV (family 5), the intragenic deletion (family 8) and four of the missense variants (family 9,10,12,16) were paternally inherited (see Figure 1 of the main text)

Families

Family 1 (p.W1158*)

Proband 1 was born after an IVF pregnancy. She first presented at the clinical genetics department at age 16 months with global developmental delay. Routine metabolic screening, SNP array and exome sequencing (trio) for intellectual disability genes and open exome analysis were normal. Methylation studies for Angelman syndrome showed no abnormalities. Re-analysis of exome sequencing data at age 4 years and 5 months revealed a maternally inherited single nucleotide PTV in *CHD3*: p.W1158*. No other variants that could explain the phenotype were identified. The phenotype at this age included global developmental delay, intellectual disability, speech delay, autism spectrum disorder, hypotonia, severe feeding problems for which the patient required a PEG catheter, hyperlaxity, and facial dysmorphism with a broad forehead, full cheeks, thin upper lip and squared face. She had a disharmonic intelligence profile. The *CHD3* variant was considered explanatory for the phenotype by the involved clinician. The mother of proband 1, carrying the same variant suffered from Graves disease and rheumatoid arthritis but had no history of developmental or speech delay. She did have a prominent forehead. Segregation analysis showed that the mother inherited the variant from her mother, who also had no history of developmental problems.

Family 2 (p.R1304*)

Proband 2 presented with profound macrocephaly (+3,2SD) global developmental delay and hardly any speech at age 3 years. She did learn sign language. At the age of 4 years, speech had improved and she made 5-6 word sentences. She had extreme stimulus-seeking behaviour and a short attention span.

The heterozygote mother followed average level education. However, she had problems doing internships and fitting herself to a working environment. She was diagnosed with autism spectrum disorder at age 20 years. She followed speech therapy during her entire life. Her head circumference was +2 SD. Maternal grandparents have not been tested for the variant.

Family 3 (p.S1384*)

Proband 3 was born after 39 weeks of pregnancy with a large head circumference (P97,7) and birth weight (>+2SD). He presented with global developmental delay at age 3 years. Speech delay was more pronounced than motor delay. At the last follow-up at 5,5 years of age, he had normal head circumference (+1,6 SD), height, and weight, and showed improved speech. Total IQ measured at age 3 years was 76. Total IQ at age 5 years was 92, but with a disharmonic profile that involved learning problems and required special education. He has regulatory difficulties, concentration problems, hyperactive and impulsive behaviour, and hypersensitivity.

The phenotype of the heterozygote mother included problems with fine motor skills, macrocephaly (+2,6 SD), vision problems and facial dysmorphism (broad forehead, squared face, coarse facies, thin upper lip, prominent chin). Intelligence was average. The CHD3 variant occurred *de novo* in the mother.

Next to the maternally inherited single nucleotide PTV, proband 3 had a paternally inherited missense variant in *CHD3* (p.Q1486E), which was considered a variant of unknown significance. The father had average intelligence, a history of stuttering, and features of ADHD and autism spectrum disorder, but with no formal diagnosis. He had a normal head circumference (slightly below +1 SD). The father inherited the missense variant from his mother.

Family 4 (p.W1427*)

Proband 4 was diagnosed at age 6. At this time she spoke very little and was very difficult to understand. The heterozygote mother of proband 4 had presumed intellectual disability and started having seizures at age 32 years. No data about facial dysmorphism were available. The uncle, maternal grandmother and maternal cousin were also suspected to be heterozygotes, but none of these individuals were available for genetic testing. Note: this individual was published as individual 18 in Ref. 1.

Family 5 (p.Q1438*)

Proband 5 first visited the genetics department at age 2 years and 8 months because of developmental delay. She additionally had speech delay, special needs in kindergarten and behavioural problems, including tantrums and aggressive behaviour. Her father had mild intellectual disability (estimated by clinician, not formally assessed), a history of developmental delay, sometimes problems to control anger, lower coordination skills, facial dysmorphism and joint hyperlaxity. A *CHD3* single nucleotide PTV was identified in both the proband and the father.

Family 6 (p.R1697*)

Proband 6 had developmental and speech delay. The protein-truncating variant was found to be mosaic in the healthy mother.

Family 7 (p.E1821*)

Proband 7 concerns a case with multiple pathogenic variants, including a maternally inherited single nucleotide PTV in *CHD3*, a *de novo* pathogenic *KCNA2* (MIM # 616366) variant and a likely pathogenic paternally inherited *FOXP2* variant (MIM # 602081). The proband had global developmental delay, so far absent speech, early infantile epileptic encephalopathy, muscular hypotonia, macrocephaly and facial and digital abnormalities. The phenotype is thought to be explained by the combination of the different pathogenic variants. Of these especially the macrocephaly and facial features fit the SNIBCPS spectrum.

Family 8 (c.2245_2978+60del)

The variant in proband 8 had a presumed loss of function (LoF) effect due to a frameshift resulting from an intragenic deletion in *CHD3* (c.2245_2978+60del, deletion exon 14-18). He had global developmental delay, and features reminiscent of Weaver syndrome. The intelligence level was unknown at age 3 years.

The heterozygote father of this individual had no features of SNIBCPS. No images were available for evaluation or facial analysis.

Family 9 (p.S477F)

Proband 9 presented with developmental coordination disorder, and both motor and speech/language delay. His intelligence was tested to be normal. He had macrocephaly, his head circumference measured 57 cm at the age of 7 years (+2,8 SD). A paternally inherited missense variant in *CHD3* was identified. The variant was also present in the oldest sister of proband 9. She had a history of substantial developmental delay, especially motor problems (hypotonia, hyperlaxity, and coordination problems), but also speech/language issues. She had difficulties writing at age 13 years. She was often ill and had recurrent upper respiratory tract/ear infections. All three individuals in family 9 that carry the *CHD3* variant showed facial features (including deep-set eyes). The father had no history of developmental delay, but did have childhood epilepsy.

He currently had no seizures without medication. The father had recurrent ear infections. He developed colon carcinoma at age 45 years.

Family 10 (p.W630R)

Proband 10 was a 9 year old boy born to non-consanguineous parents. He underwent trio clinical exome sequencing because of macrocephaly and neurodevelopmental difficulties, revealing a paternally inherited missense variant in *CHD3* (c.1888T>C), which was classified as a variant of unknown significance (class 3b). He additionally had early developmental delay, developmental coordination disorder, speech delay, a low normal IQ (~78), dyspraxia, attentional problems, posterior deformational plagiocephaly, brachycephaly, distinctive facial features (box shaped forehead, triangular face, low set ears), bilateral cryptorchidism, right sided inguinal hernia and hypotonia. He has no problems in social interaction. The heterozygote father had apparently mild literacy issues (poor spelling, reading) and maybe some mild autistic features. He had macrocephaly. He had no history of developmental problems. He did have struggles with changes in routine. Over the past 10 years the father had significant headaches. No neuroimaging was performed. The facial features of the proband were not considered the most typical, yet based on the total clinical picture (including developmental delay, speech delay, hypotonia and macrocephaly) and the mild features of the father (including macrocephaly) it was considered likely that the identified *CHD3* variant contributed to the phenotype.

Family 11 (p.F833L)

Proband 11 was a girl that was first evaluated at the genetics department at the age of 10 months. She presented with developmental delay, axial hypotonia, spastic extremities, nystagmus, myoclonic seizures, and was without teeth eruption. She had a normal body weight (8600g, p50) and small head circumference (42,5 cm, p5). Prenatal ultrasound had shown moderate ventriculomegaly. Postnatal transfontanellar ultrasound showed large ventricles. Proband 11 also had a atrial septal defect. Exome sequencing revealed a heterozygous, maternally inherited variant in *CHD3*: c.2497T>C, classified as a variant of unknown significance. There were no additional findings of genetic testing. The mother carrying the *CHD3* variant had normal intelligence and no history of developmental delay. She did have macrocephaly.

At age 13 months proband 11 was hypotonic, could not roll over, could not sit unsupported, could raise her hand, but did not grasp toys. At this time she did not say any words, but she smiled and interacted with her family. The proband was too young to evaluate cognitive impairment.

Based on the variant and the overlap of several features of the proband with those known for SNIBCPS a contribution of the *CHD3* variant to the phenotype was considered likely.

Family 12 (p.G873S)

Proband 12 was a girl with decreased foetal movements and polyhydramnios. She was born at 37 weeks and 4 days of gestation and had APGAR scores of 5, 5 and 6. She had marked neonatal hypotonia. She also experienced one episode of neonatal hypoglycemia, neonatal respiratory distress requiring intubation and surfactant, and neonatal intraventricular haemorrhages. She had macrocephaly, developmental delay, especially expressive speech delay and mild dysmorphisms (hypertelorism, downslanting palpebral fissures, broad nasal bridge, persistent glabellar nevus flammeus simplex). She also had skeletal issues: large anterior fontanel, flat feet with ankle valgus, clinodactyly of the 5th fingers bilaterally, metaphyseal striations and a mild lordosis. The *CHD3* missense variant was inherited from a father who was not known to have intellectual disability, but who had a large head circumference (60 cm, +3.4 S.D.), prominent supraorbital ridges, prominent nasal ridge, deep-set eyes, and 5th finger hypermobility. Several relatives on the paternal side had learning issues but were not assessed or tested for the *CHD3* variant as they lived abroad.

Family 13 (p.I983V)
Proband 13 was born small for gestational age (birth weight Z-1,62) after 38+4/7 weeks of pregnancy. She had global developmental delay and clear speech delay, speaking only four words at age 2 years and 5 months. Intelligence was not yet formally tested. The girl had behavioural abnormalities including tantrums and emotional dysregulation. The mother of proband 13 had developmental delay, borderline/mild intellectual disability, speech delay (dyspraxia) and facial dysmorphism. Both mother and daughter had visual abnormalities and asthma.

Family 14 (p.P1046L)
Proband 14 was a 9 year old boy with macrocephaly, developmental concerns including speech delay and autistic features, fine motor difficulties, hypermobility of joints and tall stature. He also had an immune/inflammatory disorder including periodic neutropenia and frequent daily joint and muscle pains, without a specific diagnosis forthcoming from rheumatology or immunology investigations. Intellect was estimated in the normal range but he had significant functional impairments. Proband 14 had daily headaches. Cerebral MRI did not show a Chiari Malformation.

The *CHD3* variant was maternally inherited. The mother had normal intelligence but a Chiari I Malformation with normal head circumference (57 cm, Z score +2.4). Additional features include fine motor co-ordination problems, hypermobility, frequent migraine, systemic lupus erythematosus and anti-phospholipid syndrome. Childhood features include stutter and obsessive compulsive features.

Family 15 (p.P1125A)
Proband 15 was a 5 year old boy. He had developmental delay with delayed motor milestones and speech delay. Intelligence has not yet been formally assessed. His social behaviour was normal. Body measurements were 125,5cm (P98, +2,07 SDS), 42,2kg (>P99, +3,83 SDS), head circumference 55cm (P99, +2,40 SDS). A variant of unknown significance in CHD3 was identified with diagnostic exome sequencing. This concerned a maternally inherited missense variant in the ATPase-helicase domain, close to reported *de novo* missense variants. The mother was also overweight but had no developmental delay or other clinical features. She had a normal occupation. The facial phenotype of proband 15 very well matched the classic SNIBCPS dysmorphisms, which convinced us that the *CHD3* variant contributed to his phenotype.

Family 16 (p.R1342Q)
Proband 16 had global developmental delay. At age 3 years and 3 months she spoke three words. She was estimated to have moderate to severe intellectual disability (no formal testing). Additionally, she had erethic-hyperkinetic disorder and suspected ataxia. She did not chew food properly. Little information was available about the heterozygote father. He showed frequent jerks/twitching. The father was investigated for mosaicism, which was not found. The clinician had the impression that the father was simple-minded.

Family 17 (p.R1706Q)
Proband 17 was evaluated at 3.5 and 6 years of age. She had autism spectrum disorder, global developmental delay, large stature, joint hypermobility and poor sleep. At age 3.5 years, height and weight were >97th percentile and head circumference was >75th percentile. At age 6 years, height and weight were >90th percentile, and head circumference was at the 98th percentile. The variant was inherited from a mother who had difficulty at school, mild learning delay, obsessive compulsive disorder, rheumatoid arthritis, fibromyalgia, and premature menopause. Length and head circumference of the heterozygote mother were normal.

Family 18 (p.R1759Q)
Proband 18 had mild intellectual disability and autism spectrum disorder. He had especially motor problems, but also difficulty speaking and writing. The heterozygote mother had university level education, but did have features of autism and ADHD, as well as motor problems. Several maternal cousins had features of autism spectrum disorder, behavioural problems and intellectual disability.

Family 19 (p.E1837K)
Proband 19 had autism spectrum disorder and behavioural abnormalities including running away from parents and hand flapping. He used to have self-injury and tantrums. He had speech delay and followed special education, but intelligence level was unknown. He had a stable leukodystrophy. In addition to the *CHD3* missense variant, proband 19 was found to have a 22q11.2 duplication. Since the phenotype associated with the duplication is highly variable, ranging from no developmental phenotype to severe intellectual disability[3-5], it is not possible to determine to what extend the duplication contributes to the phenotype.

Proband 19 and the heterozygote mother both had macrocephaly, but no further phenotype was documented in the mother.

Family 20 (p.R1888Q)
Proband 20 was born prematurely at 33 weeks of pregnancy. She had moderate global developmental delay. At age 3 years and 3 months she spoke 5-10 single words. She had an atrial septal defect and brain abnormalities detected by MRI (dysplastic lateral ventricles, mildly thinned splenium and blunted rostrum of corpus callosum, enlarged frontotemporal extra-axial CSF spaces, dilated posterior horns of lateral ventricles in keeping with bilateral colpocephaly). The family comprised a mother with three affected children. One of these children had a different father. A developmental phenotype was present in all, including the mother, ranging from moderate to

borderline intellectual disability. One child had severe intellectual disability, but this was presumed to be the consequence of subdural haemorrhage at age 5 weeks and subsequent acquired brain damage.

Family 21 (p.K1972T)

Proband 21 first presented at age 22 months and was genetically diagnosed at age 6 years. Head circumference at birth was normal; however, since age 1 year and 10 months he had profound macrocephaly (>+4z). He had global developmental delay, speech delay and autism spectrum disorder. His symptoms additionally included urinary incontinence. The heterozygote mother was asymptomatic. The maternal grandmother's sister had developmental delay/macrocephaly, and died at 27 years of age (no further information available). Maternal cousins were reported to have developmental delay. No further segregation analysis was performed.

References supplemental notes 1

1. Drivas, T.G., Li, D., Nair, D., Alaimo, J.T., Alders, M., Altmuller, J., Barakat, T.S., Bebin, E.M., Bertsch, N.L., Blackburn, P.R., et al. (2020). A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome. Eur J Hum Genet 28, 1422-1431.
2. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nat Commun 9, 4619.
3. Ou, Z., Berg, J.S., Yonath, H., Enciso, V.B., Miller, D.T., Picker, J., Lenzi, T., Keegan, C.E., Sutton, V.R., Belmont, J., et al. (2008). Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. Genet Med 10, 267-277.
4. Wentzel, C., Fernström, M., Ohrner, Y., Annerén, G., and Thuresson, A.C. (2008). Clinical variability of the 22q11.2 duplication syndrome. Eur J Med Genet 51, 501-510.
5. Yu, A., Turbiville, D., Xu, F., Ray, J.W., Britt, A.D., Lupo, P.J., Jain, S.K., Shattuck, K.E., Robinson, S.S., and Dong, J. (2019). Genotypic and phenotypic variability of 22q11.2 microduplications: An institutional experience. Am J Med Genet A 179, 2178-2189.

Supplemental notes 2

3D-protein modelling of inherited CHD3 ATPase-Helicase variants

We modeled the protein structure of the ATPase-Helicase domain of CHD3 in interaction with the DNA using the homology modelling script in the WHAT IF[1] & YASARA[2] Twinset with standard parameters. As a template, we used PDB file 6RYR which contains the human Nucleosome-CHD4 complex structure of a single copy of CHD4[3]. We performed variant analysis on previously published *de novo* CHD3 ATPase-Helicase domain missense variants[4; 5], and on four inherited CHD3 ATPase-Helicase domain missense variants (p.F833L, p.G873S, p.I983V, p.P1046L; p.P1125A; Figure 1).

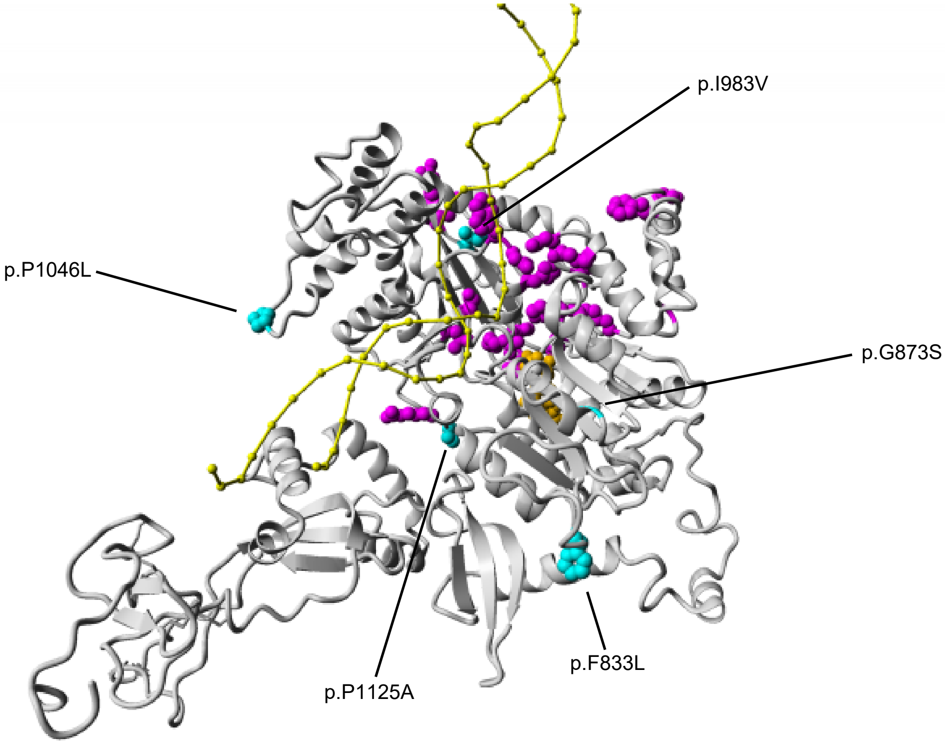


Figure 1. Overview of *de novo* and inherited CHD3 variants in the ATPase-Helicase domain. 3D model of the CHD3 ATPase-Helicase domain (homology model based on PDB: 6RYR; grey) in interaction with the DNA (yellow). ATP is depicted in orange. *De novo* CHD3 variants in the ATPase Helicase domain are shown in magenta, and cluster at DNA- and ATP-binding sites. The inherited CHD3 variants in this domain are depicted in cyan, and seem further away from DNA- and ATP-binding sites.

Variant analysis

p.F833L (Figure 2): This change concerns a semiburied residue. Leucine has a smaller sidechain and therefore some hydrophobic interactions will be lost leading to a slight destabilisation of this area. This could affect the interaction with either the DNA or another protein, but only slightly.

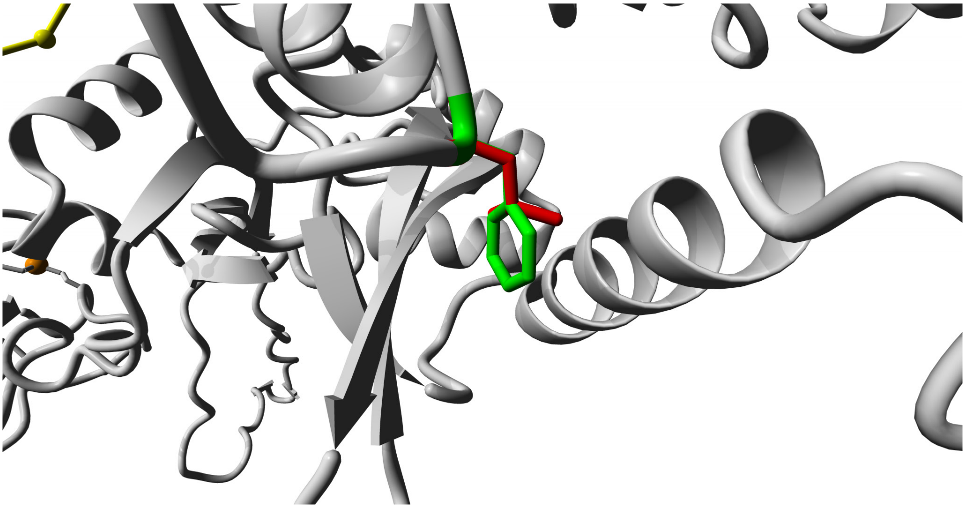


Figure 2. Close-up of the p.F833L variant. The CHD3 protein is coloured in grey and the DNA is shown in yellow. The wild-type (F) and mutant (L) sidechains are shown in green and red respectively.

p.G873S (Figure 3): Introduces a slightly bigger sidechain that is also known to interact with DNA. Since the residue is semiburied, only a small destabilisation would be expected.

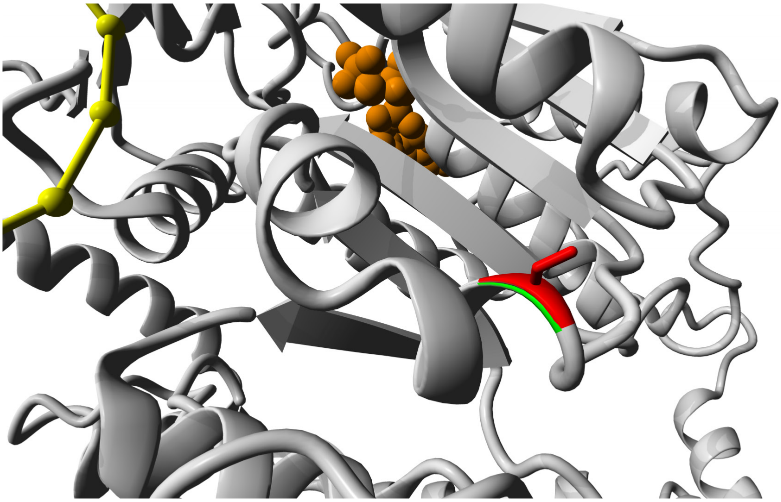


Figure 3. Close-up of the p.G873S variant. The CHD3 protein is coloured in grey, the ATP in orange and the DNA is shown in yellow. The wild-type (G) and mutant (S) sidechains are shown in green and red respectively.

p.I983V (Figure 4): Also a semiburied residue and located far away from the DNA. The loss of some hydrophobic interactions made by the sidechain does not seem to have many consequences.

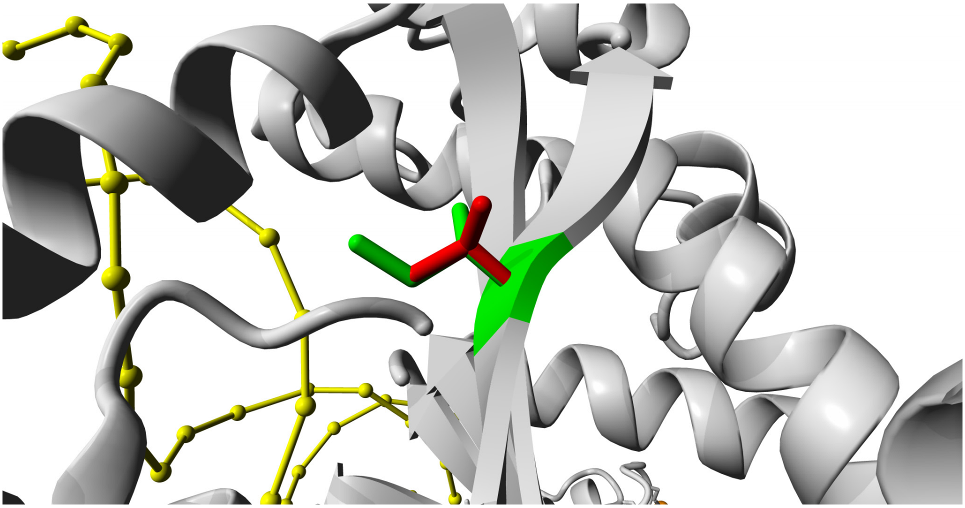


Figure 4. Close-up of the p.I983V variant. The CHD3 protein is coloured in grey and the DNA is shown in yellow. The wild-type (I) and mutant (V) sidechains are shown in green and red respectively.

p.P1046L (Figure 5): Located in a clear and possibly flexible surface loop, where the new and larger leucine sidechain easily fits. The stable structure of the loop caused by proline will be lost. The function of this loop is unclear, but might be necessary for interaction with other proteins.

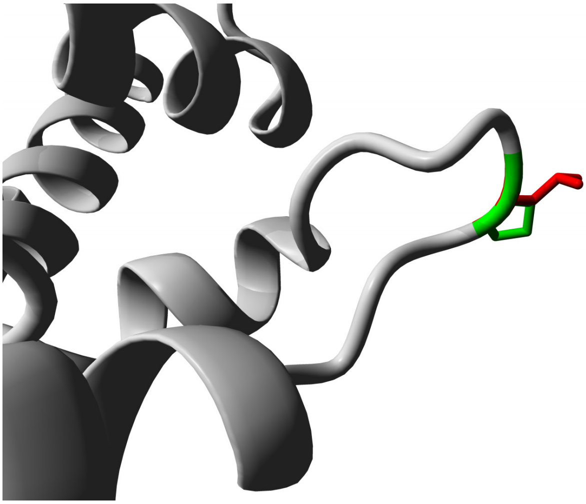


Figure 5. Close-up of the p.P1046L variant. The CHD3 protein is coloured in grey and the DNA is shown in yellow. The wild-type (P) and mutant (L) sidechains are shown in green and red respectively.

p.P1125A (Figure 6): Located at the surface of the protein, far away from the DNA and in a region that does not seem to interact with other sections of the CHD3 protein. The smaller size of the alanine residue compared to the proline residue could locally change the backbone formation of the protein, although with unclear consequences.

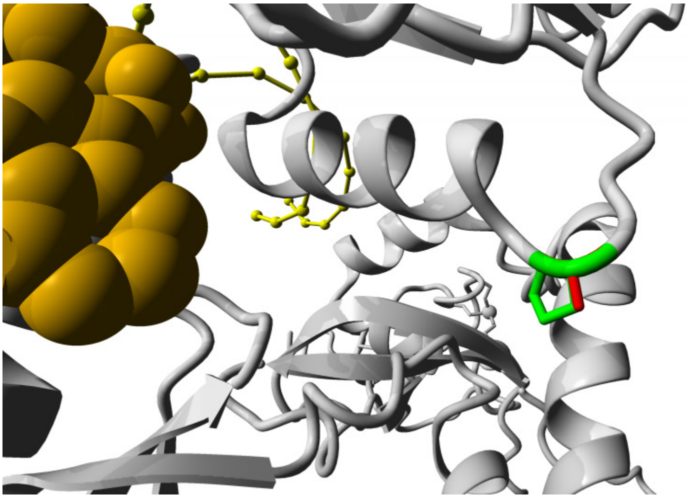


Figure 6. Close-up of the p.P1125A variant. The CHD3 protein is coloured in grey, the ATP in orange and the DNA is shown in yellow. The wild-type (P) and mutant (A) sidechains are shown in green and red respectively.

References supplemental notes 2

1. Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. J Mol Graph 8, 52-56, 29.
2. Krieger, E., Koraimann, G., and Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins 47, 393-402.
3. Farnung, L., Ochmann, M., and Cramer, P. (2020). Nucleosome-CHD4 chromatin remodeler structure maps human disease mutations. Elife 9.
4. Drivas, T.G., Li, D., Nair, D., Alaimo, J.T., Alders, M., Altmuller, J., Barakat, T.S., Bebin, E.M., Bertsch, N.L., Blackburn, P.R., et al. (2020). A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome. Eur J Hum Genet 28, 1422-1431.
5. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nat Commun 9, 4619.

Supplemental notes 3

Population-level genetic analyses in UK Biobank

1. Materials and methods
1.1 Data
1.1.1 UK Biobank

The research described in this section has been conducted using the UK Biobank (UKB) Resource under application 16066, with Clyde Francks as the principal applicant. This is a general adult population cohort[1]. The data collection in the UKB, including the consent procedure, has been described elsewhere 2. Informed consent was obtained by the UKB for all participants. We made use of imaging-derived phenotype data generated by an image-processing pipeline developed and run on behalf of the UKB[3; 4].

1.1.2 Exome sequencing data

Exome sequencing of approximately 200,000 individuals was performed through the UKB Exome Sequencing Consortium according to protocols described elsewhere[5; 6] (https://www.ukbiobank.ac.uk/media/cfulxh52/uk-biobank-exome-release-faq_v9-december-2020.pdf). Briefly, exomes were captured with the IDT xGen Exome Research Panel v1.0 including supplementary probes, targeting 39 Mbp of the human genome. Sequencing was performed on the Illumina NovaSeq 6000 platform and data was processed using the OQFE protocol[5]: raw reads were mapped to a GRCh38 reference, retaining all supplementary alignments, and duplicate reads were marked. Variants were called, restricted to exome capture regions plus the 100 basepairs flanking each capture target, resulting in a genomic variant-call file (gVCF) per sample. gVCFs were then merged into multi-sample project-level VCF (pVCF) files. We downloaded the multi-sample pVCF files and removed variants outside sequencing target regions defined by the UKB, as sequencing quality standards for variants outside these regions were not assessed. We then extracted variants in the CHD3 gene region, which we defined as chr17:7884796-7912760 (genome build GRCh38/hg38) or chr17:7788124-7816078 (GRCh37/hg19), yielding 1,056 variant sites. Variants in the CHD3 region were annotated using snpEff v5.0 (build 2020-10-04)[7]. In addition to default snpEff annotations, we annotated variants with SIFT4G[8], PolyPhen2-HDIV[9] and CADD[10] scores derived from dbNSFP (version 4.1a)[11]. SIFT4G scores ranged from 0-1, with smaller scores indicating a more likely damaging effect. Within dbNSFP, variants with a SIFT4G score < 0.05 were labelled as 'damaging'. PolyPhen2-HDIV scores ranged from 0 to 1, with higher scores indicating a more likely damaging effect. Within dbNSFP, variants with a PolyPhen-2 score [0.957-1] were labelled as 'probably damaging'. CADD scores were presented as Phred-like (log10-derived) rank scores based on the distribution of all CADD scores across the genome, with higher Phred-scores indicating higher predicted deleteriousness.

1.2 Association of rare variants with educational qualifications, fluid intelligence score and intracranial volume
1.2.1 Phenotype data

We selected three main phenotypes for association testing with rare missense variants in *CHD3*: 'Fluid intelligence score' (data field ID 20016), 'Qualifications' (data field ID 6138) and 'Volume of EstimatedTotalIntraCranial (whole brain)' (data field ID 26521). For each main phenotype, additional phenotypes were selected to use as covariates in the association tests (Table 1). Fluid intelligence was recorded at multiple instances, and for each individual the data from the first available instance was used. Qualifications was also reported at multiple instances and had six possible answering options (1: College or University degree, 2: A levels/AS levels or equivalent, 3: O levels/GCSEs or equivalent, 4: CSEs or equivalent, 5: NVQ or HND or HNC or equivalent, and 6: Other professional qualifications (e.g. nursing, teaching, etc.)), with multiple possible answers per instance. To make this phenotype suited for association testing, categories were merged to create a binary phenotype: one group comprised individuals for which the highest reported qualification was option 1 or 2 ('high' educational qualification), and the other group comprised individuals for which the highest reported qualification was option 3, 4 or 5 ('low' educational qualification). Because option 6 (other professional qualifications) can refer to diverse educational levels, we excluded individuals who only reported this option as their highest qualification. Furthermore, we excluded individuals if they were inconsistently assigned to both 'high' and 'low' categories across instances. In all datasets, we included the first 10 principal components (PCs) based on common variant data that reflect genetic ancestry[1], and the exome sequencing batch (a binary variable indicating whether an individual was sequenced in the first batch of ~50,000 individuals or in the second batch of ~150,000 individuals) as additional covariates. We made a separate dataset for each of the three main phenotypes, including only individuals with available exome sequencing data.

1.2.2 Sample-level filtering

In each of the three datasets, individuals with missing data for one or more covariates were excluded. We additionally excluded individuals with discordant phenotypic and genetically inferred sex (as reported in the exome sequencing data plink .fam file), as well as those not in the white-British ancestry subset as defined by the UKB based on self-reported ancestry and clustering in the first six genetic PCs[1]. In pairs of individuals defined as third-degree relatives or closer (kinship coefficient > 0.0442) by the UKB[1], we randomly removed one individual of a pair, prioritising removal of individuals related to multiple others.

1.2.3 Exome sequencing data filtering

Following sample-level filtering, we applied genotype- and variant site-level hard-filtering in each phenotype dataset using vcftools 0.1.17[12] and previously published thresholds[6; 13]. Genotype-level filtering changed genotypes with a low approximate read depth (DP < 7 for single-nucleotide variant [SNV] sites and DP < 10 for insertion/deletion [INDEL] sites) and/or low genotype quality (GQ < 20) to no-call. Variant-level filtering removed sites labelled as 'monoallelic', sites with low average genotype quality across individuals (average GQ < 35), a high missingness rate (> 0.12), a minor allele count (MAC) of zero, and/or a low allele balance (AB < 0.15 for SNV sites and < 0.20 for INDEL sites, calculated using GATK v4.1.9.0[14]).

1.2.4 Region-based association test of missense variants

For region-based association testing, we removed multi-allelic variants and converted data to plink binary format using plink v1.90b6.9 15. We then used MetaDome to

Table 1. UK Biobank phenotypes selected for association testing

Fluid intelligence score			
Field ID	**Phenotype**	**Instance**	**Note**
31	Sex	0	
20016	*Fluid intelligence score*	0, 1, 2 or 3	For each individual, data from the first available instance was selected.
21003	Age when attended assessment centre	0, 1, 2 or 3	For each individual, age was selected from the same instance as the one from which fluid intelligence score was obtained.
22009	Genetic principal components 1-10	0	
-	Exome batch	-	Binary variable indicating whether individuals were sequenced in the first (50k) or second (150k) batch.

Educational qualifications			
Field ID	**Phenotype**	**Instance**	**Note**
31	Sex		
34	Year of birth		
6138	*Qualifications*	0, 1, 2 or 3	
22009	Genetic principal components 1-10	0	
-	Exome batch	-	Binary variable indicating whether individuals were sequenced in the first (50k) or second (150k) batch.

Intracranial volume			
Field ID	**Phenotype**	**Instance**	**Note**
31	Sex	0	
54	UK Biobank assessment centre	2	
21003	Age when attended assessment centre	2	
22009	Genetic principal components 1-10	0	
25734	Inverted signal-to-noise ratio in T1	2	
25735	Inverted contrast-to-noise ratio in T1	2	
25756	Scanner lateral (X) brain position	2	
25757	Scanner transverse (Y) brain position	2	
25758	Scanner longitudinal (Z) brain position	2	
26521	*Volume of EstimatedTotalIntraCranial (whole brain)*	2	
-	Exome batch	-	Binary variable indicating whether individuals were sequenced in the first (50k) or second (150k) batch.
-	Nonlinear age	-	Calculated from 'age when attended assessment centre'.

The main phenotype for each analysis is highlighted in italic, other phenotypes were included as covariates. The Field ID column shows the phenotype identifier in the UK Biobank. Instance reflects the assessment centre visit(s) from which the recorded phenotype was used. 0: First assessment center visit | 1: Follow-up assessment center visit | 2: First imaging visit | 3: Second imaging visit. For some phenotypes, data from multiple instances was combined (see methods section).

select missense variants at variant-intolerant locations in functional domains of *CHD3*[16]. Only variants in the three best characterised transcripts of *CHD3* (Ensembl transcript IDs: ENST00000330494, ENST00000358181, ENST00000380358), a predicted 'damaging' (SIFT4G) or 'probably damaging' (PolyPhen2-HDIV) effect and a CADD PHRED score > 25, and a minor allele frequency (MAF) ≤ 1% were included. Optimised sequence kernel association testing (SKAT-O), implemented in the R package SKAT[17], was used to capture the combined association of the selected rare missense variants in *CHD3* with the three selected phenotypes. The SKAT-O test was designed to maximise power by optimally combining a burden and sequence-kernel association test (SKAT), making it both suitable under scenarios where rare variants within *CHD3* are associated with corresponding direction of effect, as well as in the presence of null effects or opposing effect directions. Variants were weighted for MAF using the default beta(1,25) density function implemented in SKAT, and null models for each phenotype included the covariates as shown in Table 1. For the educational qualification phenotype, we used an improved ‘robust’ version of the SKAT-O method for association testing of binary phenotypes[18].

1.2.5 Phenotypes of individuals with putative loss-of-function variant

We extracted protein coding variants with a high putative impact in the three best characterised transcripts (Ensembl transcript IDs: ENST00000330494, ENST00000358181, ENST00000380358) of *CHD3* from the remaining variants after filtering. Individuals carrying one or more putative loss-of-function (LoF) variants were identified in each of the three phenotype datasets. We then applied linear regression analysis to assess differences in fluid intelligence score and intracranial volume between individuals with LoF variants and controls (individuals that did not carry a *CHD3* LoF variant), and projected phenotypes of individuals with a LoF variant on the phenotype distribution of controls. For educational qualification, we used logistic regression to assess group differences. All regression analyses included the same set of covariates for each main phenotype as in the association test of missense variants (listed in Table 1).

1.3 Common variant association of CHD3 with head circumference and intracranial volume

We used association results of SNPs located inside the *CHD3* gene region, derived from previously published genome-wide association meta-analysis summary statistics of head circumference (N ≤ 18,881), and head circumference combined with intracranial volume (N ≤ 45,458), in child- and adulthood[19]. This study included both common and low-frequency genetic variants. In addition, we obtained results from a smaller genome-wide association meta-analysis on infant head circumference (N ≤ 10,768)[20]. For each study, we calculated gene p-values reflecting the common variant association of CHD3 with HC and ICV using the default ‘snp-wise=mean’ analysis model in MAGMA (v1.09a)[21]. We applied Bonferroni correction for three included genome-wide studies when assessing significance.

2. Results

2.1 Region-based association testing of missense variants

For the three main phenotypes, we selected individuals with available exome sequencing data and a valid reported phenotype. We then performed sample-level quality control, leaving 120,596 individuals for association analysis of educational qualification, 77,998 for fluid intelligence and 18,254 for intracranial volume (Table 2). We extracted genotype data for the individuals in each dataset and performed genotype- and variant-level filtering on 1,056 variant sites present in the defined *CHD3* region (Table 3). From variants remaining after filtering, we extracted missense variants on variant-intolerant locations in functional domains of *CHD3*. 47 selected variants overlapped with the post-QC variants in the three datasets, with 43 variants present in the educational qualification dataset, 37 variants in the fluid intelligence

Table 2. Overview sample-level filtering

Filtering step/Dataset	Qualifications	Fluid intelligence score	Intracranial volume
Available main phenotype before filtering	154,065	100,300	21,727
Missing data in one or more covariates	133	95	383
Discordant reported and genetic sex	35	36	9
Not in white British ancestry subset	26,161	17,736	2,853
Third-degree or higher relatedness to other individual(s) in dataset	7,277	4,535	291
Unique individuals flagged for removal	33,469	22,302	3,473
Individuals left after filtering	120,596	77,998	18,254
Female	65,458	41,852	9,669
Male	55,138	36,146	8,585
Age (mean) [range]	-	58.25 [40-81]	63.24 [45-81]
Year of birth (median) [range]	1952 [1936-1970]	-	-

dataset, and 13 variants in the intracranial volume dataset. SKAT-O analysis revealed no significant association of missense variants in CHD3 with educational qualification (*p*=0.35), fluid intelligence score (*p*=0.36) and intracranial volume (*p*=0.57) (Table 4).

2.2 Phenotypes of individuals with a putative LoF variant

We only identified individuals with heterozygous putative LoF variants. There was no significant effect of the group with LoF variants on educational qualification (z = 0.427, p = 0.67) and fluid intelligence score (t = 0.189, *p* = 0.85). Intracranial volumes of four individuals with a putative LoF variant were higher than the average in individuals that did not carry a *CHD3* LoF variant, with a nominally significant group effect (t = 2.37, *p* = 0.018; Table 5, Figure 1) that was not significant after a conservative Bonferroni correction for the three phenotypes tested in this analysis (adjusted *p* = 0.054).

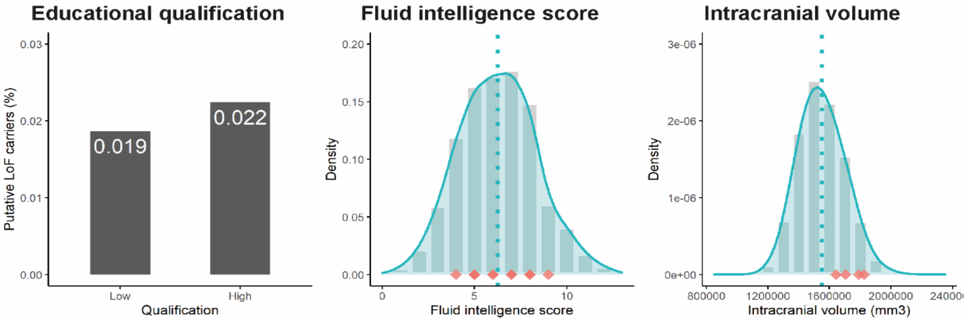


Figure 1. Phenotype frequencies and distributions for individuals with a putative loss-of-function variant and controls. The bar plot for educational qualification shows the percentage of individuals with a LoF variant among all individuals in 'low' and 'high' educational qualifications groups. For fluid intelligence score and intracranial volume, phenotype distributions are shown for controls, and phenotype values for the individuals with a LoF variant are projected on the distributions (red diamonds).

2.3 Common variant association of CHD3 with head circumference and intracranial volume

Gene-based analysis in MAGMA revealed no significant associations of common variants in *CHD3* with head circumference after Bonferroni correction for the multiple studies included in this analysis, although the combined analysis of head circumference with intracranial volume suggested a subtle association (nominal *p*=0.047; Table 6).

Table 3. Overview of exome sequencing data filtering

Filtering step/Dataset	Qualifications	Fluid intelligence score	Intracranial volume
Variant sites in target regions pre-filtering	1,056	1,056	1,056
Monoallelic sites	15	15	15
Low average genotype quality (average GQ < 35)	2	2	2
High missingness rate (> 0.12)	6	6	7
Minor allele count of zero (MAC = 0)	324	457	785
Low allele balance (SNV sites < 0.15, INDEL sites < 0.20)	0	0	0
Unique variant sites removed	345	476	802
Variant sites post-filtering	711	580	254
Ts-Tv ratio pre-filtering	4.08	4.08	4.08
Ts-Tv ratio post-filtering	4.96	4.44	5.68
Multiallelic sites removed prior to region-based testing	88	77	43
Missense variants at intolerant sites in functional regions of *CHD3*	43	37	13

Table 4. Region-based association test results

Phenotype	P-value	Q	Variants	MAC
Qualifications	0.35	9977.66	43	197
Fluid intelligence score	0.36	53090.60	37	135
Intracranial volume	0.57	9789.92	13	33

P-value: SKAT-O association test p-value | Q: SKAT-O test Q value (test statistic) | Variants: number of variants tested within CHD3 in association analysis | MAC: Total/Gene minor allele count across all tested variants.

Table 5. Main phenotypes for individuals with LoF variants versus controls

Phenotype	N total	N individuals with a LoF variant	Frequency controls	Frequency individuals with a LoF variant
Qualifications	120,596	25	N_{High} = 66,941 N_{Low} = 53,630	N_{High} = 15 N_{Low} = 10
			Mean (SD) controls	**Mean (SD) individuals with a LoF variant**
Fluid intelligence score	77,998	14	6.26 (2.09)	6.42 (1.50)
Intracranial volume	18,254	4	1549140 (150494)	1740448 (82451)

N total: total sample size for phenotype | N individuals with a LoF variant: The number of individuals with a LoF variant in the phenotype dataset | For qualifications, the frequency of controls and individuals with a LoF variant in the 'high' and 'low' groups is shown. For fluid intelligence and intracranial volume, the phenotype means (SD) are shown for controls and individuals with a LoF variant.

Table 5. Gene-level common variant associations of *CHD3* with head circumference and intracranial volume

GWAS dataset	N SNPs	N	ZSTAT	P
Haworth et al. – Head circumference	64	15564	0.826	0.20
Haworth et al. – Head circumference + intracranial volume	65	31166	1.68	0.047
Taal et al. – Infant head circumference	20	10727	0.693	0.24

N SNPs: Number of SNPs included in analysis | N: (average) sample size | ZSTAT: Z-statistic of the gene in gene-level analysis | P: Gene p-value

References supplemental notes 3

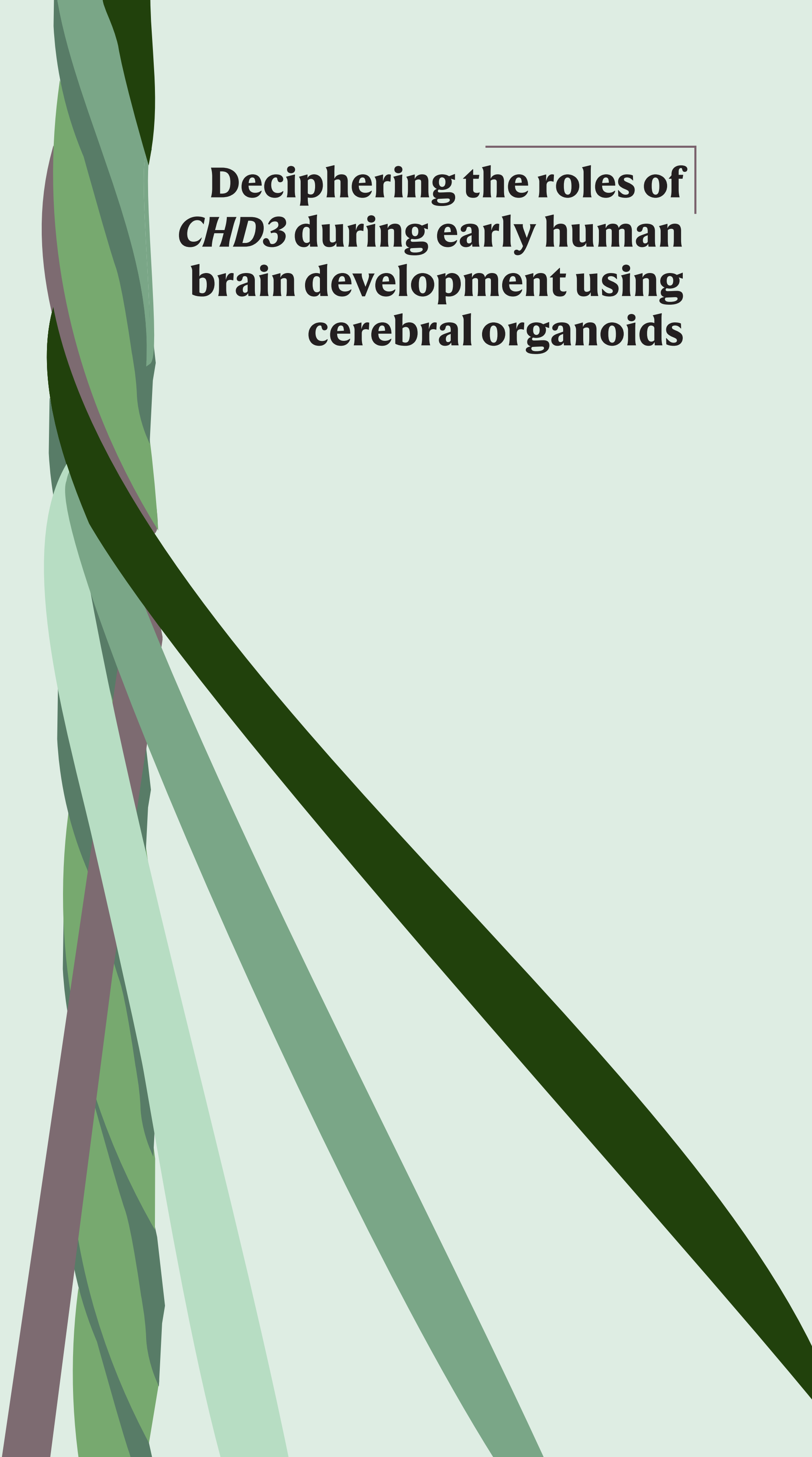
1. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203-209.
2. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 12, e1001779.
3. Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage 166, 400-424.
4. Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nature neuroscience 19, 1523-1536.
5. Szustakowski, J.D., Balasubramanian, S., Sasson, A., Khalid, S., Bronson, P.G., Kvikstad, E., Wong, E., Liu, D., Davis, J.W., Haefliger, C., et al. (2020). Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank. medRxiv, 2020.2011.2002.20222232.
6. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature 586, 749–756.
7. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6, 80-92.
8. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nature protocols 11, 1-9.
9. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics Chapter 7, Unit7.20.
10. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research 47, D886-d894.
11. Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome medicine 12, 103.
12. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics (Oxford, England) 27, 2156-2158.
13. Carson, A.R., Smith, E.N., Matsui, H., Braekkan, S.K., Jepsen, K., Hansen, J.B., and Frazer, K.A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC bioinformatics 15, 125.
14. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303.
15. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 7.
16. Wiel, L., Baakman, C., Gilissen, D., Veltman, J.A., Vriend, G., and Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. Human mutation 40, 1030-1038.
17. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American journal of human genetics 91, 224-237.
18. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L.G., and Lee, S. (2020). UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant

Test. American journal of human genetics 106, 3-12.
19. Haworth, S., Shapland, C.Y., Hayward, C., Prins, B.P., Felix, J.F., Medina-Gomez, C., Rivadeneira, F., Wang, C., Ahluwalia, T.S., Vrijheid, M., et al. (2019). Low-frequency variation in TP53 has large effects on head circumference and intracranial volume. Nature communications 10, 357.
20. Taal, H.R., Pourcain, B.S., Thiering, E., Das, S., Mook-Kanamori, D.O., Warrington, N.M., Kaakinen, M., Kreiner-Møller, E., Bradfield, J.P., Freathy, R.M., et al. (2012). Common variants at 12q15 and 12q24 are associated with infant head circumference. Nat Genet 44, 532-538.
21. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS computational biology 11, e1004219.

5

Authors:
Joery den Hoed, Maggie M. K. Wong, Willemijn J. Claassen, Lukas Lütje, Michael Heide, Wieland B. Huttner and Simon E. Fisher

Deciphering the roles of *CHD3* during early human brain development using cerebral organoids



Abstract

Changes in the dynamics of chromatin state that control spatiotemporal gene expression patterns are crucial during human brain development. CHD3 is a chromatin remodeler that is highly expressed during neurogenesis and that functions as a core member of the NuRD complex, a large multiprotein complex mediating chromatin state. *De novo* and rare inherited genetic disruptions in *CHD3* have been implicated in a neurodevelopmental disorder characterised by intellectual disability, macrocephaly and severe speech/language deficits. To study the roles of *CHD3* during early human brain development, we generated induced pluripotent stem cells with heterozygous and homozygous loss-of-function mutations, differentiated them into cerebral organoids, and analysed the organoids with immunohistochemistry, RNA-sequencing and single-cell transcriptomics. Loss of *CHD3* expression had no detectable effects on early neuroepithelial formation and organoid growth. However, in two-month-old organoids we observed a shift in cell composition, with decreased *CHD3* expression resulting in a larger neural progenitor pool, and fewer early-born neurons. In these cell types, loss of *CHD3* led to increased expression of genes involved in progenitor maintenance and decreased expression of pro-neural genes. Taken together, we identified CHD3 as an upstream regulator that facilitates neurogenesis, and controls the balance between progenitors and neurons. Our results based on genetically engineered knockout organoids pave the way for future studies modelling the neurobiological pathways affected in *CHD3*-related disorder.

Introduction

The development of the human brain is a complex process, starting from the proliferating neuroepithelium in the neural tube. The neuroepithelial cells differentiate into various types of progenitors that give rise to a large number of cell types making up functionally distinct brain regions. Tight regulation of gene expression steers temporally and spatially restricted trajectories of neuronal differentiation to allow for the emergence of the observed cellular diversity in the brain[1-3].

One of the factors underlying this spatiotemporal gene expression during development is chromatin state[4; 5], controlling both the dynamic and stable accessibility of enhancer and promoter areas in the DNA for the regulatory machinery[5; 6]. Indeed, next-generation sequencing studies on human neurodevelopmental conditions, as well as functional studies in animal and human cellular models, have identified chromatin regulatory proteins and transcription factors as key regulators of neuronal differentiation and function[5-7].

The Chromodomain helicase DNA-binding protein (CHD) family is a group of chromatin remodellers that have emerged as important regulators of brain development. From its nine members, divided in three subfamilies8, seven have been implicated in neurodevelopmental disorders[9-15]. All CHD proteins share two chromodomains used for chromatin binding, and an ATPase-Helicase domain that hydrolyses ATP to alter the physical state of chromatin by shuffling the position of the nucleosomes relative to the DNA[16]. By opening and closing the chromatin, CHD proteins can both activate and repress the expression of their targets[17].

CHD3, CHD4 and CHD5 belong to one subclass of the CHD protein family (class-II)[8] and have been found to function as core subunits of a large protein complex with both histone deacetylase and remodelling activities, called the NuRD (**Nu**cleosome **R**emodelling and **D**eacetylase) complex (Figure 1A)[18; 19]. In mouse brain development, Chd3, Chd4 and Chd5 are mutually exclusive in this chromatin remodelling complex, and were shown to have non-redundant roles during mouse cortical development. While Chd4 promotes proliferation of neural progenitors (NPCs) by activating the expression of NPC marker genes, Chd3 and Chd5 facilitate neuronal migration and cortical layer specification by repressing NPC-related markers and activating regulators of neuronal differentiation[19].

Based on bulk RNA-seq data[2], *CHD3*, *CHD4* and *CHD5* have distinct temporal expression patterns in the developing human brain, with *CHD4* having peak expression at the most early stages (8-10 post conception week), while *CHD3* is most highly expressed at mid-gestation (13-19 pcw), and *CHD5* expression levels increase later during development (24 pcw and post-natally; Figure 1B). Single-cell transcriptomic datasets of human foetal brain tissue show that *CHD4* is most highly expressed in neural progenitors, while *CHD3* expression peaks in neuronal cell populations (Figure 1C and 1D)[20; 21].

In humans, *de novo* variants in *CHD3*, *CHD4* and *CHD5* cause different neurodevelopmental conditions of variable severity, usually involving developmental delay and intellectual disability (*CHD3*: MIM #618205, *CHD4*: #617159, *CHD5*: no OMIM entry yet)[11-13]. While about 50-80% of individuals with *CHD3* and *CHD4* variants have macrocephaly[11; 12] revealing putative contributions of those genes in early brain growth, variants in *CHD5* have been associated with epilepsy[13], suggesting that *CHD5* may be crucial for functions of mature neurons. For all three class-II CHD proteins, the large majority of *de novo* aetiological variants that have been identified are missense variants, with potential to exert a dominant-negative effect. More evidence for involvement of nonsense variants has been found in familial cases of disorder, as shown for both *CHD3* (Chapter 4 of this thesis) and *CHD5*[13], suggesting variable expressivity and/or reduced penetrance for loss-of-function variation in these genes.

Although work with mouse models has shown the importance of the class-II CHD proteins in cortical development[19], their roles in the ontogeny of the human brain remain poorly understood. In the current study, we aimed to investigate this question, using a combination of CRISPR-Cas9 gene editing to create *CHD3* knockout mutations in stem cells, and generation of cerebral organoids to model early stages of human brain development[22]. Our data suggest that a complete lack of *CHD3* expression does not fully disrupt neuronal development, as all the different cell types in the organoid model, including post-mitotic neurons, were still represented in homozygous *CHD3* knockouts. However, the knockout organoids showed an increase in the relative number of NPCs, which was consistent with an upregulation of genes maintaining the NPC-state. Levels of positive regulators of neuronal differentiation were decreased in both NPCs and early-born neurons lacking *CHD3* expression. These results identify CHD3 as an important regulator controlling the timing of neuronal differentiation in human brain development.

Results

CHD3 is expressed in human induced pluripotent stem cell-derived cerebral organoids

While prior studies identified potential roles of this gene in neuronal differentiation and layer specification via knockdown experiments in mouse embryos[19], we sought to investigate the functions of *CHD3* during human brain development by employing a stem cell-based organoid model system. In analyses of developmental transcriptomic data of subdissected cortical regions, we found *CHD3* to have peak expression levels throughout all cortical regions at 16-19 pcw (Figure 1D)[2]. Therefore, we decided to perform follow-up investigations of *CHD3* function using a widely used and well-established brain organoid model that recapitulates aspects of *in vivo* cortical development (cerebral organoids)[20; 22; 23]. Following a published protocol[20] and using a commercially available induced pluripotent stem cell line (BIONi010-A; iPSC)[24], we show that *CHD3* expression levels in cerebral organoids increase during differentiation, and reach their highest levels around day 50 to day 70 of organoid development (Figure S1).

CRISPR-induced frameshift variants in exon 3 disrupt *CHD3* expression in cerebral organoids

In order to investigate whether CHD3 is crucial for neuronal differentiation, we employed CRISPR-Cas9 gene editing to generate human iPSC lines with a heterozygous or

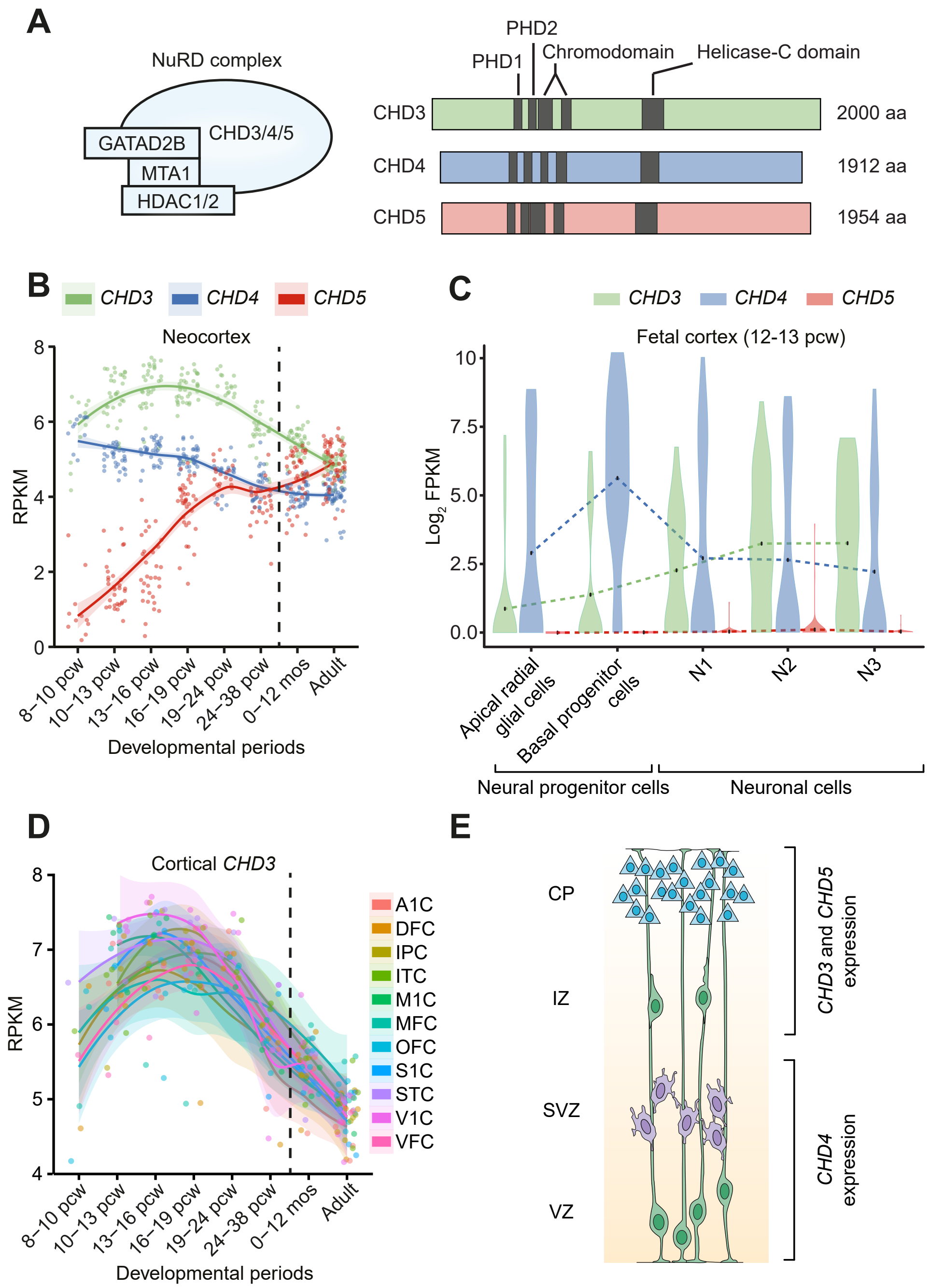


Figure 1. Expression of class II CHD proteins in the human developing brain. A) Left, a schematic of the chromatin remodelling NuRD complex with its core member proteins. Right, a schematic representation of the class II CHD of proteins (CHD3, green; CHD4, blue; CHD5, red) with their functional domains shaded in grey. **B**) Expression patterns of *CHD3*, *CHD4*, and *CHD5* in the neocortex, based on the developmental human RNA-seq dataset of BrainSpan (http://www.brainspan.org/). The lines show loess curves fitted

legend continues on next page

homozygous loss-of-function frameshift variant, and cell lines that underwent the gene editing procedure without the introduction of mutations at the target site (Figure 2A; unedited clone 1 (UE C1), unedited clone 2 (UE C2), heterozygous clone 1 (HET C1), heterozygous clone 2 (HET C2), homozygous clone 1 (HOM C1) and homozygous clone 2 (HOM C2)). We designed a guide RNA that induced a double stranded DNA break in exon 3 of the *CHD3* gene, targeting all three major isoforms (Figure 2B). Non-homologous end joining caused either a 1 bp deletion (in cell lines HET C1, HOM C1, HOM C2; c.298delG, p.G100Vfs*40; NM_001005273.2), or a 1 bp insertion (in cell line HET C2; c.298insA, p.G100Wfs*53; NM_001005273.2) at the target site (Figure 2D), in each case resulting in the formation of a premature stop-codon downstream. Selected clonal cell lines were tested for their chromosomal integrity using molecular karyotyping (Figure S2). All lines shared a 22q11.23 microduplication of approximately 1.3 Mb which was already present in the original cell line (parental: P). Furthermore, HET C2 carried a 2 Mb 1q32.1 gain (Figure S2), a commonly found aberration in stem cell culturing[25]. With the used gene-editing design, we found nineteen predicted CRISPR-Cas9 off-targets, of which only one was located in an exonic region (Table S1). We prioritised the exonic off-target, and randomly selected four intergenic predicted off-targets for screening by Sanger sequencing (5/19; Figure S3). Furthermore, the iPSC lines were tested for the expression of pluripotency makers using qPCR and immunostainings, successfully confirming their stem cell state (Figure S4).

Next, we derived cerebral organoids from these cell lines (P, UE C1, UE C2, HET C1, HET C2, HOM C1 and HOM C2), with a slightly adjusted culturing protocol that increases the time of neural induction (till day 12) and starts later with the addition of Vitamin A to the culture medium (day 18) as compared to previously published protocols[20; 22] (Figure 2C). This way, the neuroepithelium, which is visible as a bright edge with bright-field microscopy, had more time to develop during the extended neural induction period. This allowed for better quality control of the embryoid bodies before proceeding to Matrigel embedding. Analysing day-50 organoids, we confirmed reduced transcript levels of *CHD3* in lines with a heterozygous or a homozygous frameshift variant, showing a clear dosage effect (Figure 2E). These findings indicate that, as expected, the mutated transcripts are subject to nonsense-mediated decay. Moreover, investigating protein levels via Western blotting in day-30 organoids, we found that those with a homozygous knockout mutation had a clear lack of CHD3, while organoids derived from cell lines with a heterozygous *CHD3* mutation had a half-dosage of the protein compared to the parental and unedited cell lines (Figure 2F). These experiments confirm that the frameshift variants introduced into *CHD3* efficiently disrupt expression at both the transcript and protein level.

---

data points. **C**) Expression patterns of *CHD3*, *CHD4*, and *CHD5* in different cell populations in the developing human brain based on 12-13 post conception week (pcw) foetal cortex single-cell RNA-seq data. **D**) Expression patterns of *CHD3* in different cortical regions based on the BrainSpan dataset. **E**) A representation of the developing human cortex, with the radial glia cell in the ventricular zone (VZ) in green, shifting towards the intermediate zone (IZ) when undergoing cell division, intermediate/basal progenitors in the subventricular zone (SVZ) in purple, and post-mitotic neurons in blue residing in the cortical plate (CP). Based on available transcriptomic datasets, the expected expression patterns of *CHD3*, *CHD4*, and *CHD5* are indicated on the right side.

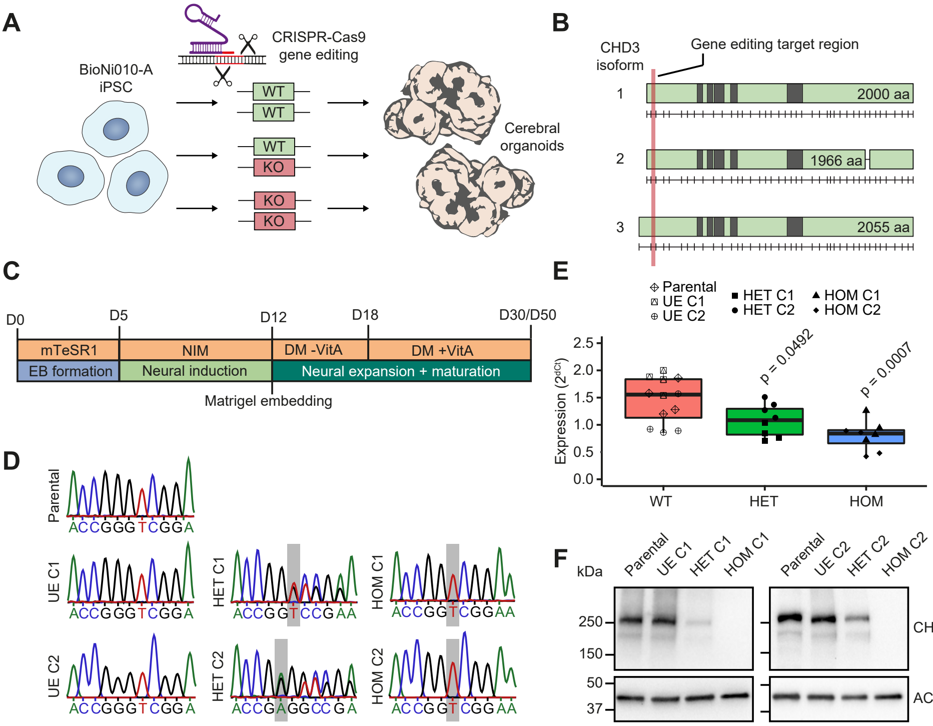


Figure 2. Disruption of *CHD3* expression in cerebral organoids. A) Overview of the gene editing approach used in this study, creating control cell lines that remained unedited after the CRISPR-Cas9 procedure and cell lines carrying a heterozygous or homozygous *CHD3* frameshift variant. **B**) Protein schematic of the three major isoforms of CHD3, with functional domains shaded in grey and the exon structure indicated below. The exon that was targeted by the CRISPR-Cas9 guide RNA is highlighted in red. **C**) The cerebral organoid culture protocol, described in detail in Methods section. **D**) Sanger traces of the CRISPR-Cas9 target region in clonal cell lines selected for further characterisation. The unedited clones show no differences from the parental target region, while the HET clones carry a heterozygous out-of-frame variant, and the HOM clones a homozygous 1 bp deletion causing a frameshift. **E**) Boxplot representing the transcript levels of *CHD3* in day-50 cerebral organoids, assayed using qPCR (*n* = 8-12, *p*- values compared to wildtype condition (WT), one-way ANOVA and post-hoc Dunnett's test). **F**) Immunoblot of whole-cell lysates of three day-30 cerebral organoids pooled together expressing CHD3 protein. Expected molecular weight is ~226 kDa. The blot was probed for ACTB to ensure equal protein loading.

iPSCs lacking CHD3 expression have normal rates of growth

Next, we assessed the morphology and growth rates of cerebral organoids derived from cells lacking the expression of one or both *CHD3* alleles. Regardless of genotype, all organoids had a similar appearance to each other during the first fifteen days of the protocol (Figure 3A and Figure S5). Measuring the surface area of the organoids, we did not find the *CHD3* genotype to influence the initial growth of the organoids (Figure S5). These results show that *CHD3* has little impact on the earliest stages of neuroepithelium formation and initial expansion, consistent with our expectations based on expression patterns of the gene in the human developing brain (Figure 1).

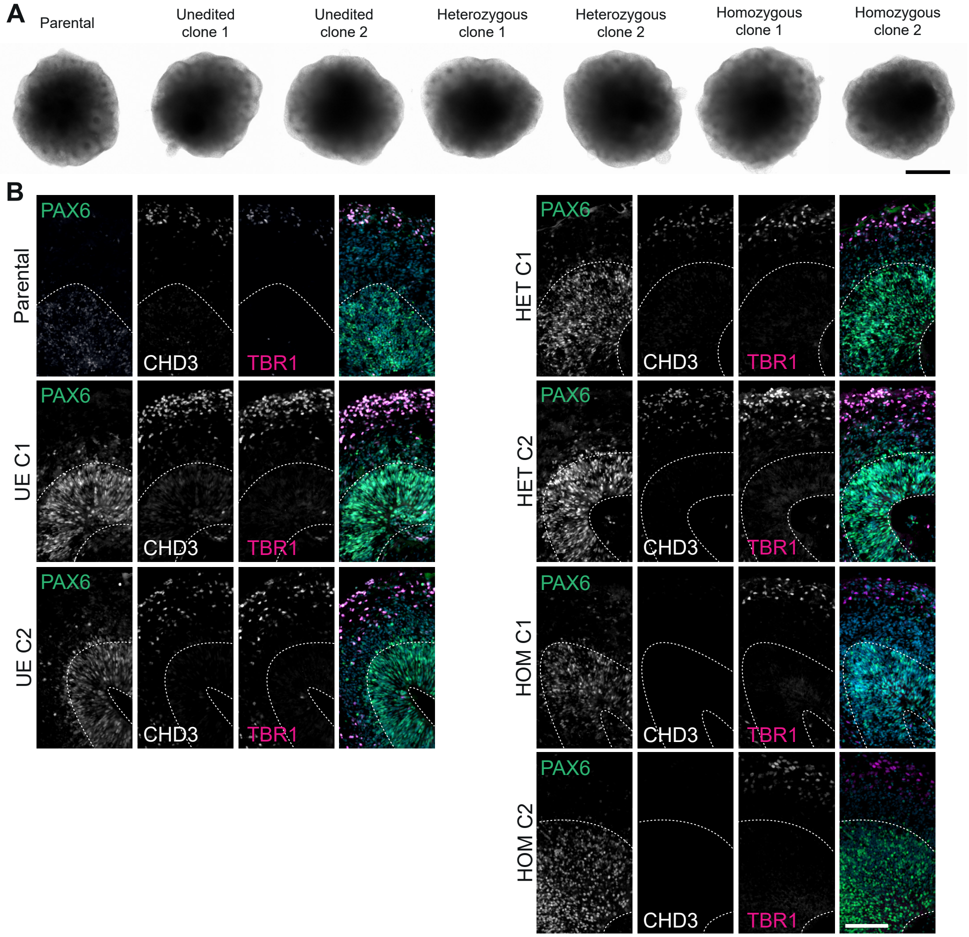


Figure 3. Growth and organisation of cerebral organoids. A) Representative bright-field images of day-15 cerebral organoids for each cell line. Scale bar = 500 µm. **B**) Immunohistochemistry micrographs of day-57 cerebral organoid ventricles for radial glia cells (PAX6, green), post-mitotic neurons (TBR1, magenta) and CHD3-positive cells (CHD3, white). Scale bar = 100 µm.

Cerebral organoids lacking *CHD3* expression contain both NPCs and mature neurons

Based on the hypothesis that it acts as a potential pro-neural factor, we sought to examine whether lower levels or a complete lack of *CHD3* would affect the generation of mature neurons. We performed immunostainings on day-50 cerebral organoids, probing for PAX6, a marker of radial glial cells that reside in the ventricular zone in the developing foetal cortex, and TBR1, a marker of early post-mitotic neurons from the cortical plate. Our immunostainings confirmed a clear loss of CHD3 expression in sections of homozygous knockout organoids, but in all genotypes we observed both PAX6- and TBR1-positive cells that organised in two subregions within the organoid ventricles, resembling the ventricular zone and the cortical plate (Figure 3B, Figure S6). Although we could not discount the possibility of subtle differences in numbers of NPCs and/or neurons between the genotypes, or effects on neuron function, these data indicate that CHD3 is not required for the differentiation of NPCs into TBR1-positive neurons.

Transcriptomic analyses reveal subtle differences between wildtype and *CHD3* knockout organoids

As *CHD3* encodes a chromatin remodelling protein, we expected that disruptions of this gene might change the dynamics of chromatin state, and subsequently, gene expression. Therefore, we performed bulk transcriptomics on four independent batches of day-50 organoids for each cell line. Principal component analysis and hierarchical clustering showed that samples from the same cell line clustered closely together, with no clear batch-to-batch effects (Figure 4A, Figure S7B). Moreover, while the variation in the transcriptomes of organoids grown from the parental, unedited and heterozygous *CHD3* knockout cell lines overlapped based on the first two principal components (Figure 4A), the organoids with the homozygous *CHD3* knockout genotype clustered away from the wildtype and heterozygous knockout samples. However, organoids derived from the two independent homozygous *CHD3* knockout clones did not show overlap in the principal component analysis, indicating that there is still significant variability between clones that is unrelated to *CHD3* genotype.

When performing differential gene expression analysis based on genotype (model: ~genotype, grouping all samples with the same genotype), we found a larger number of significant differentially expressed genes in the homozygous *CHD3* knockout versus wildtype comparison (88 genes with $p < 0.1$ and Log_2(fold change) > 0.6 or < -0.6), than when we compared heterozygous *CHD3* knockout to wildtype samples (20 genes with $p < 0.1$ and Log_2(fold change) > 0.6 or < -0.6) (Figure 4B and 4C, Figure S7C, Table S2). Based on normalised count data, *CHD3* expression levels seemed to be decreased in both heterozygous and homozygous knockout samples, consistent with our prior qPCR results from the same samples (Figure 2E), but the difference was only significant in the homozygous versus wildtype comparison (Figure 4B and Figure S7D). Observationally, based on bulk RNA-seq data the expression of *CHD4* (but not *CHD5*) appeared to show increased levels in the *CHD3* knockout cell lines, although differences were not significant (Figure S7D). As this could indicate a potential compensatory mechanism by CHD class-II family members, we followed up with qPCR using the same samples and confirmed the increased expression of *CHD4* in *CHD3* knockout organoids (Figure S8). Expression levels of well-described neural progenitor and neuronal markers were unchanged between the genotypes (*PAX6*, *EOMES*, *TBR1*; Figure S7D). These results show that there are differences between organoids with a heterozygous or homozygous *CHD3* knockout compared to the wildtype genotype at the transcriptome level. As expected, these differences seem more pronounced in cells that completely lack *CHD3* expression than in cells that have one functional allele. However, given the small number of differentially expressed genes, the consequences of *CHD3* loss seem subtle. Furthermore, in gene ontology analyses of differentially expressed genes, we did not identify well-established roles in neuronal development or convergence on specific pathways (Table S2).

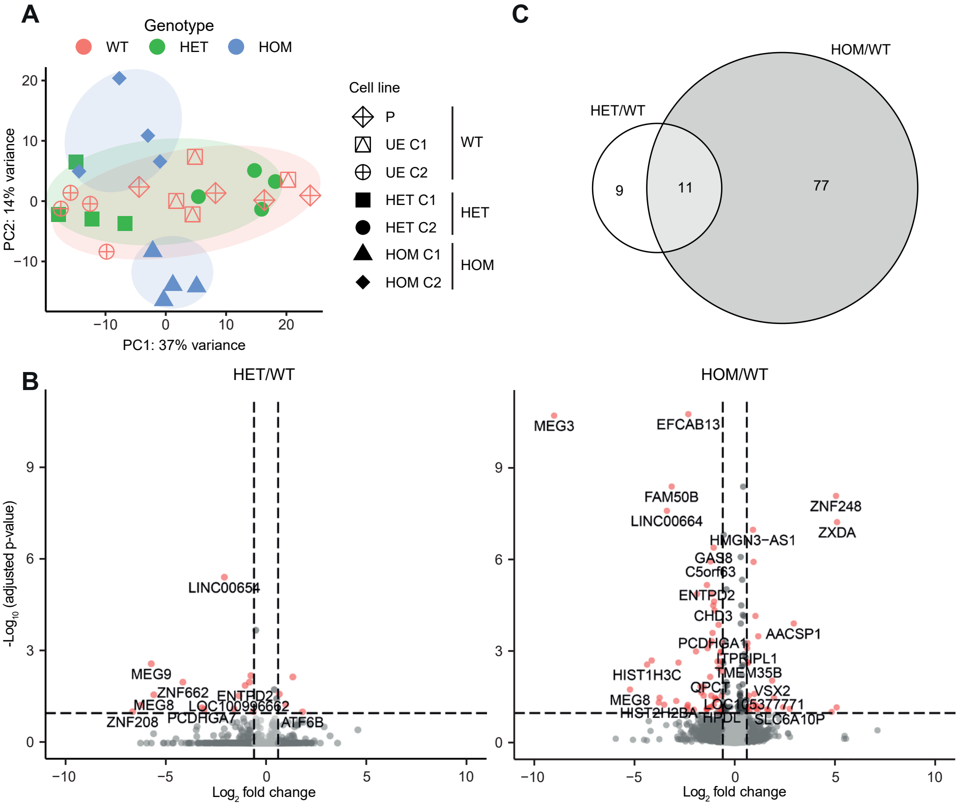


Figure 4. Bulk RNA-seq of day-50 cerebral organoids. A) Graph plotting principal component 1 (explaining 37% of the variance between the samples) and principal component 2 (explaining 14% of the variance between the samples), based on variance stabilising transformation normalised counts. Samples from cell lines carrying a homozygous *CHD3* knockout mutation (blue) cluster away from samples with a wildtype (red) or heterozygous *CHD3* knockout (green) genotype. **B**) Vulcano plots with significant differentially expressed genes (p-value < 0.1) and 0.6 < $\log_2$ fold change < 0.6 shaded in red, for both heterozygous *CHD3* knockout (HET) versus wildtype samples (left) and homozygous *CHD3* knockout (HOM) versus wildtype samples (right). **C**) Venn diagram showing the number of significant differentially expressed genes for the HET versus wildtype and HOM versus wildtype comparisons.

Disruption of *CHD3* increases the neural progenitor pool

To further examine cellular composition of the organoids, we performed single-cell RNA-seq for each cell line. We obtained data for 14,520 cells in total, of which 6,738 cells were from wildtype organoids (representing two batches of the parental cell line, and one batch for UE C1 and UE C2), 2,434 cells were from heterozygous *CHD3* knockout organoids (representing HET C1 and HET C2) and 5,348 cells were from homozygous *CHD3* knockout organoids (representing HOM C1 and HET C2) (Figure 5A). These cells clustered to nine different clusters that were annotated in a semi-automated way by mapping brain-related markers for each cluster to a cell type database (Figure 5A, Table S3)[26]. Cells from each sample were represented in all the nine clusters, independent from genotype or the total number of cells that were called (Figure S8A). As expected, *CHD3* transcript levels were decreased in cells with a heterozygous or homozygous *CHD3* knockout mutation in a dosage-dependent manner in all clusters (Figure 5B).

We identified two clusters with a neural progenitor identity, highly expressing cell cycle markers including *TOP2A*, *UBE2C* and *PCNA* (NPC1), and *CCNB2*, *CDC20* and *PTTG1* (NPC2), representing actively dividing progenitor cells (Figure 5C and Figure S9B). In addition, we found two more clusters that contained neural progenitor cells. The largest of these two clusters showed high expression of *HES1*, *FABP7* and *GLI3*, indicating a radial glia identity (RG), while the smaller cluster expressed *EOMES*, *TCF4* and *EMX1*, which are markers of intermediate progenitors (IP) (Figure 5C). Two clusters (N1 and N2) expressed markers of neuronal cells. N1 was defined by high expression of *DCX*, a marker of immature neurons, *ZFHX3*, a gene with a role in regulating neuronal terminal differentiation, and *DLX2*, a marker of inhibitory neurons (Figure 5C), suggesting that this cluster had a mixture of early-born neurons and inhibitory neurons. In contrast, markers of N2 were found to be *TBR1*, *NEUROD2* and *NEUROD6*, among others, pointing to a more mature and glutamatergic identity (Figure 5C). Two clusters could not be annotated (Table S3), and therefore may contain cells with a non-neuronal identity. A small cluster was annotated by both astrocytes and microglia labels, based on expression of *FTL*, *GADD45B* and *HLA-B* (AC/MG; Figure S9B).

Interestingly, cells with different genotypes were not equally distributed over these clusters. We found that cells from wildtype organoids had a higher fraction of neuronal cells, in particular from cluster N1. In contrast, cells from heterozygous and homozygous *CHD3* knockout organoids had larger fractions of neural progenitor clusters, and the increase seemed dependent on the number of disrupted *CHD3* alleles (Figure 5D and Figure S9C). When we estimated cell type abundances in the bulk transcriptomic samples, using a cell type signature matrix derived from the single-cell data, we also detected a higher fraction of neural progenitor clusters in the homozygous *CHD3* knockout organoids (Figure S10). These changes in cellular populations are in line with the hypothesis that *CHD3* may function as a switch for neuronal differentiation, such that lack of expression of this gene causes an accumulation of neural progenitors.

Lack of *CHD3* leads to enhanced expression of early progenitor markers and a decrease in positive regulators of neuronal differentiation

To assess gene expression differences within these clusters, we performed differential gene expression analyses. Consistent with the bulk transcriptomic data, we found that differences in gene expression between heterozygous *CHD3* knockout and wildtype cells were smaller than those between homozygous knockout and wildtype cells (Figure 4C and 6A). The numbers of differentially expressed genes in cluster NPC1 and N1 were the largest. In these two clusters, we found that lack of *CHD3* was associated with increased expression of genes involved in negative regulation of neuronal differentiation, such as *PAX6*, *VIM*, *PTN* and *TTYH1* (Figure 6B), accompanied by decreased expression of genes positively associated with neuronal differentiation, including *BCL11B*, *SOX4*, and *NR2F1* (Figure 6B). Other genes showing enhanced expression due to *CHD3* loss in multiple clusters included several that have been implicated in neural progenitor maintenance, such as *ID4*, *LHX2*, *EMX1*, while positive regulators of differentiation *ZEB2*, *DCX* and *RORB* were found to be decreased (Figure 6C and 6D, Figure S11A). These results suggest that

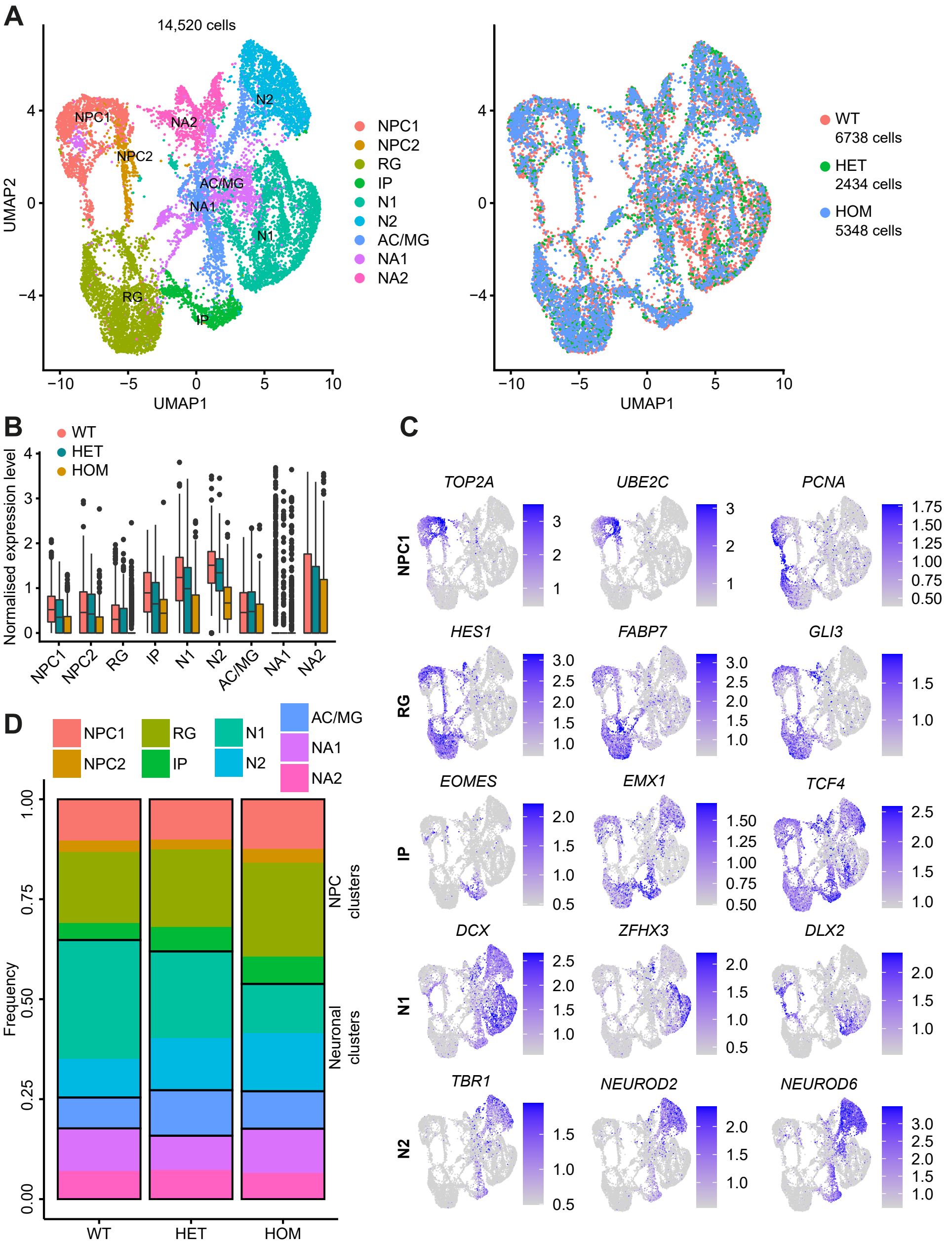


Figure 5. The effects of *CHD3* disruption on cell composition of cerebral organoids. A) Left, scatter plot of 14,520 cells from day-57 cerebral organoids after single-cell RNA-seq principal components analysis and uniform manifold approximation projection (UMAP) embedding with colouring based on the identified clusters. Right, colouring based on genotype. **B**) Boxplots of the expression levels of *CHD3* based on normalised RNA counts for each cluster. **C**) Expression of marker genes for each cluster. The gene-specific contrast levels were based on quantiles of non-zero expression (minimum = q10, maximum = q90). **D**) Relative cellular distribution for each genotype. NPC1, NPC2, RG and IP were considered NPC clusters, while N1 and N2 were grouped as neuronal clusters.

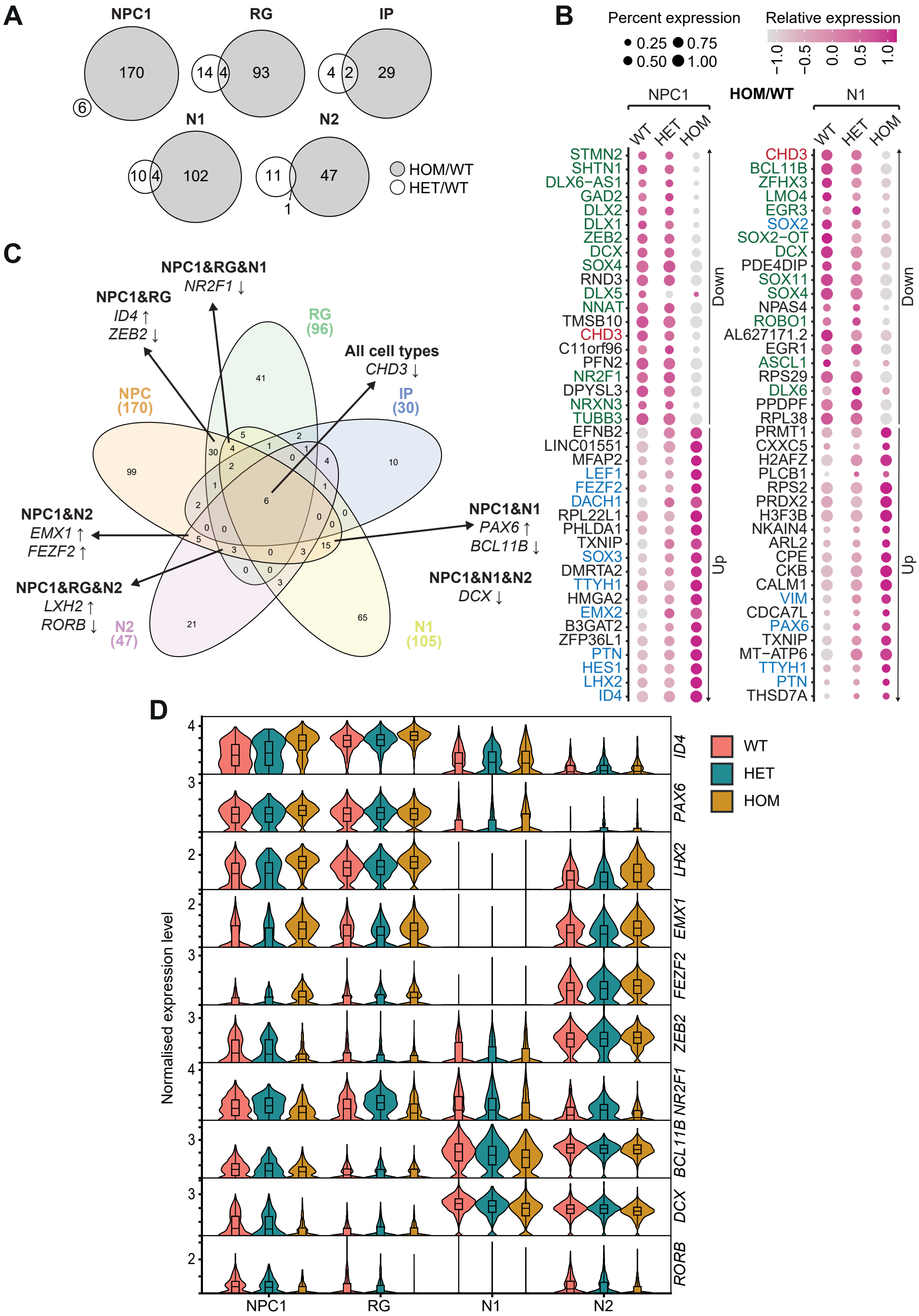


Figure 6. The effects of *CHD3* disruption on gene expression in NPCs and neurons in day-57 cerebral organoids. A) Venn diagrams showing the number of significant differentially expressed genes for the HET versus wildtype and HOM versus wildtype comparisons in different clusters. **B**) Dot plot showing the twenty strongest upregulated and the twenty most downregulated significant differentially expressed genes

legend continues on next page

CHD3 promotes neuronal differentiation, and that lower levels or lack of expression of this gene leads to a delay in differentiation of progenitor cells.

Discussion

As core members of the chromatin remodelling NuRD complex, class-II CHD proteins fulfil important roles in controlling spatiotemporal gene expression patterns during development[19; 27; 28]. We provided new insights into one of three members, *CHD3*, by introducing frameshift variants using CRISPR-Cas9 gene editing in human iPSCs, growing these cells into cerebral organoids, and analysing their development with immunohistochemistry, RNA sequencing, and single-cell transcriptomics. Our findings show that CHD3 acts as an upstream regulator of genes inhibiting neuronal differentiation as well as pro-neural genes, during early human brain development. Hence, we uncover a role for the gene within molecular networks that control the balance between NPC self-renewal and neurogenesis.

Consistent with prior studies in mice[19], our human cell data show that CHD3 expression is typically lower in progenitor cell types, and increases during neuronal differentiation, with the highest expression in post-mitotic neurons. *Chd3* knockdowns in mice were reported to result in delayed radial migration of differentiating neurons from the ventricular zone into the cortical plate and a shift towards more deep layer versus upper layer neurons, but no accumulation of NPCs was noted[19]. Our results suggest that CHD3 may already play important roles earlier during neurogenesis, controlling when cells exit the cell cycle and begin terminal differentiation. Beyond the possibility of species-related differences in functions of mouse and human orthologues, differences in the techniques used could potentially explain discrepancies between the earlier study and the present one. In the prior work, *Chd3* was knocked down using a short hairpin RNA at E13.5, when neurogenesis has already started, while we disrupted *CHD3* at the genomic level from the beginning of the experiments. By reducing *Chd3* levels only at a later stage of development, when some NPCs may already have passed the switch towards differentiation, earlier effects may have been missed. Moreover, in the mouse study, the embryonic brains were analysed at E18.5, after all cortical layers have formed. For the day 50-57 human cerebral organoids that we investigated in our work, the ventricles mainly contain neurons expressing deep layer markers, and very few neurons express upper layer markers (Figure S11B). The deep and upper layer neurons of these organoids do not organise themselves in clear separate layers, as observed in mouse and human foetal brains. Therefore, day 50-57 cerebral organoids are not well-suited to look at differences in the formation of deep and upper cortical layers.

---

in the NPC1 and N1 clusters between homozygous *CHD3* knockout and wildtype organoids. *CHD3* is highlighted in red. Genes that positively regulate neuronal differentiation are depicted in green, while genes associated with NPC maintenance are shown in blue. **C**) Venn diagram showing the overlap between differentially expressed genes in the NPC1, RG, IP, N1 and N2 clusters between *CHD3* knockout and wildtype organoids. Examples of genes differentially expressed in multiple clusters are highlighted. **D**) Violin plots for a selection of genes differentially expressed in multiple clusters between *CHD3* knockout and wildtype organoids in NPC1, RG, N1 and N2. *ID4*, *PAX6*, *LHX2*, *EMX1* and *FEZF2* are upregulated in *CHD3* knockout organoids, while *ZEB2*, *NR2F1*, *BCL11B*, *DCX* and *RORB* are downregulated.

A recent study that analysed data from 58,145 telencephalic/cortical human foetal cells identified multiple clusters of neural progenitor cells[29]. One of these clusters was marked most highly by the gene *ID4*. Notably, this progenitor cluster did not have a corresponding cluster in single-cell data from mice, but seemed unique to the human dataset[29]. Investigations of the functions of a human-specific gene, *NOTCH2NL*, have found that overexpression of this gene in mouse brain organoids leads to a delay in neuronal differentiation and increased expression of genes related to NPC maintenance, including *Id4*, *Lhx2*, *Ttyh1* and *Fezf2*, among others[30]. Moreover, in another study that employed brain organoids to investigate human-specific neurodevelopmental mechanisms, the researchers identified decreased expression of *ZEB2* in human versus gorilla organoids at early stages, associated with a delay in neuronal differentiation[31]. In our homozygous *CHD3* knockout organoids we also identified increases in *ID4*, *LHX2*, *TTYH1* and *FEZF2* expression, while we observed a decrease in *ZEB2*. Hence, our conclusion that CHD3 may serve as a switch for neurogenesis seems consistent with related studies that focused on early human brain development. Whether *CHD3* or its promoter/enhancer regions underwent any evolutionary changes in recent evolution, or whether its target regions may have acquired changes in humans versus other species, potentially impacting the timing of neurogenesis, have not yet been studied.

While we observed effects of a complete lack of *CHD3* expression on cell composition and gene expression in our cerebral organoid model, differences between wildtype and heterozygous *CHD3* knockout organoids were much smaller. Many differentially expressed genes identified in the NPC and neuronal clusters in our single-cell dataset showed a dosage effect dependent on *CHD3* expression levels. This dosage effect was also reflected in the number of differentially expressed genes in both the bulk RNAseq and single-cell RNAseq data. However, most of the differences that we found between wildtype and homozygous *CHD3* knockout cells were too small to be reliably detected in our comparisons between wildtype and heterozygous *CHD3* knockouts. While an increase in the number of cell lines and/or batches could potentially increase the power of our experimental design and resolve this issue, our data show that the effects of a heterozygous loss of *CHD3* on early neurodevelopmental processes should be subtle.

Interestingly, the large majority of *de novo CHD3* variants identified in individuals with a neurodevelopmental disorder are missense variants[11; 32]. Heterozygous *CHD3* variants with a predicted loss-of-function effect have mostly been identified in familial cases, with variable expressivity and/or reduced penetrance as an underlying mechanism, suggesting that a second hit or additional mutational load may be required for a *CHD3* loss-of-function variant to result in a neurodevelopmental phenotype (Chapter 4 of this thesis). Moreover, to our knowledge no cases have so far been reported of people carrying homozygous loss-of-function variants in this gene. Hence, our genetically engineered homozygous knockout cell lines are not modelling the genetic condition of individuals with CHD3-related disorder. However, by providing insights into mechanistic pathways the organoid data can enhance our understanding of disorder. For example, the increase in the NPC populations and delay of neuronal differentiation in our knockout organoids appear consistent with macrocephaly as one of the characterising features of CHD3-related disorder[11]. Therefore, some of the downstream effects of a complete disruption of the gene may overlap with pathways that are implicated in *CHD3*-related disorder.

In addition, while we expected the largest effects of decreased *CHD3* expression in post-mitotic neuronal populations, we identified most differences in cycling NPCs (cluster NPC1) and immature/inhibitory neurons (cluster N1). These findings suggest that, although the gene is only expressed at low levels during these early stages, disruption of *CHD3* may have the strongest effects on progenitor cells and immature neurons. For future modelling of *CHD3*-related disorder using CRISPR-Cas9 engineered cell lines and patient-derived cells, it may therefore be possible to screen for effects on brain development at earlier time points of cerebral organoid culturing. Performing the experiments in day-30 organoids instead of day-50-57 organoids would significantly decrease culturing times, and potentially speed up future studies.

Our bulk transcriptomic analyses hinted at an upregulation of *CHD4* expression in *CHD3* knockout organoids. Moreover, in the single-cell transcriptomics data, *CHD4* showed significantly increased expression in the neuronal N2 cluster as a consequence of *CHD3* loss (Figure S11C). Elevated *CHD4* expression could have in part reflected the increased number of NPCs in the *CHD3* knockout organoids, given the different expression patterns of these genes in NPCs and neurons. However, our finding that *CHD4* levels are only significantly increased in neuronal cells of the knockout organoids, where *CHD3* would normally have been most highly expressed, points towards a potential compensation mechanism to maintain the availability of (alternative) NuRD complexes. In previously reported work, *Chd4* disruption in satellite cells from mouse skeletal muscles induced compensatory upregulation of *Chd3* and *Chd5* as well[33]. It remains unclear whether these effects involve a disrupted inhibitory feedback loop of class-II CHDs onto each other. Although CHD3 and CHD4 have distinct functions during development, we also do not know if a CHD4-NuRD complex can mitigate the loss of *CHD3* expression and subsequently the loss of CHD3-NuRD complexes. Based on our data, it would be interesting to study *CHD4* expression levels in cases of *CHD3*-related disorder, in particular in individuals with loss-of-function variants.

Although our work made use of a well-characterised commercial iPSC line from a healthy male individual, during the quality controls of our CRISPR-Cas9 experiments we discovered a 1.3 Mb microduplication involving 22q11.23 in all the cell lines, including the original line obtained from the company. The microduplication had not been detected in the prior G-banding and comparative genomic hybridisation arrays of this line, most likely since it is on the edge of the resolution that can be observed with the techniques used. This shows the importance of examining genomic integrity of cell lines at high resolution for research involving iPSC models. 22q11 microduplications, some of which span the region duplicated in the cell line that we used, have been described in neurodevelopmental conditions with reduced penetrance and variable expressivity[34]. Our iPSC line came from a documented healthy donor without neurodevelopmental issues, so it seems very likely to have been an unaffected carrier. Our analyses involved comparisons of CRISPR-Cas9 induced *CHD3* knockout alleles to unedited and original lines all with identical genetic background, meaning we can be confident that observed differences relate to altered *CHD3* expression, rather than copy number variant status. However, we cannot completely exclude the possibility that the effects we observe might be modified by presence of the 22q11.23 microduplication. Future work could resolve this point by assessing the effects on another genomic background.

Overall, our results identify CHD3 as a potential regulator of NPC proliferation and neurogenesis. As homozygous *CHD3* knockout cells are still able to differentiate into post-mitotic neurons, the gene seems to be involved in the timing and facilitation of neurogenesis, rather than being an essential factor for differentiation. In future studies, comparing gene expression profiles of different genotypes at multiple time points will help to better establish whether decreased *CHD3* levels cause a delay in neurodevelopment. Experiments that assess neuronal functions, such as calcium imaging or multi-electrode arrays, could yield insights into whether neurons that are ultimately formed in the absence of *CHD3* are fully functional. Furthermore, establishing direct targets of CHD3, and uncovering effects on DNA accessibility, will be crucial to understand how this important regulatory factor modifies the molecular pathways to control the balance between progenitor state and neuronal differentiation.

Methods

Cell line and cell culture

The BIONi010-A (K1P53) iPSC line (male, 15-19y, European Bank for induced pluripotent Stem Cells) derived from a healthy donor[24], was cultured on plates coated with Matrigel (Corning) in mTeSR1 medium (Stem Cell Technologies) at 37 °C with 5% CO2. Medium was replaced daily and cells were passaged using Versene solution (Gibco) when confluency of 70-80% was reached. The chromosomal integrity of the cell line was confirmed by Cell Guidance Systems with an aCGH array before initiating CRISPR-Cas9 experiments.

Gene editing with CRISPR-Cas9

In order to target exon 3 of the human *CHD3* gene, present in all three major isoforms (NM_005852.3, NM_001005273.2 and NM_001005271.2), we designed a guide RNA using CRISPOR[35] with a high specificity (0.97), a high predicted efficiency (0.62-0.70) and a small number of predicted off-targets (19 off-targets, Table S1): 5'-AATATGGAACCGGACCGGGT CGG-3'. BIONi010-A cells were pre-treated with Y-27632 (10 µM; Selleckchem) for 30 min, disassociated using TrypLE Express (Gibco), and passaged through a 40 µm strainer to obtain a single-cell suspension. The guide was delivered as an Alt-R CRISPR-Cas9 sgRNA (IDT) after forming a protein complex with the Alt-R *Streptococcus pyogenes* Cas9 Nuclease V3 (IDT), using the P3 Primary Cell 4D-Nucleofector TM X kit (Lonza Biosciences) in combination with the Alt-R Cas9 Electroporation Enhancer (IDT). The electroporation was performed with an AMAXA 4D CoreUnit (CA137 programme; Lonza Biosciences). Cells were maintained in mTeSR1 medium supplemented with 10 µM Y-27632 for 4-6 days and afterwards passaged one time to recover from the electroporation. To isolate colonies derived from single cells, cells were disassociated with TrypLE, passaged through a 40 µm strainer and seeded at a low density in Matrigel-coated 100 mm dishes in mTeSR1 supplemented with 1x CloneR (Stem Cell Technologies). After 7-8 days, iPSC colonies were manually picked and transferred to a 96-well plate in medium supplemented with 1x CloneR until ready to passage. Clones were split to three wells, to prepare cryovials for freezing in mFreSR (Stem Cell Technologies) and to isolate gDNA using Squishing buffer (10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 25 mM NaCl, 1:5 1 mg/ml Proteinase K) to screen for the introduced mutations and CRISPR-Cas9 off-target effects.

Screening of CRISPR-Cas9 edited cell lines

The iPSC clones were screened for mutations introduced by CRISPR-Cas9 gene editing in exon 3 of the *CHD3* gene by amplifying the target region of the guide RNA using PCR. For each clone, a PCR reaction was prepared with isolated gDNA, Phusion Green Hot Start II High-Fidelity PCR Master Mix (Thermo Fisher) and primers annealing to the target region (Table S5). PCR products were sent for Sanger sequencing (Eurofins Scientific) and the resulting Sanger traces were analysed using the ICE CRISPR analysis tool[36] to identify heterozygous and (compound) homozygous out-of-frame mutations. Positive clones were selected for expansion and further characterisation. To assess off-target effects, five off-targets were selected for screening with PCR followed by Sanger sequencing (Table S1; primers used: Table S5). All selected CRISPR-Cas9 edited clones underwent molecular karyotyping using the KaryoStat HD Assay (Thermo Fisher).

Cerebral organoid differentiation

Cerebral organoids were cultured as previously described[22], with minor adjustments. Single-cell suspensions were prepared using TrypLE, and 9000 cells were seeded per well in an ultra-low attachment U-bottom 96-well plate (Corning) in mTeSR1 medium supplemented with 50 μM Y-27632. The day of seeding was considered day 0 (Figure 2A). Half of the medium was replaced every other day. On day 5, neural induction was started by changing the medium to neural induction medium: DMEM/F12, 1x N2, 1x GlutaMAX (all Gibco), 1x minimum essential media-nonessential amino acids (MEM-NEAA) and 1:1000 1 mg/ml heparin from porcine intestinal mucosa (both Sigma). At day 12, embryoid bodies were transferred to drops of 20 μL of ice-cold Matrigel, and the Matrigel was allowed to solidify at 37 °C for > 30 min. Afterwards the embedded embryoid bodies were transferred to 60 mm dishes in differentiation medium: 50% DMEM/F12, 50% Neurobasal medium, 0.5x N2, 1x B27 minus VitA, 1x GlutaMAX, 0.5x MEM-NEAA, 50 μM 2-Mercaptoethanol, 1x Pen/Strep, 1:4000 human insulin (9.5-11.5 mg/mL; Sigma). The next day, the dishes were moved to a CO2 Resistant Shaker (Thermo Fisher), and organoids were cultured for the rest of the protocol under shaking conditions (40 rpm, orbit of 19 mm). Medium was replaced every other day. On day 18, the medium was changed to differentiation medium containing B27 with VitA (Gibco). On day 50, organoids were collected for RNA isolation by pooling three organoids together and snap-freezing on dry-ice (for qPCR and bulk RNA-seq), or fixed in 4% paraformaldehyde (Electron Microscopy Supplies Ltd) for 30 min at room temperature followed by 90 min at 4 °C. On day 57, four organoids were pooled and prepared for single cell RNA-seq.

Immunostainings

iPSCs were grown on Matrigel-coated coverslips and fixed with 4% paraformaldehyde in the culture medium for 15 min at room temperature. Cells were blocked and permeabilised with 5% horse serum (Vector) and 0.1% Triton-X100 in PBS for 1 h at room temperature. Antibodies were diluted in blocking buffer (5% horse serum in PBS). Primary antibodies were incubated overnight at 4 °C and secondary antibodies for 1.5 h at room temperature. Nuclei were stained with Hoechst 33342 (Invitrogen) before cells were mounted in DAKO fluorescent mounting medium (DAKO). Primary antibodies: rabbit-anti-OCT4 (1:1000, AB19857, Abcam); mouse-anti-SSEA4 (1:500, AB16287, Abcam); goat-anti-SOX2 (1:500, af2018, R&D systems); mouse anti-TRA-1-60 (1:200, AB16288, Abcam). Secondary antibodies: donkey-anti-rabbit Alexa Fluor 488 (1:1000, Invitrogen, A21206); chicken-anti-mouse Alexa Fluor 594 (1:1000, Invitrogen, A21201); donkey-anti-goat Alexa Fluor 647 (1:250, Jackson Immuno Research, 705-605-147). Fluorescence images were acquired with an Axiovert A-1 fluorescent microscope and ZEN Image Software (Zeiss).

Fixed organoids were cryoprotected in 30% sucrose overnight at 4 °C, embedded in Neg-50 Frozen Section Medium (Thermo Fisher) and cryosectioned at 8 µm. Sections selected for immunostainings were rehydrated in PBS at room temperature for 20 min. Antigen retrieval was performed with citric acid buffer (10 mM, pH 6.0) at 65 °C for 20 min. Afterwards, sections were blocked and permeabilised with 5% horse serum and 0.25% Triton-X100 in PBS for 1 h at room temperature. Antibodies were diluted in blocking buffer (5% horse serum and 0.25% Triton-X100 in PBS). Primary antibodies were incubated overnight at 4 °C and secondary antibodies for 2 h at room temperature. Nuclei were stained with Hoechst 33342 (Invitrogen) before cells were mounted in DAKO fluorescent mounting medium (DAKO). Primary antibodies: mouse-anti-PAX6 (1:500, 862002, BioLegend); chicken-anti-TBR1 (1:500, AB2261, Millipore); rabbit-anti-CHD3 (1:500, AB109195, Abcam). Secondary antibodies: donkey-anti-mouse Alexa Fluor 488 (1:1000, Invitrogen, A21202); donkey-anti-rabbit Alexa Fluor 594 (1:1000, Invitrogen, A21207); donkey anti-chicken Alexa Fluor 647 (1:250, Jackson Immuno Research, 703-605-155). Fluorescence images were acquired with an AxioScan Z1 microscope and ZEN Image Software (Zeiss).

RT-PCR and qPCR

Total RNA was isolated from snap-frozen iPSC pellets or pooled cerebral organoids (*n* = 3) using the RNeasy Mini kit (Qiagen). gDNA was removed with on-column incubation with RNase-free DNase I (Qiagen). cDNA was generated from 1-2 µg RNA by reverse-transcription using the SuperScript III Reverse Transcriptase (ThermoFisher) with random primers. qPCR was performed in technical duplicates or triplicates in 10 µl reaction volumes, containing iQ SYBR Green mastermix (BioRad), forward and reverse primers (Table S6) and cDNA, on the CFX Real-Time PCR Detection System (BioRad). *TBP* and *PPIA* were used as internal normalising controls.

Immunoblotting

Whole-cell lysates were prepared by treating three snap-frozen day 30 cerebral organoids pooled together with 1x RIPA buffer (ThermoFisher) supplemented with 1x PIC (Roche) and 1% PMSF (Sigma). Cells were lysed for 20 min at 4 °C followed by centrifugation for 20 min at 12,000 rpm. Samples were loaded on 4–15% Mini-PROTEAN TGX Precast Gels (Bio-Rad) and transferred onto polyvinylidene fluoride membranes. Proteins were resolved on 4–15% Mini-PROTEAN TGX Precast Gels (Bio-Rad) and transferred onto polyvinylidene fluoride membranes. Membranes were blocked in 5% milk for 1 h at room temperature and then probed with rabbit-anti-CHD3 antibody (1:1000; Abcam, ab109195) overnight at 4 °C. Next, membranes were incubated with HRP-conjugated goat-anti-rabbit (1:10,000; Jackson ImmunoResearch) for 1.5 h at room temperature. Bands were visualised with the SuperSignal West Femto Maximum Sensitivity Substrate Reagent Kit (Thermo Fisher) using a ChemiDoc XRS + System (Bio-Rad). Equal protein loading was confirmed by probing with mouse-anti-β-actin antibody (1:10,000; Sigma, A5441), followed by incubation with HRP-conjugated goat-anti-mouse (1:10,000; Jackson ImmunoResearch) and visualisation with the Novex ECL Chemiluminescent Substrate Reagent Kit (Invitrogen).

Bulk RNA-sequencing and analyses

RNA from day 50 cerebral organoids was sent to BGI (Hong Kong) for library preparation and paired-end 150 bp stranded RNA-seq on the BGI DNBseq platform. HISAT2 (version 4.8.2 (GCC))[37] was used to map RNA-seq reads to the human reference genome (GCF_000001405.38_GRCh38.p12; NCBI) with the settings: --phred33, --sensitive, --no-discordant, --no-mixed and --rna-strandness RF. The quality of the aligned data was assessed using Picard tools v2.26.3 and MultiQC v1.11[38], identifying 35.8 to 65.6 million unique reads mapped to the human genome. Next, the reads were counted with the featureCount function of the Subread (v2.0.3) package[39]. For featureCount the gene transfer format file from NCBI (GCF_000001405.38_GRCh38.p12_genomic.gtf) was used as the GTF annotation, and multi-mapping, multi-overlapping and chimeric reads were excluded. DESeq2 v1.30.1[40] was used to normalise the read counts and perform differential gene expression analysis. Principal component analysis and hierarchical relationship analysis was done on the variance stabilising transformed (VST) counts. Our data set included three cell lines with a wildtype genotype (the parental BIONi010-A cell line, UE C1 and UE C2), two with a heterozygote *CHD3* knockout mutation (HET C1, HET C2) and two with a homozygous knockout *CHD3* mutation (HOM C1, HOM C2), with four batches per cell line. The differential gene expression analysis was run with the design: ~ genotype, grouping the cell lines and batches with the same genotype. Differentially expressed genes were filtered for an adjusted *p*-value <0.1 (FDR/Benjamini-Hochberg method). Gene ontology enrichment was done using GOrilla, comparing the lists of differentially expressed genes to a list of all expressed genes[41]. Cell type abundance was estimated using CIBERSORTx[42], by creating a signature matrix based on the single-cell data sets of the cell lines with a wildtype genotype (P, UE C1, UE C2; identifying 300-500 marker genes per cluster), and performing subsequent deconvolution on the DESeq2 library-size normalised count data of the bulk transcriptomic samples.

Single-cell RNA sequencing and analyses

Day-57 organoids were pooled ($n = 4$) and cut in pieces using a sterile blade. The cut organoids were placed in pre-warmed accutase supplemented with 1:2000 DNAseI and 1:2000 RNAase Inhibitor (both NEB), and slowly pipetted up and down using wide-bore pipette tips. After 30-60 minutes, cells were centrifuged for 5 min at 300x g, washed with PBS, and then passed through 30 and 20 µm filters (Miltenyi Biotec). The filtered cell suspension was centrifuged for 5 min at 300x g, and cells were resuspended in 500 µL differentiation medium. The resuspended cells were pipetted on top of a three-layered Percoll (Sigma) gradient which was centrifuged for 5 min at 300x g, to separate the cells from debris. The fraction that contained the cells was centrifuged 5 min at 300x g, and the cell pellet was resuspended in ice-cold PBS with

0.04% BSA. Cells were counted and cell viability was assessed based on Trypan Blue (BioRad) staining. Afterwards, ~7,000 cells per sample were loaded on a Chromium Next GEM Chip G using the Chromium Next GEM Single Cell 3' Kit v3.1 (both 10x Genomics). The dual-index library was prepared using the Library Construction Kit (10x Genomics), and quality control was performed using the RNA 6000 Nano Kit on a 2100 Bioanalyser Instrument (both Agilent). Libraries were sent to Novogene for 150 bp paired-end sequencing on the Illumina Novoseq 6000 platform (400 million reads per sample, 120 Gb clean data per sample). Cellranger v3.1[43] was used for demultiplexing of the data. Then, the count function from Cellranger v6.0.1[43] was used to map the reads to the refdata-gex-GRCh38-2020-A. The filtered feature matrix output was then loaded as a SeuratObject using the Seurat v4 package[44]. Samples were combined in a single object using the Seurat merge function, and after SCTransformation[45] of each sample individually including regression of mitochondrial gene content, the data set was integrated with the wildtype samples set as the reference to increase computational processing speed. The UMAP dimensionality reduction was based on the first thirty principal components, and clusters were identified with a resolution of 0.2. Cluster annotation was performed using the scCATCH package, identifying marker genes per cluster from a curated data set containing genes expressed in brain-related tissues, and mapping these to brain-related cell types[26]. Differentially expressed genes per cluster were identified with the Seurat FindMarkers function (Wilcoxon Rank Sum test), comparing cells with a wildtype genotype (from samples P1, P2, UE C1 and UE C2) with cells with a heterozygous (HET C1, HET C2) or homozygous (HOM C1, HOM C2) genotype. Differentially expressed genes were filtered for an adjusted *p*-value <0.05 (Bonferroni corrected).

Quantification and statistical analyses

Statistical analysis of qPCR data was performed with GraphPad Prism software using a one-way ANOVA followed by a Dunnett's *post-hoc* test. To calculate the total area or embryoid bodies, bright-field images were analysed using ImageJ by applying a threshold followed by automated particle size analysis.

References

1. Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M., Pletikos, M., Meyer, K.A., and Sedmak, G. (2011). Spatio-temporal transcriptome of the human brain. Nature 478, 483-489.
2. Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., and Aiona, K. (2014). Transcriptional landscape of the prenatal human brain. Nature 508, 199-206.
3. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., and Velmeshev, D. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science 358, 1318-1323.
4. Won, H., de La Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., and Lu, D. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538, 523-527.
5. Ronan, J.L., Wu, W., and Crabtree, G.R. (2013). From neural development to cognition: unexpected roles for chromatin. Nature Reviews Genetics 14, 347-359.
6. de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The dynamic landscape of open chromatin during human cortical neurogenesis. Cell 172, 289-304. e218.
7. Nord, Alex S., Pattabiraman, K., Visel, A., and Rubenstein, John L.R. (2015). Genomic

Perspectives of Transcriptional Regulation in Forebrain Development. Neuron 85, 27-47.
8. Kolla, V., Zhuang, T., Higashi, M., Naraparaju, K., and Brodeur, G.M. (2014). Role of CHD5 in Human Cancers: 10 Years Later. Cancer Research 74, 652.
9. Pilarowski, G.O., Vernon, H.J., Applegate, C.D., Boukas, L., Cho, M.T., Gurnett, C.A., Benke, P.J., Beaver, E., Heeley, J.M., Medne, L., et al. (2018). Missense variants in the chromatin remodeler CHD1 are associated with neurodevelopmental disability. Journal of Medical Genetics 55, 561.
10. Carvill, G.L., Heavin, S.B., Yendle, S.C., McMahon, J.M., O'Roak, B.J., Cook, J., Khan, A., Dorschner, M.O., Weaver, M., and Calvert, S. (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nature genetics 45, 825-830.
11. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nature Communications 9, 4619.
12. Weiss, K., Terhal, P.A., Cohen, L., Bruccoleri, M., Irving, M., Martinez, A.F., Rosenfeld, J.A., Machol, K., Yang, Y., Liu, P., et al. (2016). De Novo Mutations in CHD4, an ATP-Dependent Chromatin Remodeler Gene, Cause an Intellectual Disability Syndrome with Distinctive Dysmorphisms. American journal of human genetics 99, 934-941.
13. Parenti, I., Lehalle, D., Nava, C., Torti, E., Leitão, E., Person, R., Mizuguchi, T., Matsumoto, N., Kato, M., Nakamura, K., et al. (2021). Missense and truncating variants in CHD5 in a dominant neurodevelopmental disorder with intellectual disability, behavioral disturbances, and epilepsy. Human Genetics 140, 1109-1120.
14. Vissers, L.E., van Ravenswaaij, C.M., Admiraal, R., Hurst, J.A., de Vries, B.B., Janssen, I.M., van der Vliet, W.A., Huys, E.H., de Jong, P.J., and Hamel, B.C. (2004). Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. Nature genetics 36, 955-957.
15. Bernier, R., Golzio, C., Xiong, B., Stessman, H.A., Coe, B.P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., and Vulto-van Silfhout, A.T. (2014). Disruptive CHD8 mutations define a subtype of autism early in development. Cell 158, 263-276.
16. Hargreaves, D.C., and Crabtree, G.R. (2011). ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. Cell research 21, 396-420.
17. Marfella, C.G.A., and Imbalzano, A.N. (2007). The Chd family of chromatin remodelers. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 618, 30-40.
18. Zhang, Y., LeRoy, G., Seelig, H.-P., Lane, W.S., and Reinberg, D. (1998). The Dermatomyositis-Specific Autoantigen Mi2 Is a Component of a Complex Containing Histone Deacetylase and Nucleosome Remodeling Activities. Cell 95, 279-289.
19. Nitarska, J., Smith, J.G., Sherlock, W.T., Hillege, M.M.G., Nott, A., Barshop, W.D., Vashisht, A.A., Wohlschlegel, J.A., Mitter, R., and Riccio, A. (2016). A Functional Switch of NuRD Chromatin Remodeling Complex Subunits Regulates Mouse Cortical Development. Cell Rep 17, 1683-1698.
20. Camp, J.G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., and Lancaster, M. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proceedings of the National Academy of Sciences 112, 15672-15677.
21. Kageyama, J., Wollny, D., Treutlein, B., and Camp, J.G. (2018). ShinyCortex: Exploring Single-Cell Transcriptome Data From the Developing Human Cortex. Frontiers in Neuroscience 12.
22. Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature 501, 373-379.
23. Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., et al. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature 574, 418-422.
24. Rasmussen, Mikkel A., Holst, B., Tümer, Z., Johnsen, Mads G., Zhou, S., Stummann, Tina C., Hyttel, P., and Clausen, C. (2014). Transient p53 Suppression Increases Reprogramming of Human Fibroblasts without Affecting Apoptosis and DNA Damage. Stem Cell Reports 3, 404-413.
25. Amps, K., Andrews, P.W., Anyfantis, G., Armstrong, L., Avery, S., Baharvand, H., Baker, J., Baker, D., Munoz, M.B., Beil, S., et al. (2011). Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. Nature

Biotechnology 29, 1132-1144.

26. Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., and Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience 23, 100882.
27. Pierson, T.M., Otero, M.G., Grand, K., Choi, A., Graham, J.M., Jr., Young, J.I., and Mackay, J.P. (2019). The NuRD complex and macrocephaly associated neurodevelopmental disorders. American journal of medical genetics Part C, Seminars in medical genetics 181, 548-556.
28. Hoffmann, A., and Spengler, D. (2019). Chromatin Remodeling Complex NuRD in Neurodevelopment and Neurodevelopmental Disorders. Frontiers in Genetics 10.
29. Eze, U.C., Bhaduri, A., Haeussler, M., Nowakowski, T.J., and Kriegstein, A.R. (2021). Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. Nature Neuroscience 24, 584-594.
30. Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M., van den Bout, A., Bishara, A., Rosenkrantz, J.L., et al. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 173, 1356-1369.e1322.
31. Benito-Kwiecinski, S., Giandomenico, S.L., Sutcliffe, M., Riis, E.S., Freire-Pritchett, P., Kelava, I., Wunderlich, S., Martin, U., Wray, G.A., McDole, K., et al. (2021). An early cell shape transition drives evolutionary expansion of the human forebrain. Cell 184, 2084-2102.e2019.
32. Drivas, T.G., Li, D., Nair, D., Alaimo, J.T., Alders, M., Altmüller, J., Barakat, T.S., Bebin, E.M., Bertsch, N.L., Blackburn, P.R., et al. (2020). A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome. European journal of human genetics : EJHG 28, 1422-1431.
33. Sreenivasan, K., Rodríguez-delaRosa, A., Kim, J., Mesquita, D., Segalés, J., Arco, P.G.-d., Espejo, I., Ianni, A., Di Croce, L., Relaix, F., et al. (2021). CHD4 ensures stem cell lineage fidelity during skeletal muscle regeneration. Stem Cell Reports 16, 2089-2098.
34. Wincent, J., Bruno, D.L., van Bon, B.W.M., Bremer, A., Stewart, H., Bongers, E.M.H.F., Ockeloen, C.W., Willemsen, M.H., Keays, D.D.A., Baird, G., et al. (2010). Sixteen New Cases Contributing to the Characterization of Patients with Distal 22q11.2 Microduplications. Mol Syndromol 1, 246-254.
35. Concordet, J.-P., and Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Research 46, W242-W245.
36. Hsiau, T., Maures, T., Waite, K., Yang, J., Kelso, R., Holden, K., and Stoner, R. (2018). Inference of CRISPR Edits from Sanger Trace Data. bioRxiv, 251082.
37. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37, 907-915.
38. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England) 32, 3047-3048.
39. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England) 30, 923-930.
40. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550.
41. Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48.
42. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology 37, 773-782.
43. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nature Communications 8, 14049.
44. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e3529.
45. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology 20, 296.

Supplemental information

- Figure S1	Expression of CHD3 in cerebral organoids.
- Figure S2	Molecular karyotyping of CRISPR-Cas9 gene edited iPSC lines.
- Figure S3	CRISPR-Cas9 off-target analysis.
- Figure S4	Characterisation of CRISPR-Cas9 gene-edited iPSC lines.
- Figure S5	Growth of embryoid bodies.
- Figure S6	Organisation of cerebral organoids.
- Figure S7	Bulk RNA-seq of day-50 cerebral organoids.
- Figure S8	Expression of CHD4 and CHD5 in day-50 cerebral organoids.
- Figure S9	The effects of CHD3 disruption on cell composition of cerebral organoids.
- Figure S10	Cell type abundance estimation in bulk transcriptomics data using CIBERSORTx.
- Figure S11	The effects of CHD3 disruption on gene expression in NPCs and neurons in day-57 cerebral organoids.

- Table S1	Predicted off-targets by CRISPOR 4.97 for the gRNA used to target the CHD3 gene.
- Table S2	Differentially expressed genes of day-50 cerebral organoids based on bulk transcriptomics data.
- Table S3	scCATCH annotation of UMAP identified clusters of the scRNAseq data.
- Table S4	Differentially expressed genes of day-57 cerebral organoids based on single-cell RNAseq data.
- Table S5	Primer sequences for CRISPR-Cas9 target and off-target regions.
- Table S6	qPCR primer sequences.

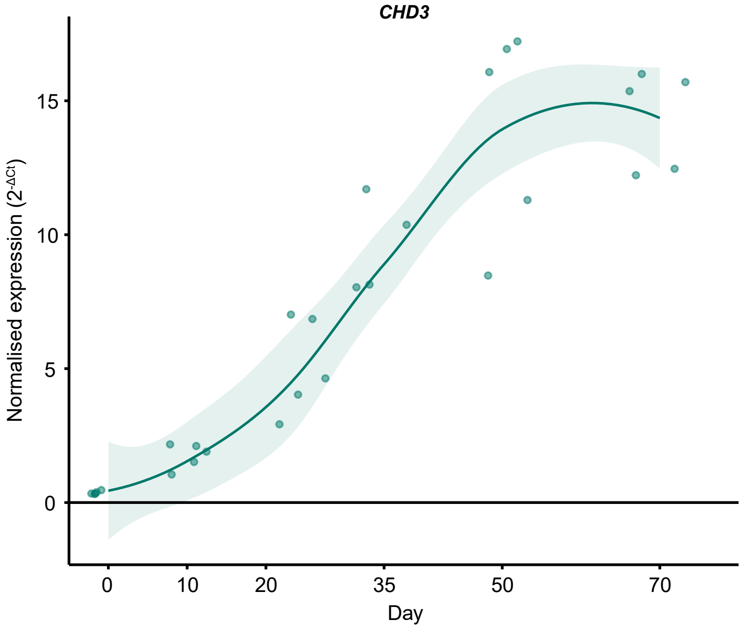
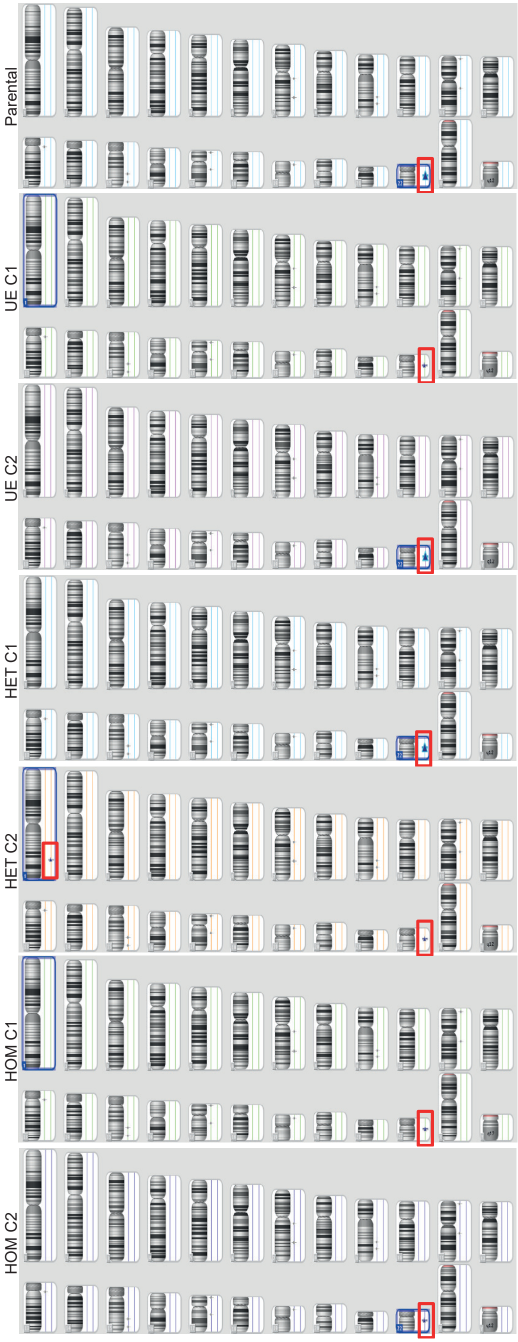


Figure S1. Expression of *CHD3* in cerebral organoids. Transcript levels of *CHD3* in iPSCs and cerebral organoids from day 10 to day 70, assayed using qPCR (*n* = 5). The line shows a loess curve fitted through the data points.



Chromosome	Type	Cytoband Start	CN State	Size (kbp)
22	Gain	q11.23	3.00	1,446

Location of 22q11.23 duplication:
arr[GRCh38] 22q11.23(23,324,045_24,770,353)x3

Chromosome	Type	Cytoband Start	CN State	Size (kbp)
22	Gain	q11.23	3.00	1,331

Chromosome	Type	Cytoband Start	CN State	Size (kbp)
22	Gain	q11.23	3.00	1,317

Chromosome	Type	Cytoband Start	CN State	Size (kbp)
22	Gain	q11.23	3.00	1,331

Chromosome	Type	Cytoband Start	CN State	Size (kbp)
1	Gain	q32.1	3.00	2,042
22	Gain	q11.23	3.00	1,331

Location of 1q32.1 duplication:
arr[GRCh38] 1q32.1(202,453,699_204,495,318)x3

Chromosome	Type	Cytoband Start	CN State	Size (kbp)
22	Gain	q11.23	3.00	1,315

Chromosome	Type	Cytoband Start	CN State	Size (kbp)
22	Gain	q11.23	3.00	1,317

Figure S2. Molecular karyotyping of CRISPR-Cas9 gene edited iPSC lines. Molecular karyotyping was performed with the Cytoscan HT-CMA 96 array using the KaryoStat+ method. The size of structural aberration that could be detected were > 1 Mb for chromosomal gains and > 1 Mb for chromosomal losses (the resolution depended on the location of the aberration in the chromosome. Due to a lower probe density on the telomere ends and centromeres, the resolution in those locations may have been closer to 5 Mb).

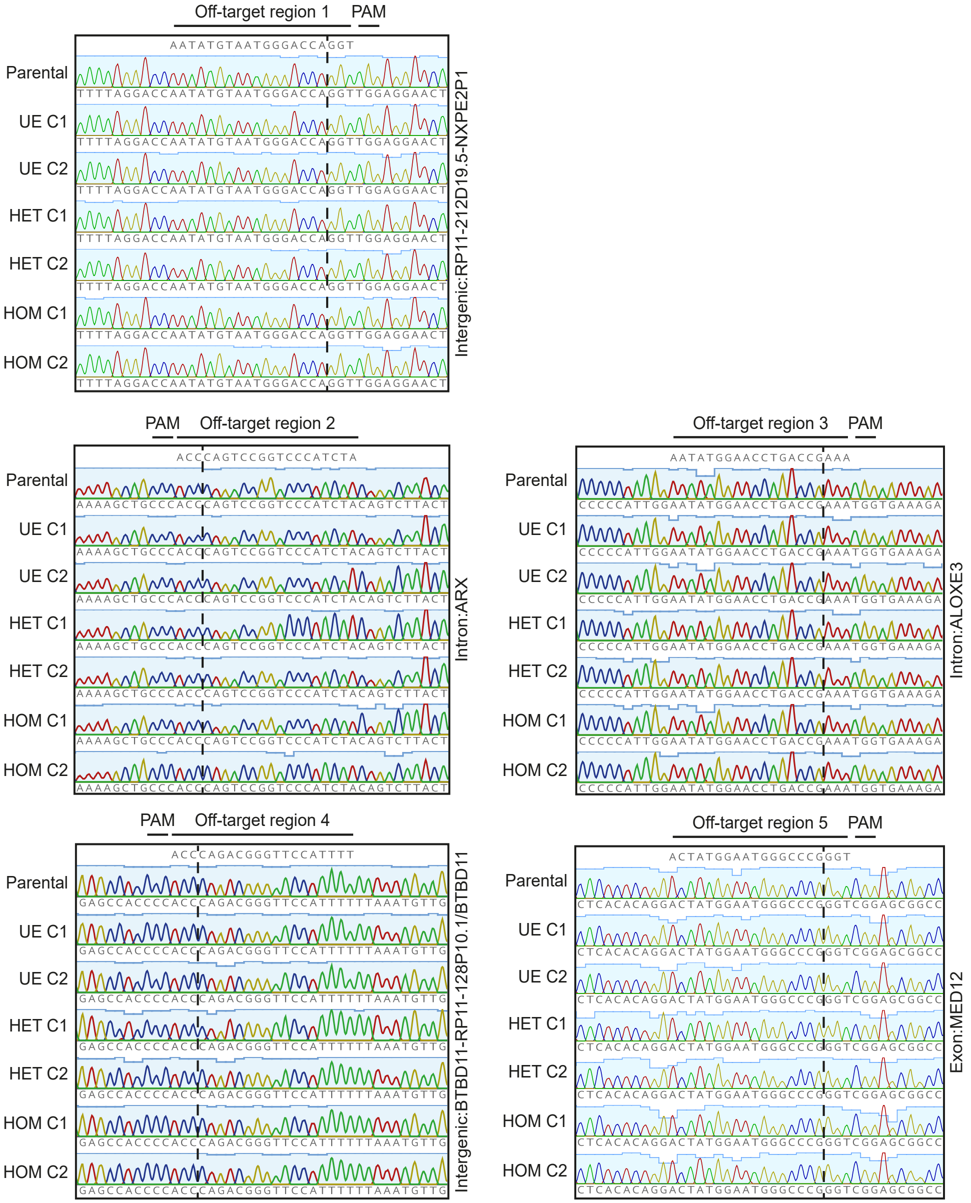


Figure S3. CRISPR-Cas9 off-target analysis. Five predicted off-targets for the *CHD3* CRISPR-Cas9 gene editing design were selected (Table S1) for Sanger sequencing and subsequent analysis. Graphs show the Sanger traces for the off-target regions for each cell line, and the dashed lines indicate the predicted cut sites.

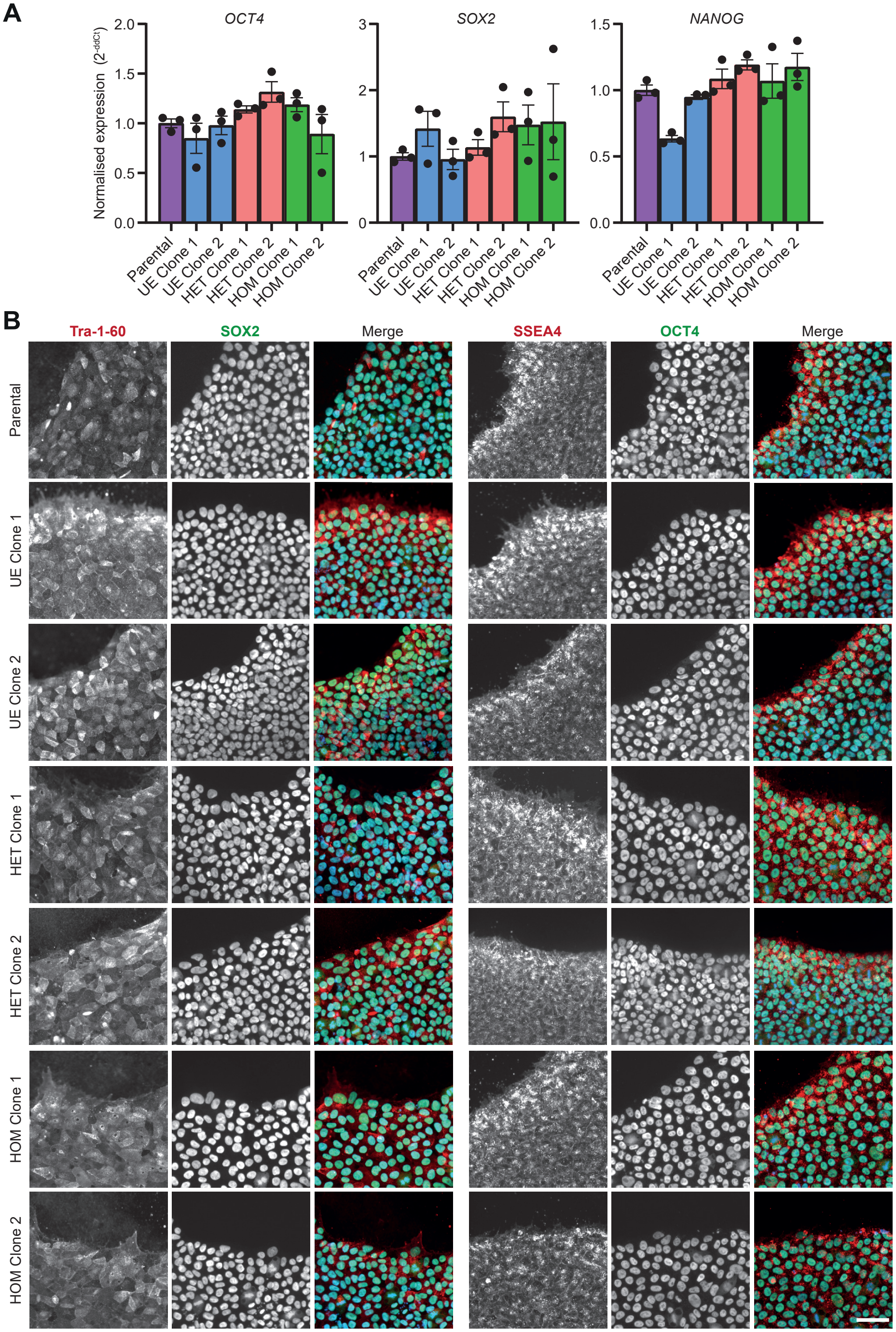


Figure S4. Characterisation of CRISPR-Cas9 gene-edited iPSC lines. A) Bar plot representing the mean ± S.E.M transcript levels of pluripotency markers *OCT4*, *SOX2* and *NANOG* in iPSCs, assayed using qPCR (*n* = 3). **B**) Immunohistochemistry micrographs of iPSC colonies for pluripotency markers TRA1-60 (left, red), SOX2 (left, green), SSEA4 (right, red) and OCT4 (right, green). Scale bar = 50 μm.

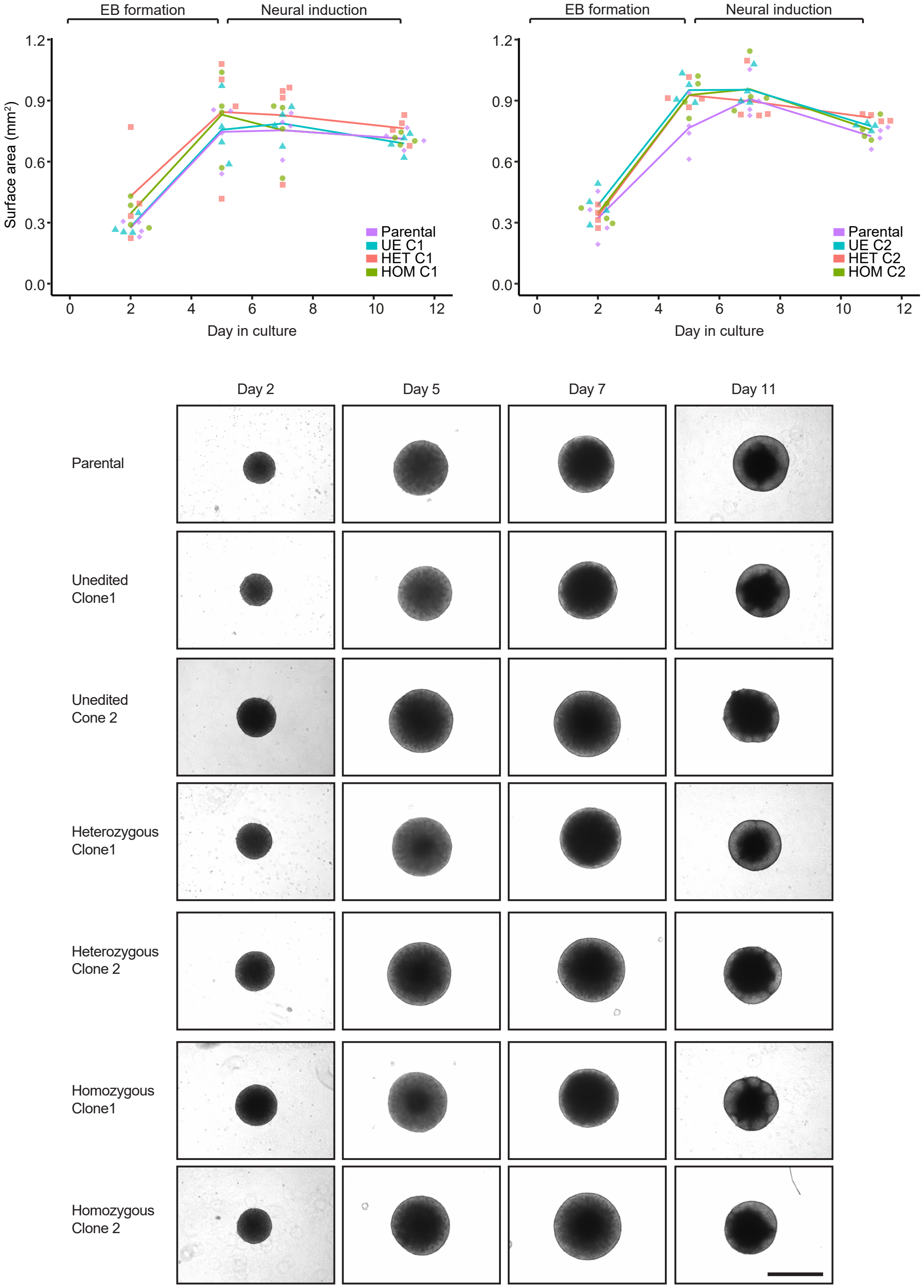


Figure S5. Growth of embryoid bodies. Top, a graph showing the surface area of embryoid bodies of each cell line at day 2, 5, 7 and 11 of the cerebral organoid protocol. The lines connect the mean surface area at each day. Bottom, representative bright-field images of day 2, 5, 7 and 11 cerebral organoids for each cell line. Scale bar = 1000 µm.

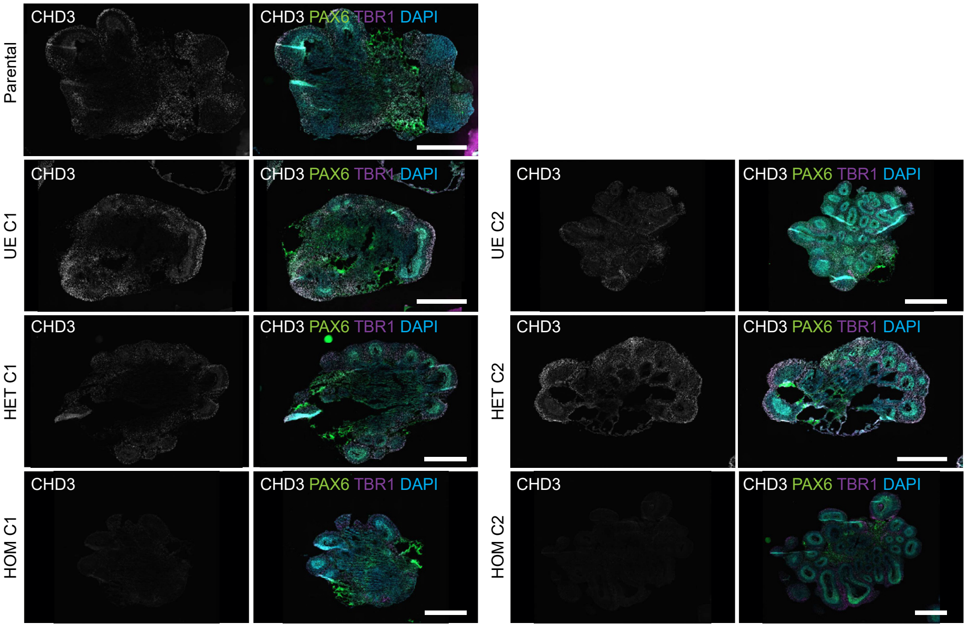


Figure S6. Organisation of cerebral organoids. Immunohistochemistry micrographs of whole day 57 cerebral organoid ventricles for radial glia cells (PAX6, green), post-mitotic neurons (TBR1, magenta) and CHD3-positive cells (CHD3, white). Scale bar = 1000 µm.

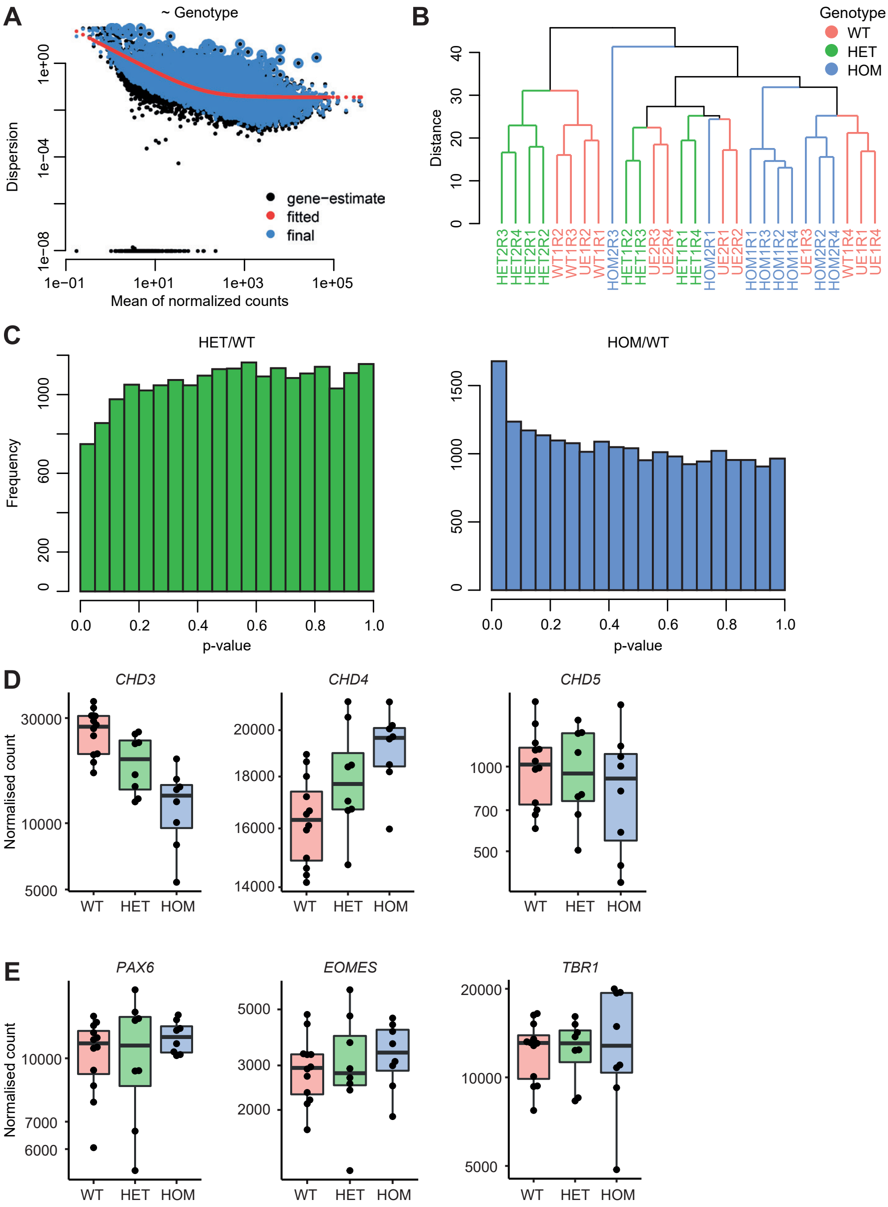


Figure S7. Bulk RNA-seq of day-50 cerebral organoids. A) Plot of dispersion estimates over the average expression strength for the ~genotype model. The red curve shows the overall trend of the dispersion-mean dependence, and the blue dots depict the final estimates after shrinkage based on the fit (in red). **B**) Dendogram showing the hierarchical clustering results based on variance stabilising transformation normalised counts. Samples with the wildtype genotype are shown in red, samples with a heterozygous *CHD3* knockout mutation in green, and homozygous CHD3 knockout samples in blue. **C**) Histogram distributing the p value (between 0 and 1) in twenty bins, and showing the number of differentially expressed genes with the corresponding p value for each bin. The left graph shows the data for the heterozygous CHD3 knockout versus wildtype comparison (green), while the right graph shows the results for the homozygous CHD3 knockout versus wildtype comparison (blue). **D**) Boxplots of normalised counts for *CHD3*, *CHD4* and *CHD5*. **E**) Boxplots of normalised counts for *PAX6*, *EOMES* and *TBR1*.

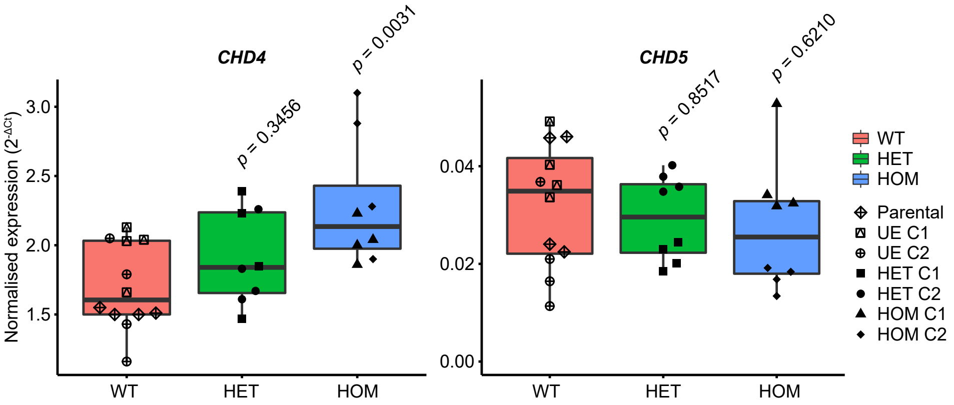


Figure S8. Expression of *CHD4* and *CHD5* in day-50 cerebral organoids. Boxplot representing the transcript levels of *CHD4* (left) and *CHD5* (right) in day-50 cerebral organoids, assayed using qPCR (*n* = 8-12, p values compared to wildtype condition (WT), one-way ANOVA and *post-hoc* Dunnett's test).

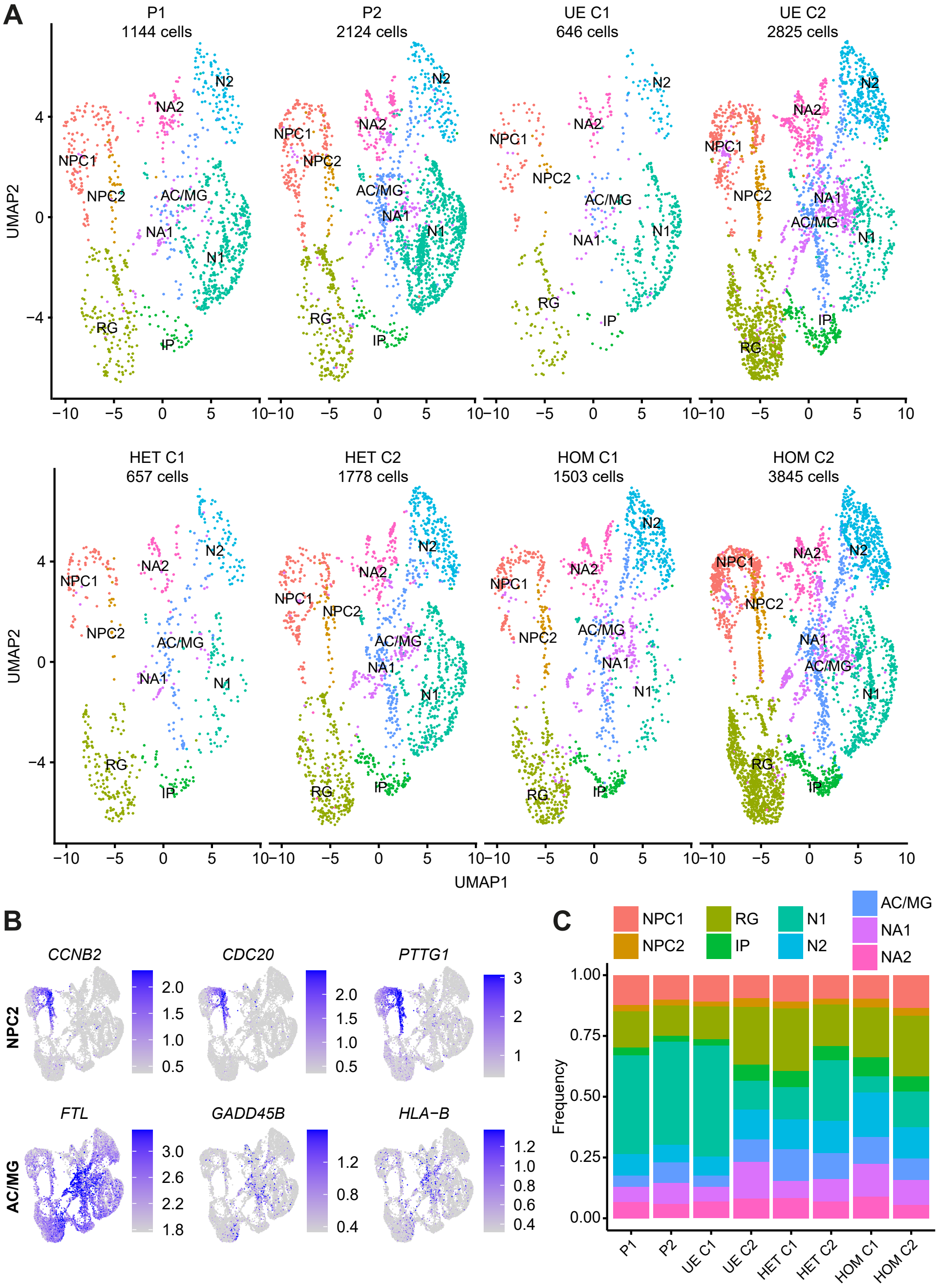


Figure S9. The effects of *CHD3* disruption on cell composition of cerebral organoids. A) Scatter plots of day-57 cerebral organoids for each sample after single-cell RNA-seq principal components analysis and uniform manifold approximation projection (UMAP) embedding with colouring based on the identified clusters. **B**) Expression of marker genes for cluster NPC2 and AC/MG. The gene-specific contrast levels were based on quantiles of non-zero expression (minimum = q10, maximum = q90). **C**) Relative cellular distributions for each sample.

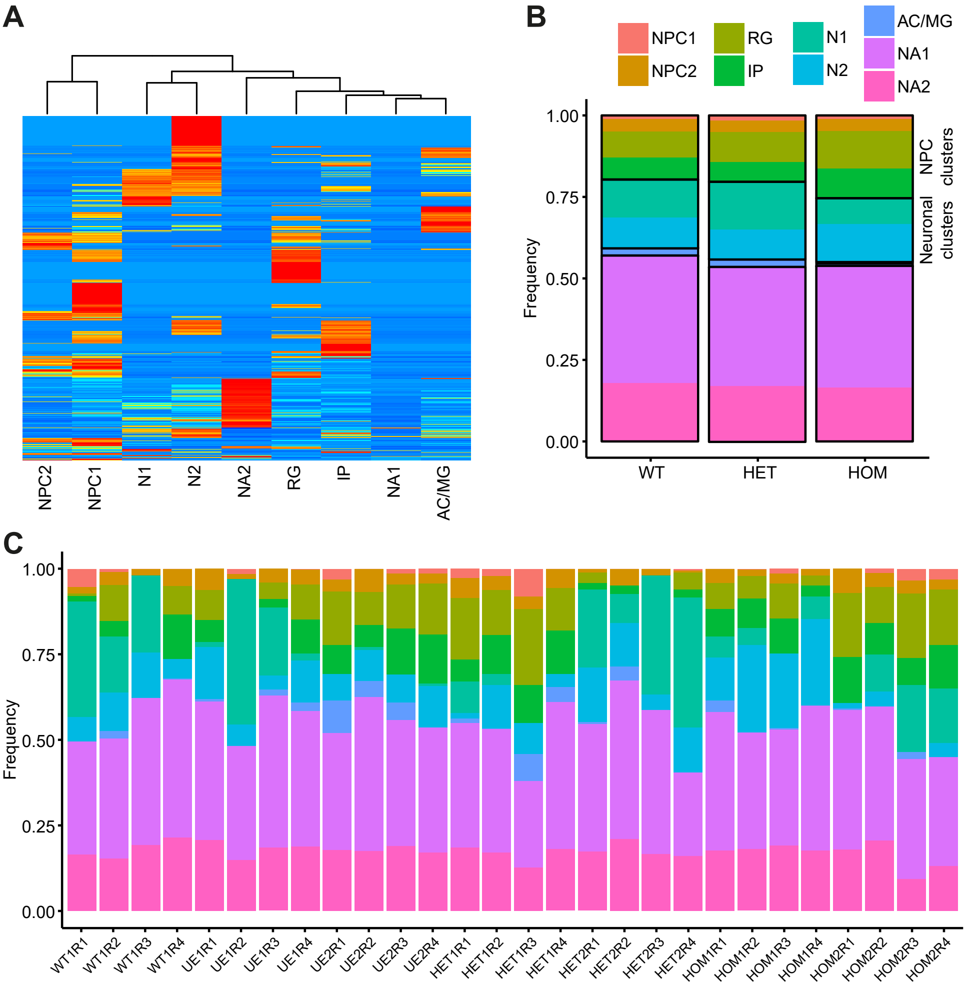


Figure S10. Cell type abundance estimation in bulk transcriptomics data using CIBERSORTx. A) A signature profile of 300-500 marker genes was created for each cluster identified in the single-cell RNAseq data set based on the expression data of the parental (P), UE C1 and UE C2 cell lines. The signature matrix is shown as a heatmap, horizontally presenting the different clusters, and vertically the expression of marker genes. **B**) Using the signature matrix, cell type abundance in the bulk transcriptomic samples was estimated using CIBERSORTx. Bars represent relative cellular distributions for each genotype. **C**) Relative cellular distributions for each bulk transcriptomic sample.

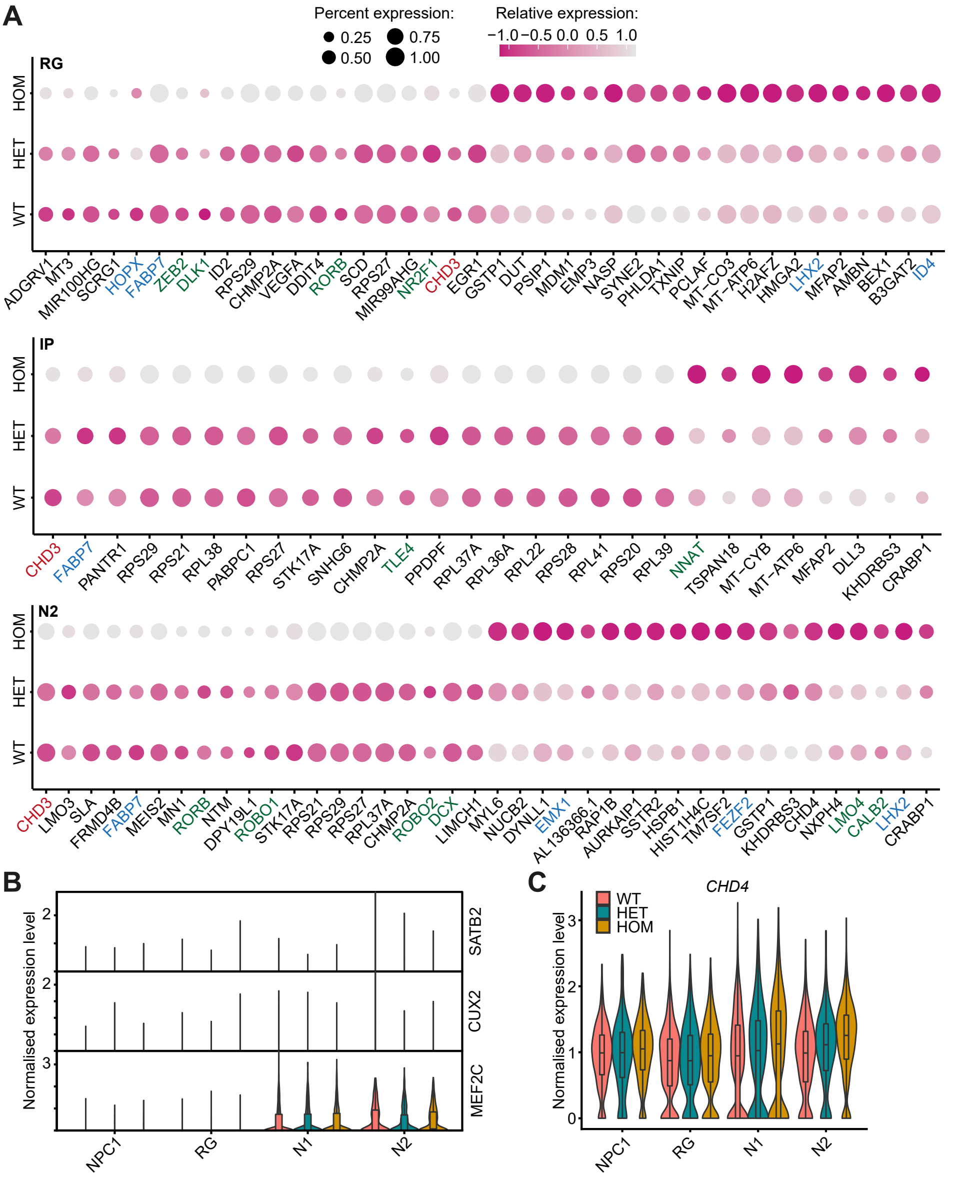


Figure S11. The effects of *CHD3* disruption on gene expression in NPCs and neurons in day-57 cerebral organoids. A) Dot plots showing the twenty strongest upregulated and the twenty most downregulated significant differentially expressed genes in the RG, IP and N2 clusters between homozygous *CHD3* knockout and wildtype organoids. For cluster IP, only 8 genes were found to be downregulated. *CHD3* is highlighted in red. Genes that positively regulate neuronal differentiation are depicted in green, while genes associated with NPC maintenance are shown in blue. **B**) Violin plots for three markers of upper layer cortical neurons (*SATB2*, *CUX2* and *MEF2C*) in day-57 cerebral organoids based on single cell RNA-seq data in the NPC1, RG, N1 and N2 clusters. **C**) Violin plots for *CHD4* in day-57 cerebral organoids based on single cell RNA-seq data in the NPC1, RG, N1 and N2 clusters.

Table S1. Predicted off-targets by CRISPOR 4.97 for the gRNA used to target the CHD3 gene. Genomic locations are based on hg19.

Predicted off-targets	Mismatch pattern	Chromosome	Start	End	Strand	Tested off-target
intron:IPPK	....**..........*.*.	chr9	95401680	95401702	+	-
intron:USP9X	*.**.......*........	chrX	40945767	40945789	-	-
intergenic:RP11-669M16.1-AC006296.1	...*......*......*..	chr4	14361036	14361058	-	-
intergenic:RP11-212D19.5-NXPE2P1	......*..**.....*...	chr11	114328247	114328269	+	Off-target 1
intron:ARX	*.*....*.......*....	chrX	25027102	25027124	+	Off-target 2
intron:ALOXE3	...........*.....***	chr17	8009937	8009959	+	Off-target 3
intergenic:LINC00963/RP11-492E3.51-RP11-492E3.2	...*....*..*........	chr9	132320170	132320192	-	-
intron:ZNF131	......*........*..**	chr5	43182693	43182715	-	-
intron:RAP1B	...*.......*......**	chr12	69014656	69014678	-	-
intergenic:BTBD11-RP11-128P10.1/BTBD11	..*........*.*.*....	chr12	107982587	107982609	-	Off-target 4
exon:MED12	.*.......**..*......	chrX	70356730	70356752	+	Off-target 5
intron:LIN54	.........*.*.*.....*	chr4	83892147	83892169	-	-
intergenic:RP11-973F15.2-AC016405.1	...*.*........*.*...	chr8	123755989	123756011	+	-
intergenic:FAM78B-RP11-375H19.2	....*.........*.**..	chr1	166045486	166045508	-	-
intergenic:PNPLA5-Z97055.1	.......*.**.........	chr22	44290344	44290366	-	-
intron:MECOM	.......*..*.*.....*.	chr3	168973006	168973028	-	-
intergenic:AL691479.1-PGBD5	...........*..*.**..	chr1	230455958	230455980	+	-
intergenic:RP3-497J21.1-RPS6KA2	...**.........*...*.	chr6	167146435	167146457	+	-
intron:SH3RF3	..*...*.......**....	chr2	110010848	110010870	-	-

Table S2. Differentially expressed genes of cerebral organoids with a heterozygous knockout variant (HET C1 and HET C2; top), or a homozygous knockout variant (HOM C1 and HOM C2; bottom) compared to organoids with a wildtype genotype (P, UE C1 and UE C2) at day 50. Genes were filtered for an adjusted *p*-value of < 0.1.

HET/WT

GeneID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
MEG3	1168.185	-9.69532	1.137236	-8.52534	1.52E-17	1.61E-13
SVIL-AS1	120.5159	-7.79735	0.579256	-13.461	2.65E-41	5.62E-37
ZNF208	17.83823	-6.63947	1.698824	-3.90827	9.30E-05	0.08954
ZNF257	8.604197	-6.21861	1.520274	-4.09045	4.31E-05	0.053667
MEG9	9.211912	-5.71193	1.139211	-5.01394	5.33E-07	0.00226
MEG8	4.92342	-5.58556	1.290359	-4.32869	1.50E-05	0.024451
ZNF662	6.322812	-4.13822	0.893568	-4.63112	3.64E-06	0.00929
PCDHGA7	162.9519	-3.19168	0.792377	-4.02798	5.63E-05	0.066232
BCO1	6.75151	-3.08325	0.77646	-3.9709	7.16E-05	0.075864
LINC00654	30.43506	-2.07289	0.331762	-6.24814	4.15E-10	2.93E-06
SPATC1L_1	18.5533	-1.59481	0.399106	-3.99596	6.44E-05	0.071862
LOC100996662	15.98727	-1.34741	0.313722	-4.29491	1.75E-05	0.026453
ENTPD2	33.48505	-1.04038	0.229248	-4.53825	5.67E-06	0.01202
TICAM1	36.51054	-0.82648	0.179115	-4.61422	3.95E-06	0.00929
CLDN6	108.0962	-0.75575	0.157448	-4.80001	1.59E-06	0.005604
GAS8	327.3522	-0.63899	0.163307	-3.91285	9.12E-05	0.08954
RAD51C	231.8914	-0.51355	0.092908	-5.5275	3.25E-08	0.000172
ADIPOR1	2455.191	0.409828	0.092081	4.450738	8.56E-06	0.016486
LOC101926887	106.0342	0.528655	0.125004	4.229105	2.35E-05	0.033146
LOC440173	143.3727	0.662235	0.151959	4.358	1.31E-05	0.023179
PLEKHM1	14.39399	0.980385	0.238369	4.112894	3.91E-05	0.05175
SLCO1A2	94.54709	1.333453	0.28086	4.747749	2.06E-06	0.006227
ATF6B	59.41439	1.814559	0.465986	3.894024	9.86E-05	0.09084

HOM/WT

GeneID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
MEG3	1168.185	-9.00159	1.124022	-8.00837	1.16E-15	1.21E-11
MEG8	4.92342	-5.22	1.290359	-4.04538	5.22E-05	0.016447
HIST1H3C	3.475971	-4.37285	0.965671	-4.5283	5.95E-06	0.002423
LOC101928335	2.672542	-4.14076	0.895218	-4.62542	3.74E-06	0.001766
HIST2H2BA	1.970124	-3.78814	1.017586	-3.72267	0.000197	0.044604
MEG9	9.211912	-3.74691	0.974263	-3.84589	0.00012	0.030838
HTR1A	3.412394	-3.5184	0.95736	-3.6751	0.000238	0.052636
LINC00664	25.85959	-3.38344	0.490334	-6.90028	5.19E-12	1.80E-08
FAM50B	109.8501	-3.1453	0.433575	-7.25434	4.04E-13	2.80E-09
ZNF662	6.322812	-2.91755	0.775295	-3.76315	0.000168	0.039621
FAM157B	6.814564	-2.80625	0.613514	-4.57406	4.78E-06	0.002071
KCNS1	8.728336	-2.36834	0.656204	-3.60915	0.000307	0.061578
CYSLTR2	61.96452	-2.3516	0.671512	-3.50195	0.000462	0.075055
GPR179_1	4.457343	-2.34124	0.638905	-3.66446	0.000248	0.053099
EFCAB13	25.68276	-2.31668	0.285755	-8.10722	5.18E-16	1.08E-11
OTOG	5.560105	-2.03903	0.584351	-3.48939	0.000484	0.077387
FCER2	11.89835	-1.93347	0.403431	-4.79258	1.65E-06	0.000877
LINC00654	30.43506	-1.88472	0.328057	-5.7451	9.19E-09	1.08E-05
LOC107985534	15.75538	-1.70811	0.415182	-4.11411	3.89E-05	0.01292
LINC02449	47.80376	-1.66862	0.419495	-3.9777	6.96E-05	0.020657
BCL6B	15.15493	-1.55253	0.379338	-4.09274	4.26E-05	0.013628
SPATC1L_1	18.5533	-1.5506	0.398221	-3.89381	9.87E-05	0.026289
PCK1	11.32895	-1.40235	0.399353	-3.51157	0.000445	0.075055
ENTPD2	33.48505	-1.3877	0.235624	-5.88947	3.87E-09	5.37E-06
PCDHB17P	36.47513	-1.38545	0.393776	-3.51837	0.000434	0.073957
TRPM2	12.46953	-1.37967	0.389715	-3.54019	0.0004	0.070411
FHIT	23.74252	-1.35506	0.279588	-4.84663	1.26E-06	0.000687
ZNRD1	37.37031	-1.25007	0.349633	-3.57537	0.00035	0.063968

COX6B2	173.7013	-1.23634	0.250665	-4.93225	8.13E-07	0.000469
QPCT	42.7355	-1.22566	0.28501	-4.3004	1.70E-05	0.006216
C5orf63	145.3805	-1.20749	0.194209	-6.21747	5.05E-10	8.75E-07
SCN4A	21.18855	-1.17678	0.302159	-3.89457	9.84E-05	0.026289
PCDHB18P	216.6827	-1.1752	0.327463	-3.5888	0.000332	0.063332
NECAB1	100.4376	-1.15134	0.230563	-4.9936	5.93E-07	0.000397
CPNE7	96.38558	-1.14906	0.200094	-5.74262	9.32E-09	1.08E-05
PCDHGA1	84.59397	-1.10613	0.215804	-5.1256	2.97E-07	0.000213
CHD3	20379.17	-1.07386	0.192962	-5.56511	2.62E-08	2.59E-05
GAS8	327.3522	-1.05273	0.164251	-6.40924	1.46E-10	3.04E-07
PTP4A3_1	191.3797	-1.04825	0.285773	-3.66812	0.000244	0.053099
TICAM1	36.51054	-1.03175	0.18307	-5.63583	1.74E-08	1.91E-05
PNMA5	45.71488	-1.00933	0.183822	-5.49081	4.00E-08	3.78E-05
F7	39.33339	-0.9848	0.256077	-3.8457	0.00012	0.030838
FAM83H	23.22521	-0.95534	0.250272	-3.81721	0.000135	0.034203
TRIM7	77.91976	-0.90387	0.237547	-3.80503	0.000142	0.035498
HSPA2	65.90027	-0.8603	0.186912	-4.60269	4.17E-06	0.001911
FIBCD1	92.48586	-0.85321	0.251484	-3.39271	0.000692	0.098499
INPP5F	5410.537	-0.80965	0.154264	-5.24847	1.53E-07	0.000114
DPY19L2P4	72.37179	-0.78696	0.176415	-4.46082	8.16E-06	0.003142
PLD6	130.8396	-0.71368	0.155157	-4.5997	4.23E-06	0.001911
TRIM47	111.9329	-0.71335	0.206839	-3.4488	0.000563	0.087977
NSMCE1	543.3166	-0.71012	0.149197	-4.75962	1.94E-06	0.000983
LOC105376063	24.48222	-0.69585	0.203682	-3.41637	0.000635	0.093262
STAC2	342.3298	-0.69321	0.144893	-4.7843	1.72E-06	0.000891
ELN	801.6984	-0.64956	0.166436	-3.90279	9.51E-05	0.026289
HPDL	52.61472	-0.64388	0.179854	-3.57999	0.000344	0.063752
LOC100129484	37.06847	-0.63498	0.181083	-3.50657	0.000454	0.075055
DPF3	242.2411	-0.61803	0.124735	-4.95474	7.24E-07	0.000456
DPP7	1334.508	-0.61368	0.139511	-4.39881	1.09E-05	0.004112
RGS11	319.2116	-0.59527	0.148653	-4.00441	6.22E-05	0.019283
CLDN6	108.0962	-0.58608	0.156353	-3.74844	0.000178	0.041546
FBLN5	165.5408	-0.57757	0.157525	-3.66654	0.000246	0.053099
ZFYVE28	85.00977	-0.56923	0.124934	-4.55624	5.21E-06	0.002164
AGAP2-AS1	123.395	-0.55078	0.154141	-3.57319	0.000353	0.063968
PCDHB5	2160.793	-0.54817	0.083405	-6.57243	4.95E-11	1.14E-07
ACTG1P20	38.30898	-0.51836	0.151514	-3.42117	0.000624	0.092548
SOX15	140.8765	-0.51686	0.152291	-3.39389	0.000689	0.098499
SMTNL2	104.0547	-0.49601	0.136715	-3.62806	0.000286	0.058752
FBXW4	301.0674	-0.47306	0.11518	-4.10713	4.01E-05	0.013007
RAD51C	231.8914	-0.46121	0.092675	-4.97663	6.47E-07	0.00042
EPS8L2	156.1503	-0.45486	0.133187	-3.41522	0.000637	0.093262
LRRC75B	1011.336	-0.42105	0.100172	-4.2033	2.63E-05	0.009265
SOCS2	880.0535	-0.41592	0.071404	-5.82492	5.71E-09	7.42E-06
YBX2	137.5218	-0.41438	0.1203	-3.44451	0.000572	0.088059
TNS3	877.75	-0.38181	0.092371	-4.13342	3.57E-05	0.012175
PPP1R9A	1314.14	-0.37127	0.09308	-3.98871	6.64E-05	0.020301
DBNDD1	1586.998	-0.33557	0.093599	-3.58519	0.000337	0.063523
KLHL21	601.754	-0.21616	0.06013	-3.59494	0.000324	0.062429
ORMDL3	1570.543	0.142462	0.03955	3.602116	0.000316	0.061877
ZFAND3	4318.886	0.188251	0.051591	3.648886	0.000263	0.055847
RNF11	4020.647	0.19065	0.056288	3.387014	0.000707	0.098863
ZNF275	1205.528	0.193699	0.055296	3.502933	0.00046	0.075055
UBE2J1	2781.322	0.210334	0.061379	3.426789	0.000611	0.091968
SAP30L	1453.355	0.219185	0.0533	4.112316	3.92E-05	0.01292
SGPL1	2361.724	0.227251	0.066807	3.401614	0.00067	0.097345
GALNT1	2303	0.243513	0.065421	3.722225	0.000197	0.044604
EDEM1	708.9171	0.24876	0.068529	3.630013	0.000283	0.058752
IGF1R	4502.647	0.25064	0.073864	3.393281	0.000691	0.098499
FHL1	8418.49	0.253249	0.064291	3.939113	8.18E-05	0.023604
TRIM52-AS1	473.7506	0.256959	0.065832	3.903257	9.49E-05	0.026289

MAPK4	1622.868	0.268064	0.070737	3.789587	0.000151	0.036776
TMEM200C	1016.402	0.268192	0.068596	3.909719	9.24E-05	0.026289
KLHL5	1817.581	0.272094	0.068352	3.980801	6.87E-05	0.020657
BRPF3	3351.932	0.275438	0.076868	3.583271	0.000339	0.063523
RXRA	1617.973	0.277627	0.08199	3.38612	0.000709	0.098863
MAPKAPK2	2280.786	0.294842	0.04692	6.283946	3.30E-10	6.24E-07
FAM210B	2068.081	0.297733	0.056464	5.272944	1.34E-07	0.000103
ZNF559	583.7879	0.302315	0.079984	3.779678	0.000157	0.037507
UBE2E2	765.2836	0.317528	0.09247	3.433837	0.000595	0.090291
LBH	2926.918	0.323712	0.058063	5.575156	2.47E-08	2.57E-05
SLC6A8	4642.184	0.33779	0.098373	3.433756	0.000595	0.090291
PTGIS	828.5361	0.353309	0.081536	4.333147	1.47E-05	0.005454
GALNT16	1816.088	0.360141	0.080195	4.490813	7.10E-06	0.002782
AR	287.4325	0.367082	0.099894	3.674725	0.000238	0.052636
JUP	4200.056	0.368216	0.061715	5.966402	2.43E-09	3.60E-06
KIAA1324L	1465.898	0.389122	0.093518	4.160946	3.17E-05	0.010976
BEND7	684.6849	0.391488	0.086635	4.518838	6.22E-06	0.002485
PTPN21	996.6778	0.400723	0.073909	5.421877	5.90E-08	5.33E-05
PATZ1	3628.01	0.417782	0.057913	7.213925	5.44E-13	2.82E-09
PCDHGA8	95.3368	0.426265	0.123677	3.446605	0.000568	0.088032
RCBTB2	610.2943	0.459892	0.085343	5.388768	7.09E-08	5.90E-05
LINC01351	118.549	0.460383	0.097298	4.731658	2.23E-06	0.001102
WIPF1	678.3046	0.489973	0.14461	3.388244	0.000703	0.098863
DUSP22	1118.492	0.500358	0.10133	4.937883	7.90E-07	0.000469
MAML1	446.0414	0.501385	0.140986	3.55628	0.000376	0.066806
TRPC3	135.7156	0.541221	0.115625	4.68083	2.86E-06	0.001381
PACSIN3	262.1253	0.545137	0.143668	3.794412	0.000148	0.036611
FGFR3	5362.498	0.573793	0.163862	3.501688	0.000462	0.075055
SLC45A3	239.1738	0.578094	0.164269	3.519184	0.000433	0.073957
BRINP3	297.1349	0.593153	0.15661	3.787451	0.000152	0.036776
ITPRIPL1	198.3513	0.618738	0.127539	4.85137	1.23E-06	0.000687
CHRD	204.0617	0.639049	0.129518	4.934049	8.05E-07	0.000469
NR6A1	220.7265	0.641656	0.14057	4.564661	5.00E-06	0.002122
TMEM35B	66.09425	0.679485	0.148258	4.583119	4.58E-06	0.002025
TPM2	425.1239	0.688518	0.191467	3.596005	0.000323	0.062429
HRK	124.0961	0.692093	0.191889	3.606729	0.00031	0.061578
LRRC37A6P	23.43249	0.695944	0.178609	3.896463	9.76E-05	0.026289
XKRX	28.66275	0.729399	0.210429	3.466247	0.000528	0.083085
ADAMTS18	69.17383	0.904377	0.241829	3.739734	0.000184	0.042533
HMGN3-AS1	51.55712	0.907853	0.136569	6.647555	2.98E-11	7.74E-08
LOC440173	143.3727	0.936284	0.151045	6.1987	5.69E-10	9.10E-07
LOC105377771	80.30896	0.975469	0.24703	3.948795	7.85E-05	0.022988
SLCO1A2	94.54709	1.019608	0.281681	3.619722	0.000295	0.060083
ZFHX4-AS1	218.3028	1.030197	0.190808	5.399132	6.70E-08	5.80E-05
LOC112268460	27.65933	1.059944	0.290829	3.64456	0.000268	0.056221
LINC01508	15.73985	1.067781	0.30189	3.53699	0.000405	0.070672
ABCA13	13.77567	1.137923	0.324958	3.501753	0.000462	0.075055
SPINK5	199.0406	1.168248	0.230402	5.07048	3.97E-07	0.000275
LOC100507351	84.63368	1.315137	0.364724	3.605837	0.000311	0.061578
LOC105369205	460.9398	1.617418	0.462218	3.49925	0.000467	0.075157
LOC339260	66.60968	1.683052	0.484766	3.471886	0.000517	0.08198
VSX2	40.16918	1.862699	0.440124	4.232217	2.31E-05	0.00829
AKT3_1	28.89636	1.952279	0.507143	3.849561	0.000118	0.030838
LOC107985079	3.016008	2.405322	0.673347	3.57219	0.000354	0.063968
SLC6A10P	2.57835	2.73758	0.775975	3.527924	0.000419	0.072528
AACSP1	66.98783	2.938052	0.556402	5.280446	1.29E-07	0.000103
ZNF728	3.876981	4.826866	1.40958	3.424329	0.000616	0.092138
ZNF248	172.0153	5.04503	0.711977	7.08594	1.38E-12	5.74E-09
ANKRD30B	2.847497	5.075608	1.422918	3.567042	0.000361	0.064675
ZXDA	46.46016	5.097031	0.754805	6.752778	1.45E-11	4.31E-08

Table S3. scCATCH annotation of UMAP identified clusters of the scRNAseq data. Marker genes were identified per cluster, selected from brain-related tissues from the CellMatch database. Afterwards, these cluster markers were mapped to cell type-related markers (also selected from brain-related tissues), and a cell type score was calculated. Based on the annotations from scCATCH, the clusters were relabeled as indicated in the top table.

Cluster	Relabeling	Abbreviation
0	Neuronal cluster 1	N1
1	Radial glial cells	RG
2	Neuronal cluster 2	N2
3	Neural progenitor cells 1	NPC1
4	Not annotated	NA1
5	Astrocytes/microglia cells	AC/MG
6	Not annotated	NA2
7	Intermediate progenitor cells	IP
8	Neural progenitor cells 2	NPC2

1184 selected brain-related markers
(tissues: Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Mapped to 40 cell types (from Brain, Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Cluster	Cluster markers	Cell type	Cell type score	Cell type-related markers	PMID
0	DCX, DLX2, DLX5, EGR3, GAD1, GAD2, INA, MEIS2, NRXN3, SOX2-OT, ZFHX3	Interneuron	0.66	DLX2, DLX5, GAD1, GAD2, INA, NRXN3, SOX2-OT	29539641
1	CLU, CREB5, DOK5, FABP7, GLI3, HES1, ID4, IFITM3, MIR9-1HG, PEA15, PLPP3, PON2, PTN, QKI, RCN1, RPS27L, SCD, SFRP1, SLC1A3, TCIM, TTYH1, ZFP36L1	Astrocyte	0.69	MIR9-1HG, CLU, DOK5, FABP7, HES1, ID4, IFITM3, PEA15, PON2, PTN, QKI, RCN1, SCD, SLC1A3, TTYH1, ZFP36L1, TCIM, PLPP3	29539641

2	BCHE, BCL11A, BEX1, BEX3, CSRP2, FXYD6, GAP43, GPM6A, GSTA4, LMO3, LMO4, MYT1L, NEUROD2, NEUROD6, NFIB, NRXN1, NSG1, OLFM2, PCSK1N, PPP2R2B, RTN1, SCD5, SLA, SLC1A2, SNAP25, SYT1, TBR1, TSPAN13, TUBA1A, UCHL1, ZEB2	Neuron	0.76	TBR1, CSRP2, NEUROD2, NEUROD6, PPP2R2B, SLA	12513943, 29539641
3	ASPM, ATAD2, AURKB, BTG3, BUB3, C21orf58, CCNA2, CDCA3, CDCA4, CDCA5, CDK1, CENPK, CENPM, CENPU, CENPW, CKAP2, CKAP2L, CKS1B, CKS2, CLSPN, DHFR, FBXO5, GMNN, GTSE1, H1-2, H2AX, H4C3, HJURP, IER2, ITGB3BP, KIF11, KIF15, KIF22, KIF23, KPNA2, MAD2L1, MELK, MIS18BP1, MKI67, NDC80, NUSAP1, ORC6, PBK, PCLAF, PCNA, PRC1, RPA3, RRM1, RRM2, SGO1, SGO2, SMC2, SMC4, SPC25, TMPO, TOP2A, TUBA1B, TUBB, TUBB4B, UBE2C, UBE2T, ZWINT	Neural Progenitor Cell	0.7	ASPM, ATAD2, AURKB, BTG3, BUB3, C21orf58, CCNA2, CDCA3, CDCA4, CDCA5, CDK1, CENPK, CENPM, CENPU, CENPW, CKAP2, CKAP2L, CKS1B, CKS2, CLSPN, DHFR, FBXO5, GMNN, GTSE1, H2AX, H4C3, HJURP, ITGB3BP, PCLAF, KIF11, KIF15, KIF22, KIF23, KPNA2, MAD2L1, MELK, MIS18BP1, MKI67, NDC80, NUSAP1, ORC6, PBK, PCNA, PRC1, RPA3, RRM1, RRM2, SGO1, SGO2, SMC2, SMC4, SPC25, TMPO, TOP2A, TUBA1B, TUBB4B, UBE2C, UBE2T, ZWINT	29539641
5	FTL, GADD45B, HLA-B, SLC3A2	Microglial Cell	0.61	FTL, GADD45B, HLA-B	29539641
7	ADGRG1, EMX1, ENC1, EOMES, KLHDC8A,	Astrocyte	0.63	ADGRG1, KLHDC8A, SOX11, TAGLN3	29539641

	PHLDA1, SOX11, TAGLN3, TCF4				
8	CCNB2, CDC20, DCXR, HES5, PTTG1	Neural Progenitor Cell	0.61	CCNB2, CDC20, PTTG1	29539641

1184 selected brain-related markers
(tissues: Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Mapped to 10 cell types (from Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Cluster	Cluster markers	Cell type	Cell type score	Cell type-related markers	PMID
0	DCX, DLX2, DLX5, EGR3, GAD1, GAD2, INA, MEIS2, NRXN3, SOX2-OT, ZFHX3	Neuron, Progenitor Cell	0.71, 0.71	DCX, SOX2-OT, DLX2, DLX5	21275797, 29539641, 26060301
1	CLU, CREB5, DOK5, FABP7, GLI3, HES1, ID4, IFITM3, MIR9-1HG, PEA15, PLPP3, PON2, PTN, QKI, RCN1, RPS27L, SCD, SFRP1, SLC1A3, TCIM, TTYH1, ZFP36L1	Astrocyte	0.85	MIR9-1HG, CLU, DOK5, ID4, PEA15, PON2, SLC1A3, ZFP36L1, FABP7, HES1, IFITM3, PTN, QKI, RCN1, SCD, TTYH1, GLI3, TCIM, PLPP3	27806376, 29539641, 26060301
2	BCHE, BCL11A, BEX1, BEX3, CSRP2, FXYD6, GAP43, GPM6A, GSTA4, LMO3, LMO4, MYT1L, NEUROD2, NEUROD6, NFIB, NRXN1, NSG1, OLFM2, PCSK1N, PPP2R2B, RTN1, SCD5, SLA, SLC1A2, SNAP25, SYT1, TBR1, TSPAN13, TUBA1A, UCHL1, ZEB2	Astrocyte	0.78	SLC1A2, BCHE, BCL11A, BEX1, CSRP2, NEUROD6, NSG1, OLFM2, RTN1, SYT1, UCHL1	27806376, 29539641
3	ASPM, ATAD2, AURKB, BTG3, BUB3, C21orf58, CCNA2, CDCA3, CDCA4, CDCA5, CDK1, CENPK, CENPM, CENPU, CENPW,	Neural Progenitor Cell	0.7	ASPM, ATAD2, AURKB, BTG3, BUB3, C21orf58, CCNA2, CDCA3, CDCA4, CDCA5, CDK1, CENPK, CENPM, CENPU, CENPW,	29539641

	CKAP2, CKAP2L, CKS1B, CKS2, CLSPN, DHFR, FBXO5, GMNN, GTSE1, H1-2, H2AX, H4C3, HJURP, IER2, ITGB3BP, KIF11, KIF15, KIF22, KIF23, KPNA2, MAD2L1, MELK, MIS18BP1, MKI67, NDC80, NUSAP1, ORC6, PBK, PCLAF, PCNA, PRC1, RPA3, RRM1, RRM2, SGO1, SGO2, SMC2, SMC4, SPC25, TMPO, TOP2A, TUBA1B, TUBB, TUBB4B, UBE2C, UBE2T, ZWINT			CKAP2, CKAP2L, CKS1B, CKS2, CLSPN, DHFR, FBXO5, GMNN, GTSE1, H2AX, H4C3, HJURP, ITGB3BP, PCLAF, KIF11, KIF15, KIF22, KIF23, KPNA2, MAD2L1, MELK, MIS18BP1, MKI67, NDC80, NUSAP1, ORC6, PBK, PCNA, PRC1, RPA3, RRM1, RRM2, SGO1, SGO2, SMC2, SMC4, SPC25, TMPO, TOP2A, TUBA1B, TUBB4B, UBE2C, UBE2T, ZWINT	
5	FTL, GADD45B, HLA-B, SLC3A2	Astrocyte	0.71	GADD45B, HLA-B, SLC3A2	27806376, 29539641
7	ADGRG1, EMX1, ENC1, EOMES, KLHDC8A, PHLDA1, SOX11, TAGLN3, TCF4	Neural Progenitor Cell	0.73	ADGRG1, EMX1, EOMES	19525879, 29539641
8	CCNB2, CDC20, DCXR, HES5, PTTG1	Neural Progenitor Cell	0.61	CCNB2, CDC20, PTTG1	29539641

1877 selected brain-related markers
(tissues: Brain, Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Mapped to 40 cell types (from Brain, Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Cluster	Cluster markers	Cell type	Cell type score	Cell type-related markers	PMID
0	CXXC4, DCX, DLX2, DLX5, DLX6, EGR3, GAD1, GAD2, INA, LHFPL6, MEIS2, NRXN3, SOX2-OT, ZFHX3	Neuron	0.73	SOX2-OT, DLX2, DLX5, DLX6	29539641, 26060301
1	CLU, CREB5, DOK5, FABP7, FGFBP3, GLI3,	Astrocyte	0.85	MIR9-1HG, CLU, DOK5, ID4, PEA15, PON2,	27806376, 29539641, 26060301

	HES1, ID4, IFITM3, MIR9-1HG, NME2, PEA15, PLPP3, PON2, PTN, QKI, RCN1, RPS2, RPS27L, SCD, SFRP1, SLC1A3, TCIM, TTYH1, ZFP36L1			SLC1A3, ZFP36L1, FABP7, HES1, IFITM3, PTN, QKI, RCN1, SCD, TTYH1, GLI3, TCIM, PLPP3	
2	BCHE, BCL11A, BCL11B, BEX1, BEX3, CSRP2, CYRIA, FEZF2, FXYD6, FXYD7, GAP43, GPM6A, GRIA1, GRIA2, GRIA3, GSTA4, KIT, LMO3, LMO4, MAP2, MYT1L, NEUROD2, NEUROD6, NFIB, NFIL3, NRXN1, NSG1, OLFM2, PCSK1N, PHACTR3, PPP2R2B, RAB2A, RTN1, RTN4, SCD5, SLA, SLC1A2, SNAP25, SYP, SYT1, TBR1, TSPAN13, TUBA1A, TUBB4A, UCHL1, ZEB2	Neuron	0.87	TBR1, MAP2, SYP, CSRP2, NEUROD2, NEUROD6, PPP2R2B, SLA, GRIA3	12513943, 24142904, 27587997, 29539641, 26060301
3	ANLN, ASPM, ATAD2, AURKB, BTG3, BUB3, C21orf58, CCNA2, CDCA3, CDCA4, CDCA5, CDK1, CENPK, CENPM, CENPU, CENPW, CKAP2, CKAP2L, CKS1B, CKS2, CLSPN, DHFR, FBXO5, GMNN, GTSE1, H1-2, H2AX, H4C3, HJURP, IER2, ITGB3BP, KIF11, KIF15, KIF22, KIF23, KPNA2, MAD2L1, MELK, MIS18BP1, MKI67, NDC80, NUSAP1, ORC6, PBK,	Astrocyte	0.71	ANLN, IER2, TUBB	27806376, 29539641

	PCLAF, PCNA, PRC1, RPA3, RRM1, RRM2, SGO1, SGO2, SMC2, SMC4, SPC25, TMPO, TOP2A, TUBA1B, TUBB, TUBB4B, UBE2C, UBE2T, ZWINT				
5	ENO2, FTL, GADD45B, HLA-B, SLC3A2	Astrocyte	0.71	GADD45B, HLA-B, SLC3A2	27806376, 29539641
7	ADGRG1, DLL3, EMX1, ENC1, EOMES, EZR, KLHDC8A, MYT1, PHLDA1, RGMB, SOX11, TAGLN3, TCF4, TMSB4X	Astrocyte	0.75	EZR, ADGRG1, KLHDC8A, SOX11, TAGLN3	27806376, 29539641
8	CCNB2, CDC20, DCXR, HES5, NMU, PTTG1	Neural Progenitor Cell	0.61	CCNB2, CDC20, PTTG1	29539641

1877 selected brain-related markers
(tissues: Brain, Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Mapped to 10 cell types (from Dorsolateral prefrontal cortex, Embryonic brain, Embryonic prefrontal cortex, Fetal brain)

Cluster	Cluster markers	Cell type	Cell type score	Cell type-related markers	PMID
0	CXXC4, DCX, DLX2, DLX5, DLX6, EGR3, GAD1, GAD2, INA, LHFPL6, MEIS2, NRXN3, SOX2-OT, ZFHX3	Interneuron	0.66	DLX2, DLX5, GAD1, GAD2, INA, NRXN3, SOX2-OT	29539641
1	CLU, CREB5, DOK5, FABP7, FGFBP3, GLI3, HES1, ID4, IFITM3, MIR9-1HG, NME2, PEA15, PLPP3, PON2, PTN, QKI, RCN1, RPS2, RPS27L, SCD, SFRP1, SLC1A3, TCIM, TTYH1, ZFP36L1	Astrocyte	0.69	MIR9-1HG, CLU, DOK5, FABP7, HES1, ID4, IFITM3, PEA15, PON2, PTN, QKI, RCN1, SCD, SLC1A3, TTYH1, ZFP36L1, TCIM, PLPP3	29539641

2	BCHE, BCL11A, BCL11B, BEX1, BEX3, CSRP2, CYRIA, FEZF2, FXYD6, FXYD7, GAP43, GPM6A, GRIA1, GRIA2, GRIA3, GSTA4, KIT, LMO3, LMO4, MAP2, MYT1L, NEUROD2, NEUROD6, NFIB, NFIL3, NRXN1, NSG1, OLFM2, PCSK1N, PHACTR3, PPP2R2B, RAB2A, RTN1, RTN4, SCD5, SLA, SLC1A2, SNAP25, SYP, SYT1, TBR1, TSPAN13, TUBA1A, TUBB4A, UCHL1, ZEB2	Neuron	0.76	TBR1, CSRP2, NEUROD2, NEUROD6, PPP2R2B, SLA	12513943, 29539641
3	ANLN, ASPM, ATAD2, AURKB, BTG3, BUB3, C21orf58, CCNA2, CDCA3, CDCA4, CDCA5, CDK1, CENPK, CENPM, CENPU, CENPW, CKAP2, CKAP2L, CKS1B, CKS2, CLSPN, DHFR, FBXO5, GMNN, GTSE1, H1-2, H2AX, H4C3, HJURP, IER2, ITGB3BP, KIF11, KIF15, KIF22, KIF23, KPNA2, MAD2L1, MELK, MIS18BP1, MKI67, NDC80, NUSAP1, ORC6, PBK, PCLAF, PCNA, PRC1, RPA3, RRM1, RRM2, SGO1, SGO2, SMC2, SMC4, SPC25, TMPO, TOP2A, TUBA1B, TUBB, TUBB4B, UBE2C, UBE2T, ZWINT	Neural Progenitor Cell	0.7	ASPM, ATAD2, AURKB, BTG3, BUB3, C21orf58, CCNA2, CDCA3, CDCA4, CDCA5, CDK1, CENPK, CENPM, CENPU, CENPW, CKAP2, CKAP2L, CKS1B, CKS2, CLSPN, DHFR, FBXO5, GMNN, GTSE1, H2AX, H4C3, HJURP, ITGB3BP, PCLAF, KIF11, KIF15, KIF22, KIF23, KPNA2, MAD2L1, MELK, MIS18BP1, MKI67, NDC80, NUSAP1, ORC6, PBK, PCNA, PRC1, RPA3, RRM1, RRM2, SGO1, SGO2, SMC2, SMC4, SPC25, TMPO, TOP2A, TUBA1B, TUBB4B, UBE2C, UBE2T, ZWINT	29539641

5	ENO2, FTL, GADD45B, HLA-B, SLC3A2	Microglial Cell	0.61	FTL, GADD45B, HLA-B	29539641
7	ADGRG1, DLL3, EMX1, ENC1, EOMES, EZR, KLHDC8A, MYT1, PHLDA1, RGMB, SOX11, TAGLN3, TCF4, TMSB4X	Astrocyte	0.63	ADGRG1, KLHDC8A, SOX11, TAGLN3	29539641
8	CCNB2, CDC20, DCXR, HES5, NMU, PTTG1	Neural Progenitor Cell	0.61	CCNB2, CDC20, PTTG1	29539641

Table S4. Differentially expressed genes in homozygous cells (HOM1 and HOM2) compared to wildtype (P1, P2, UE1 and UE2) cells in individual clusters identified in the UMAP of the scRNAseq data (NPC1, RG, IP, N1 and N2).

NPC1	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
STMN2	1.58E-12	-1.10499	0.266	0.415	3.74E-08
SHTN1	3.02E-37	-1.08779	0.223	0.54	7.17E-33
DLX6-AS1	1.18E-24	-1.07836	0.122	0.357	2.81E-20
GAD2	2.90E-32	-1.06203	0.123	0.409	6.89E-28
DLX2	6.65E-32	-0.93247	0.198	0.496	1.58E-27
DLX1	1.29E-28	-0.86744	0.183	0.451	3.06E-24
ZEB2	6.29E-22	-0.82234	0.526	0.671	1.50E-17
DCX	3.46E-20	-0.79596	0.483	0.647	8.22E-16
SOX4	1.97E-11	-0.789	0.997	0.996	4.68E-07
RND3	6.30E-16	-0.75996	0.8	0.87	1.50E-11
DLX5	3.19E-21	-0.69037	0.114	0.331	7.58E-17
NNAT	6.84E-15	-0.64543	0.932	0.859	1.63E-10
TMSB10	1.17E-32	-0.63773	1	0.999	2.78E-28
CHD3	3.49E-64	-0.61585	0.386	0.772	8.30E-60
C11orf96	8.83E-29	-0.58922	0.247	0.527	2.10E-24
PFN2	1.07E-20	-0.56691	0.941	0.96	2.54E-16
NR2F1	4.01E-13	-0.52594	0.726	0.804	9.52E-09
DPYSL3	3.98E-17	-0.4986	0.823	0.892	9.47E-13
NRXN3	6.25E-20	-0.47093	0.224	0.434	1.49E-15
TUBB3	2.79E-08	-0.46577	0.973	0.986	0.000664
PPDPF	2.03E-36	-0.46546	0.994	0.991	4.82E-32
SCRG1	7.62E-20	-0.4569	0.2	0.409	1.81E-15
EPHA3	4.23E-25	-0.45532	0.194	0.432	1.01E-20
RPL38	2.89E-58	-0.45435	1	0.999	6.87E-54
DYNC1I2	1.41E-22	-0.45113	0.955	0.968	3.35E-18
GAP43	2.45E-12	-0.44493	0.692	0.762	5.83E-08
DST	4.92E-16	-0.44192	0.764	0.833	1.17E-11
ARL4C	4.29E-25	-0.43754	0.436	0.66	1.02E-20
CAMK2N1	1.62E-17	-0.43622	0.714	0.823	3.86E-13
RBP1	1.36E-08	-0.42612	0.374	0.486	0.000323
ASCL1	9.57E-08	-0.42526	0.614	0.676	0.002274
INSM1	8.24E-19	-0.42013	0.244	0.487	1.96E-14
CCDC88A	4.34E-09	-0.40615	0.952	0.957	0.000103
RORB	9.82E-26	-0.39909	0.383	0.638	2.33E-21
RPS29	6.13E-53	-0.39509	1	0.996	1.46E-48
ELAVL4	5.07E-09	-0.39294	0.528	0.622	0.00012
RPS27	4.25E-45	-0.38338	1	1	1.01E-40
AL627171.2	2.04E-18	-0.37749	0.926	0.951	4.86E-14
RPL37A	2.19E-50	-0.37091	1	0.999	5.21E-46
RGS16	3.13E-10	-0.36823	0.313	0.48	7.44E-06
CD24	1.10E-11	-0.36172	0.902	0.929	2.61E-07
MT2A	1.67E-10	-0.36135	0.436	0.568	3.98E-06
RPS21	1.65E-38	-0.3594	0.997	1	3.92E-34
ADGRV1	4.22E-17	-0.35628	0.478	0.654	1.00E-12
NR2F2	2.06E-20	-0.35518	0.113	0.316	4.90E-16
CCNE2	2.04E-12	-0.34893	0.347	0.501	4.84E-08

INA	4.72E-11	-0.34178	0.236	0.382	1.12E-06
GSX2	8.55E-25	-0.33775	0.105	0.347	2.03E-20
STK39	5.36E-14	-0.32664	0.617	0.739	1.27E-09
MLLT11	3.88E-07	-0.32657	0.892	0.905	0.00922
HIST1H2AM	1.34E-13	-0.32532	0.465	0.621	3.17E-09
DCLK2	8.48E-12	-0.32207	0.347	0.486	2.02E-07
ZNF704	6.93E-12	-0.32084	0.779	0.853	1.65E-07
ID2	8.51E-12	-0.31576	0.433	0.586	2.02E-07
HSPA1A	2.84E-08	-0.31519	0.168	0.29	0.000675
KLF7	6.71E-13	-0.31183	0.414	0.565	1.59E-08
CITED2	4.48E-09	-0.30833	0.534	0.633	0.000107
ACTB	3.70E-22	-0.30286	1	1	8.79E-18
STMN1	8.12E-13	-0.2997	1	0.999	1.93E-08
NDUFA3	1.93E-18	-0.29957	0.932	0.958	4.58E-14
CEP170	1.20E-10	-0.29925	0.854	0.919	2.85E-06
HDAC2	1.71E-19	-0.29584	0.974	0.987	4.06E-15
CDK6	7.62E-14	-0.29257	0.362	0.548	1.81E-09
RAB3IP	7.81E-16	-0.29196	0.343	0.537	1.86E-11
CRMP1	6.75E-12	-0.2889	0.889	0.924	1.61E-07
DPYSL2	7.35E-18	-0.2867	0.983	0.993	1.75E-13
GNG4	1.79E-16	-0.28546	0.847	0.914	4.27E-12
GADD45G	7.26E-09	-0.28351	0.411	0.553	0.000173
SPECC1	1.36E-14	-0.2822	0.49	0.674	3.24E-10
HES6	3.83E-07	-0.27991	0.974	0.986	0.009104
LNPK	3.22E-12	-0.27971	0.635	0.768	7.66E-08
DLEU2	7.74E-10	-0.2765	0.654	0.752	1.84E-05
BCL11B	1.97E-06	-0.27635	0.705	0.755	0.04682
CHMP2A	4.11E-16	-0.27519	0.62	0.761	9.76E-12
LIMA1	8.83E-08	-0.27203	0.391	0.493	0.002099
SEPTIN11	8.95E-12	-0.27132	0.971	0.988	2.13E-07
SMC3	2.30E-14	-0.27043	0.977	0.988	5.47E-10
IGFBPL1	3.06E-10	-0.26939	0.376	0.546	7.28E-06
ELAVL3	6.64E-07	-0.26236	0.785	0.831	0.015791
C4orf48	9.44E-11	-0.26196	0.953	0.974	2.24E-06
SCG3	1.21E-11	-0.26177	0.191	0.357	2.87E-07
SRRM4	2.01E-13	-0.25915	0.191	0.373	4.77E-09
OLA1	1.74E-10	-0.25912	0.914	0.942	4.13E-06
DCLK1	3.59E-09	-0.25802	0.427	0.546	8.52E-05
MGST1	4.42E-08	-0.25764	0.212	0.323	0.00105
CFL1	1.30E-12	-0.25409	0.997	0.997	3.08E-08
LMO4	1.42E-09	-0.25019	0.642	0.751	3.38E-05
PNRC1	1.17E-11	0.250632	0.949	0.909	2.79E-07
RAB3B	1.04E-13	0.25104	0.367	0.195	2.47E-09
FGFBP3	1.23E-10	0.254011	0.844	0.719	2.93E-06
CNTLN	2.89E-09	0.254578	0.887	0.785	6.87E-05
KDM5B	4.42E-09	0.254759	0.854	0.748	0.000105
TMEM161B-AS1	1.31E-14	0.254919	0.97	0.937	3.10E-10
MPPED2	3.56E-15	0.255039	0.911	0.767	8.47E-11
RAP1B	5.28E-15	0.256731	0.856	0.795	1.26E-10
NFIA	1.17E-07	0.257737	0.97	0.947	0.002792

PDCD4	1.75E-14	0.259015	0.818	0.659	4.15E-10
ARGLU1	2.38E-15	0.259783	0.985	0.984	5.65E-11
PGM1	1.88E-16	0.260739	0.672	0.504	4.47E-12
MAGED2	3.58E-16	0.26226	0.917	0.878	8.51E-12
FHL1	2.87E-17	0.262453	0.795	0.666	6.83E-13
PAX6	1.64E-13	0.264206	0.952	0.84	3.90E-09
MEIS2	2.53E-13	0.264843	0.964	0.929	6.02E-09
RASGRP1	1.40E-20	0.265445	0.457	0.241	3.33E-16
CD63	2.03E-17	0.265543	0.964	0.937	4.83E-13
FILIP1	2.69E-19	0.26588	0.477	0.248	6.38E-15
MYO10	1.09E-16	0.267424	0.836	0.697	2.58E-12
CYP51A1	4.63E-13	0.269207	0.959	0.921	1.10E-08
CMBL	2.98E-17	0.2695	0.729	0.54	7.07E-13
REST	3.93E-17	0.269564	0.699	0.509	9.35E-13
CDON	7.23E-18	0.271461	0.666	0.464	1.72E-13
VCAN	7.03E-12	0.275958	0.925	0.869	1.67E-07
NCKAP5	3.82E-24	0.277049	0.538	0.294	9.07E-20
COL4A6	2.12E-25	0.277723	0.549	0.293	5.04E-21
EML1	1.43E-18	0.27914	0.641	0.434	3.40E-14
SYNE2	2.36E-11	0.285163	0.992	0.988	5.62E-07
EMX2OS	7.57E-27	0.286397	0.591	0.308	1.80E-22
PNISR	1.19E-18	0.286754	0.982	0.974	2.84E-14
MT-ND2	6.32E-12	0.289235	0.991	0.988	1.50E-07
MASP1	1.68E-20	0.293138	0.577	0.349	3.99E-16
FOXP1	5.95E-14	0.294984	0.89	0.785	1.42E-09
FZD8	1.17E-15	0.298809	0.713	0.493	2.78E-11
CPNE3	1.43E-17	0.299063	0.863	0.782	3.39E-13
CREB5	2.42E-12	0.30111	0.847	0.749	5.75E-08
IFI44L	1.69E-18	0.303056	0.627	0.406	4.02E-14
TLE4	3.78E-18	0.306854	0.857	0.637	8.98E-14
NEUROG2	7.08E-15	0.308322	0.586	0.375	1.68E-10
MT-CO3	4.19E-19	0.309873	1	0.999	9.95E-15
GSTP1	5.55E-25	0.312735	0.997	0.987	1.32E-20
MDM1	3.34E-20	0.313036	0.675	0.468	7.94E-16
NKAIN3	8.18E-19	0.316131	0.744	0.591	1.94E-14
ISYNA1	8.03E-22	0.317252	0.842	0.712	1.91E-17
DOK5	6.32E-19	0.318754	0.838	0.673	1.50E-14
GAS1	5.21E-22	0.322869	0.744	0.546	1.24E-17
SFRP1	3.66E-13	0.329195	0.974	0.957	8.69E-09
CNTNAP2	1.57E-23	0.331656	0.617	0.357	3.74E-19
RAB11FIP2	1.20E-24	0.333292	0.758	0.545	2.86E-20
NTRK3	1.74E-26	0.334317	0.687	0.427	4.14E-22
EMP3	1.68E-22	0.344582	0.549	0.313	4.00E-18
INTU	5.02E-23	0.348498	0.862	0.712	1.19E-18
EMX1	1.09E-22	0.349688	0.795	0.496	2.60E-18
LIX1	8.28E-19	0.358882	0.714	0.503	1.97E-14
GNAI2	4.71E-22	0.360469	0.958	0.95	1.12E-17
CLU	4.80E-19	0.36121	0.851	0.669	1.14E-14
LRRC3B	1.24E-25	0.362442	0.654	0.409	2.94E-21
HMGN3	3.31E-22	0.365597	0.98	0.973	7.86E-18

C1orf61	1.95E-21	0.367432	0.995	0.996	4.62E-17
PLCB1	4.47E-30	0.372574	0.72	0.437	1.06E-25
VIM	7.26E-15	0.407627	0.964	0.934	1.72E-10
AMBN	3.03E-34	0.408631	0.633	0.307	7.20E-30
EFNB2	1.89E-22	0.417163	0.904	0.833	4.49E-18
LINC01551	1.40E-24	0.422594	0.971	0.939	3.33E-20
MFAP2	6.23E-31	0.425948	0.726	0.509	1.48E-26
LEF1	1.23E-41	0.431509	0.785	0.486	2.93E-37
FEZF2	3.68E-35	0.446332	0.669	0.36	8.75E-31
DACH1	5.82E-29	0.446958	0.869	0.646	1.38E-24
RPL22L1	4.16E-23	0.449336	0.907	0.807	9.88E-19
PHLDA1	2.64E-29	0.455389	0.8	0.526	6.28E-25
TXNIP	8.88E-17	0.461227	0.795	0.666	2.11E-12
SOX3	1.16E-33	0.465562	0.788	0.533	2.76E-29
DMRTA2	8.63E-32	0.475404	0.764	0.451	2.05E-27
TTYH1	2.02E-26	0.552614	0.904	0.793	4.80E-22
HMGA2	2.17E-36	0.567005	0.841	0.595	5.15E-32
EMX2	2.98E-40	0.572626	0.887	0.628	7.09E-36
B3GAT2	4.25E-41	0.59902	0.734	0.467	1.01E-36
ZFP36L1	1.09E-29	0.611207	0.871	0.71	2.59E-25
PTN	4.59E-29	0.624675	0.929	0.725	1.09E-24
HES1	2.12E-26	0.642114	0.85	0.615	5.04E-22
LHX2	1.17E-50	0.720298	0.92	0.671	2.79E-46
ID4	4.55E-57	1.182786	0.968	0.879	1.08E-52

RG	**p_val**	**avg_log2FC**	**pct.1**	**pct.2**	**p_val_adj**
ADGRV1	2.05E-37	-0.64027	0.343	0.564	4.87E-33
MT3	7.15E-16	-0.6375	0.229	0.365	1.70E-11
MIR100HG	1.76E-33	-0.52147	0.5	0.721	4.18E-29
SCRG1	7.35E-23	-0.48495	0.128	0.282	1.75E-18
HOPX	4.64E-27	-0.46414	0.244	0.45	1.10E-22
FABP7	7.25E-14	-0.45977	0.968	0.966	1.72E-09
ZEB2	5.72E-10	-0.42957	0.336	0.433	1.36E-05
DLK1	7.09E-21	-0.42728	0.174	0.329	1.68E-16
ID2	1.78E-21	-0.40481	0.38	0.555	4.23E-17
RPS29	2.10E-79	-0.39424	0.999	0.999	4.99E-75
CHMP2A	1.36E-37	-0.39326	0.646	0.826	3.22E-33
VEGFA	6.39E-14	-0.38325	0.617	0.749	1.52E-09
DDIT4	1.07E-10	-0.3776	0.728	0.816	2.54E-06
RORB	6.00E-27	-0.37466	0.202	0.388	1.43E-22
SCD	7.08E-16	-0.37273	0.953	0.98	1.68E-11
RPS27	1.52E-75	-0.3706	0.999	0.999	3.62E-71
MIR99AHG	1.53E-26	-0.36889	0.61	0.777	3.65E-22
NR2F1	4.09E-12	-0.3636	0.61	0.704	9.71E-08
CHD3	1.30E-45	-0.36324	0.246	0.513	3.09E-41
EGR1	6.67E-09	-0.35864	0.914	0.935	0.000158
RPL38	8.47E-68	-0.35449	0.997	0.998	2.01E-63
RPL37A	1.60E-80	-0.32621	0.999	1	3.80E-76
STK39	2.16E-12	-0.32553	0.459	0.576	5.14E-08

TMEM158	2.98E-19	-0.32272	0.384	0.548	7.07E-15
CCN1	4.34E-11	-0.32179	0.74	0.811	1.03E-06
EPHA3	4.07E-17	-0.32077	0.13	0.256	9.67E-13
CDO1	1.08E-12	-0.31614	0.534	0.661	2.58E-08
FAM107A	2.00E-19	-0.3141	0.113	0.251	4.75E-15
RPS21	4.40E-52	-0.31406	1	0.997	1.05E-47
SHTN1	2.82E-17	-0.31169	0.108	0.229	6.69E-13
SEPTIN11	3.42E-14	-0.30994	0.962	0.967	8.12E-10
TOMM7	5.18E-36	-0.3039	0.964	0.987	1.23E-31
SNHG25	2.47E-30	-0.29966	0.446	0.663	5.86E-26
HMGCS1	2.36E-12	-0.2989	0.987	0.993	5.60E-08
PPDPF	2.64E-21	-0.29645	0.969	0.979	6.26E-17
MT2A	1.89E-09	-0.29083	0.32	0.418	4.49E-05
MEST	5.09E-20	-0.28946	0.27	0.438	1.21E-15
MGST1	4.80E-23	-0.28937	0.175	0.345	1.14E-18
AL139246.5	3.00E-17	-0.2879	0.219	0.372	7.13E-13
RAB31	3.80E-28	-0.28763	0.14	0.327	9.03E-24
C11orf96	1.30E-19	-0.28435	0.122	0.263	3.09E-15
TNC	3.15E-18	-0.28279	0.113	0.248	7.49E-14
VIM	3.65E-26	-0.27311	0.999	0.999	8.68E-22
SALL3	1.87E-22	-0.26015	0.101	0.249	4.43E-18
RGS16	7.16E-13	-0.25877	0.213	0.343	1.70E-08
HERPUD1	9.40E-08	-0.25631	0.586	0.681	0.002233
RHOB	6.88E-08	-0.25592	0.646	0.728	0.001635
SNHG6	2.30E-25	-0.2545	0.991	0.993	5.48E-21
P4HA1	1.07E-06	-0.25355	0.697	0.76	0.025525
MCM3	2.19E-11	0.251415	0.567	0.459	5.21E-07
SIVA1	2.53E-15	0.255669	0.814	0.783	6.02E-11
NCKAP5	1.30E-17	0.259134	0.538	0.395	3.09E-13
NUCKS1	1.60E-26	0.264242	0.994	0.99	3.81E-22
MSH6	5.66E-17	0.264826	0.821	0.757	1.35E-12
PLCB1	1.71E-19	0.265165	0.578	0.437	4.07E-15
TUBA1B	1.67E-14	0.265399	0.979	0.977	3.98E-10
PCSK1N	5.20E-18	0.267016	0.703	0.602	1.24E-13
NUDT1	2.22E-16	0.26974	0.699	0.593	5.28E-12
PRDX2	3.81E-29	0.269935	0.993	0.986	9.05E-25
NKAIN3	1.96E-18	0.273254	0.623	0.5	4.66E-14
ORC6	5.67E-12	0.274595	0.503	0.385	1.35E-07
SYT1	6.56E-21	0.27735	0.594	0.437	1.56E-16
H3F3A	9.10E-41	0.280878	1	0.997	2.16E-36
SOX3	4.24E-18	0.281471	0.811	0.718	1.01E-13
BEX2	8.70E-18	0.286226	0.75	0.651	2.07E-13
LRRC3B	9.01E-24	0.287503	0.69	0.532	2.14E-19
MTLN	4.62E-43	0.289613	0.378	0.14	1.10E-38
RASGRP1	2.56E-18	0.290269	0.469	0.328	6.09E-14
SNRPB	5.06E-19	0.292381	0.947	0.93	1.20E-14
GINS2	6.77E-13	0.293057	0.639	0.54	1.61E-08
MPPED2	8.45E-24	0.293683	0.839	0.738	2.01E-19
LEF1	9.01E-23	0.295312	0.666	0.521	2.14E-18
HMGB3	7.52E-20	0.30001	0.844	0.773	1.79E-15

VCAN	4.48E-19	0.301717	0.908	0.883	1.07E-14
CNTNAP2	1.04E-23	0.303111	0.6	0.421	2.48E-19
MT-CYB	1.27E-20	0.303294	0.978	0.986	3.03E-16
GSTP1	1.04E-38	0.313378	0.997	0.992	2.47E-34
DUT	6.09E-12	0.315159	0.883	0.87	1.45E-07
PSIP1	1.07E-32	0.317433	0.974	0.964	2.54E-28
MDM1	5.32E-29	0.327051	0.517	0.332	1.27E-24
EMP3	3.63E-28	0.329077	0.497	0.31	8.62E-24
NASP	8.47E-23	0.332008	0.942	0.935	2.01E-18
SYNE2	1.91E-19	0.343882	0.957	0.956	4.55E-15
PHLDA1	1.04E-21	0.344419	0.775	0.645	2.48E-17
TXNIP	9.96E-21	0.345966	0.856	0.79	2.37E-16
PCLAF	1.51E-06	0.352368	0.501	0.428	0.03586
MT-CO3	2.49E-28	0.354844	0.99	0.992	5.92E-24
MT-ATP6	2.89E-27	0.354947	0.99	0.99	6.86E-23
H2AFZ	5.62E-28	0.37142	0.994	0.988	1.34E-23
HMGA2	2.31E-29	0.404231	0.838	0.701	5.49E-25
LHX2	4.09E-41	0.410102	0.946	0.894	9.72E-37
MFAP2	5.05E-37	0.436216	0.717	0.553	1.20E-32
AMBN	1.79E-54	0.483255	0.521	0.241	4.26E-50
BEX1	5.89E-41	0.489302	0.888	0.769	1.40E-36
B3GAT2	1.50E-43	0.493028	0.774	0.601	3.57E-39
ID4	8.33E-56	0.529342	0.996	0.982	1.98E-51

IP	**p_val**	**avg_log2FC**	**pct.1**	**pct.2**	**p_val_adj**
CHD3	1.09E-21	-0.73634	0.552	0.817	2.60E-17
FABP7	2.32E-07	-0.48121	0.591	0.779	0.005524
PANTR1	3.32E-07	-0.46511	0.727	0.862	0.007884
RPS29	7.58E-29	-0.44588	0.989	1	1.80E-24
RPS21	1.17E-22	-0.37591	0.992	1	2.78E-18
RPL38	3.84E-24	-0.3637	0.992	0.997	9.12E-20
PABPC1	2.13E-08	-0.3614	0.859	0.948	0.000506
RPS27	2.38E-21	-0.35187	0.994	1	5.65E-17
STK17A	8.62E-07	-0.34341	0.575	0.761	0.020497
SNHG6	4.06E-07	-0.33341	0.936	0.986	0.00965
CHMP2A	6.61E-10	-0.31303	0.599	0.799	1.57E-05
TLE4	1.14E-06	-0.306	0.425	0.623	0.026976
PPDPF	2.22E-10	-0.30262	0.936	0.976	5.27E-06
RPL37A	1.29E-21	-0.30085	0.994	1	3.05E-17
RPL36A	4.00E-10	-0.28352	0.978	0.993	9.51E-06
RPL22	1.24E-11	-0.26982	0.989	0.997	2.95E-07
RPS28	8.25E-14	-0.26957	0.989	1	1.96E-09
RPL41	1.17E-08	-0.26646	0.997	1	0.000279
RPS20	9.10E-09	-0.26213	0.986	0.997	0.000216
RPL39	4.98E-15	-0.25989	0.986	0.997	1.18E-10
COMT	3.62E-07	-0.255	0.376	0.592	0.0086
RPL31	2.26E-10	-0.25199	0.986	0.993	5.37E-06
NNAT	2.87E-07	0.335627	0.986	0.796	0.006824
TSPAN18	2.49E-07	0.350592	0.577	0.436	0.005917

MT-CYB	1.86E-07	0.362481	0.975	0.986	0.004417
MT-ATP6	1.17E-07	0.374262	0.981	0.993	0.002778
MFAP2	1.47E-06	0.383602	0.547	0.436	0.035035
DLL3	3.41E-07	0.467907	0.859	0.82	0.008092
KHDRBS3	1.46E-09	0.509568	0.456	0.26	3.47E-05
CRABP1	3.41E-08	0.618814	0.591	0.381	0.00081

N1	**p_val**	**avg_log2FC**	**pct.1**	**pct.2**	**p_val_adj**
CHD3	2.74E-102	-1.14441	0.399	0.81	6.50E-98
BCL11B	3.21E-23	-0.6462	0.808	0.898	7.62E-19
ZFHX3	1.48E-08	-0.54288	0.583	0.664	0.000352
LMO4	2.88E-18	-0.52832	0.311	0.498	6.85E-14
EGR3	1.16E-21	-0.51846	0.231	0.44	2.76E-17
SOX2	3.76E-15	-0.51056	0.471	0.634	8.94E-11
SOX2-OT	2.70E-15	-0.49638	0.583	0.74	6.41E-11
DCX	5.73E-35	-0.45626	0.893	0.956	1.36E-30
PDE4DIP	1.53E-18	-0.45576	0.393	0.577	3.63E-14
SOX11	1.56E-26	-0.43776	0.962	0.981	3.70E-22
SOX4	5.63E-30	-0.43651	0.986	0.996	1.34E-25
NPAS4	2.09E-13	-0.40061	0.175	0.326	4.98E-09
ROBO1	1.89E-12	-0.39582	0.29	0.438	4.49E-08
AL627171.2	7.24E-28	-0.39305	0.789	0.903	1.72E-23
EGR1	2.15E-09	-0.39181	0.743	0.81	5.11E-05
ASCL1	4.00E-08	-0.39122	0.215	0.326	0.000951
RPS29	2.61E-43	-0.38158	0.955	0.976	6.20E-39
DLX6	4.71E-18	-0.37957	0.316	0.529	1.12E-13
PPDPF	1.49E-22	-0.3771	0.819	0.92	3.54E-18
RPL38	1.74E-39	-0.37491	0.955	0.972	4.13E-35
LSAMP	3.93E-13	-0.37367	0.27	0.426	9.33E-09
RPS27	2.44E-34	-0.37235	0.977	0.982	5.80E-30
NSG2	1.50E-11	-0.37206	0.61	0.738	3.57E-07
ZNF503	1.63E-07	-0.36065	0.163	0.262	0.003875
MIR100HG	1.35E-13	-0.35951	0.512	0.665	3.22E-09
PID1	8.33E-15	-0.3579	0.142	0.288	1.98E-10
AC004158.1	4.42E-10	-0.35756	0.56	0.682	1.05E-05
SETBP1	3.55E-13	-0.34901	0.37	0.535	8.44E-09
CCDC88A	3.91E-13	-0.34884	0.802	0.905	9.30E-09
PCDH10	1.29E-08	-0.34794	0.296	0.417	0.000306
ACTB	5.36E-17	-0.3442	0.965	0.981	1.27E-12
SOX1	1.37E-11	-0.34206	0.151	0.28	3.27E-07
RPS21	3.75E-29	-0.34015	0.974	0.984	8.92E-25
NR2F1	1.63E-11	-0.33927	0.431	0.588	3.88E-07
MN1	5.00E-16	-0.33912	0.31	0.497	1.19E-11
CHMP2A	5.59E-19	-0.33903	0.495	0.682	1.33E-14
DPYSL2	6.43E-22	-0.33705	0.867	0.944	1.53E-17
GPI	3.14E-07	-0.33562	0.435	0.555	0.007452
DSEL	7.79E-09	-0.33071	0.101	0.198	0.000185
PLXNA2	1.76E-14	-0.32782	0.231	0.4	4.19E-10
RPL37A	6.29E-35	-0.3229	0.986	0.986	1.49E-30

RPL37	2.31E-29	-0.32158	0.977	0.981	5.48E-25
CCND2	5.26E-14	-0.31672	0.506	0.703	1.25E-09
H1F0	5.44E-15	-0.31559	0.267	0.45	1.29E-10
DPYSL3	1.46E-15	-0.31554	0.807	0.904	3.47E-11
INPP5F	8.42E-13	-0.31384	0.273	0.421	2.00E-08
EPHA5	5.36E-11	-0.31075	0.142	0.267	1.27E-06
HNRNPH1	3.58E-19	-0.31065	0.787	0.914	8.50E-15
LIMA1	7.34E-09	-0.30773	0.199	0.31	0.000174
GRIK3	9.14E-16	-0.30762	0.115	0.271	2.17E-11
PEG10	1.20E-07	-0.30567	0.724	0.819	0.002849
TLE5	2.24E-16	-0.30339	0.798	0.889	5.33E-12
DLX1	2.67E-10	-0.30336	0.337	0.5	6.34E-06
INA	4.65E-16	-0.29831	0.748	0.876	1.10E-11
LY6H	7.37E-11	-0.2969	0.32	0.474	1.75E-06
PCBP2	2.08E-17	-0.29639	0.743	0.887	4.95E-13
RGS2	3.54E-08	-0.29195	0.343	0.469	0.00084
IRS2	1.15E-09	-0.28725	0.292	0.43	2.74E-05
GAD1	1.12E-09	-0.28715	0.296	0.444	2.67E-05
DLX2	7.86E-09	-0.28607	0.467	0.611	0.000187
PPP1CB	6.18E-15	-0.28195	0.577	0.752	1.47E-10
ATP5F1E	2.32E-16	-0.27883	0.921	0.957	5.52E-12
UBB	9.01E-12	-0.27861	0.621	0.796	2.14E-07
HIST1H4D	5.82E-09	-0.2758	0.24	0.366	0.000138
QKI	4.99E-12	-0.27528	0.642	0.783	1.19E-07
SCD	3.20E-08	-0.27335	0.579	0.703	0.00076
MTSS1	9.85E-10	-0.27077	0.586	0.711	2.34E-05
HEXIM1	5.14E-11	-0.27045	0.251	0.393	1.22E-06
GNG2	2.38E-12	-0.26761	0.489	0.65	5.65E-08
MAFB	1.66E-07	-0.26454	0.192	0.3	0.003941
APC2	1.10E-08	-0.26232	0.329	0.466	0.000261
RPL34	1.39E-16	-0.26081	0.974	0.976	3.30E-12
CCNI	2.43E-13	-0.26011	0.894	0.949	5.78E-09
ATP5MD	9.77E-17	-0.25936	0.784	0.882	2.32E-12
RIPOR2	8.57E-07	-0.25747	0.34	0.445	0.020374
TRIB2	2.63E-08	-0.25717	0.289	0.411	0.000625
ZKSCAN1	1.88E-08	-0.25651	0.521	0.647	0.000447
SPATS2L	5.28E-08	-0.25609	0.177	0.286	0.001254
EIF4G2	5.49E-13	-0.25573	0.894	0.953	1.30E-08
TMEM123	1.02E-07	-0.25298	0.444	0.564	0.002413
AKAP9	8.54E-08	-0.2527	0.801	0.885	0.002029
RPL39	6.17E-19	-0.25253	0.964	0.983	1.47E-14
RBFOX1	1.63E-08	-0.25112	0.172	0.282	0.000388
ELAVL3	4.17E-13	-0.251	0.77	0.887	9.90E-09
PFKFB3	1.84E-07	-0.25002	0.201	0.31	0.004379
PRMT1	1.96E-06	0.252892	0.733	0.752	0.046611
CXXC5	9.30E-08	0.266044	0.32	0.228	0.002211
H2AFZ	9.94E-11	0.272621	0.9	0.923	2.36E-06
PLCB1	1.17E-08	0.30081	0.192	0.11	0.000277
RPS2	6.51E-19	0.301393	0.982	0.974	1.55E-14
PRDX2	3.96E-13	0.30787	0.909	0.926	9.42E-09

H3F3B	6.47E-20	0.322865	0.977	0.984	1.54E-15
NKAIN4	2.67E-08	0.323734	0.35	0.26	0.000635
ARL2	2.67E-07	0.335961	0.385	0.303	0.006355
CPE	1.37E-07	0.344508	0.68	0.659	0.003259
CKB	2.00E-11	0.349806	0.947	0.966	4.75E-07
CALM1	2.84E-15	0.361789	0.914	0.925	6.74E-11
VIM	3.52E-11	0.375545	0.573	0.465	8.38E-07
CDCA7L	1.01E-06	0.391935	0.272	0.198	0.024056
PAX6	1.67E-10	0.403293	0.437	0.32	3.97E-06
TXNIP	4.50E-07	0.412674	0.633	0.584	0.010694
MT-ATP6	1.53E-06	0.434485	0.956	0.959	0.03626
TTYH1	1.52E-13	0.516065	0.45	0.34	3.62E-09
PTN	2.05E-11	0.534608	0.267	0.162	4.86E-07
THSD7A	3.62E-11	0.588904	0.248	0.146	8.59E-07

N2	**p_val**	**avg_log2FC**	**pct.1**	**pct.2**	**p_val_adj**
CHD3	1.03E-107	-1.15338	0.784	0.931	2.45E-103
LMO3	8.51E-10	-0.71362	0.411	0.564	2.02E-05
SLA	1.43E-13	-0.47528	0.715	0.824	3.39E-09
FRMD4B	4.27E-10	-0.45056	0.48	0.622	1.01E-05
FABP7	5.41E-12	-0.44165	0.468	0.622	1.29E-07
MEIS2	1.89E-08	-0.42243	0.746	0.8	0.0004503
MN1	1.29E-13	-0.3919	0.282	0.463	3.08E-09
RORB	5.71E-14	-0.37529	0.352	0.52	1.36E-09
NTM	2.13E-08	-0.36632	0.248	0.384	0.0005073
DPY19L1	1.58E-06	-0.36511	0.171	0.266	0.0375305
ROBO1	1.73E-16	-0.36063	0.4	0.593	4.11E-12
STK17A	9.16E-07	-0.34724	0.732	0.772	0.0217652
RPS21	9.91E-25	-0.32417	0.983	0.977	2.36E-20
RPS29	2.06E-31	-0.30971	0.988	0.977	4.89E-27
RPS27	8.75E-25	-0.30409	0.997	0.988	2.08E-20
RPL37A	1.09E-22	-0.29679	0.997	0.992	2.58E-18
CHMP2A	2.26E-12	-0.29374	0.741	0.815	5.37E-08
ROBO2	9.80E-08	-0.29338	0.23	0.352	0.0023297
DCX	9.28E-10	-0.28821	0.973	0.962	2.20E-05
LIMCH1	1.32E-10	-0.28326	0.528	0.68	3.13E-06
RPL39	7.15E-21	-0.28206	0.991	0.989	1.70E-16
GPM6A	2.72E-24	-0.27741	0.999	0.988	6.46E-20
RPL36A	5.88E-18	-0.26491	0.991	0.979	1.40E-13
SOX4	6.82E-15	-0.26157	1	0.995	1.62E-10
INPP5F	1.34E-09	-0.25857	0.554	0.673	3.18E-05
HNRNPA1	3.60E-21	-0.2581	0.984	0.972	8.55E-17
EMX2	5.75E-10	0.252813	0.489	0.344	1.37E-05
MYL6	3.98E-15	0.252972	0.978	0.943	9.46E-11
NUCB2	1.21E-16	0.253067	0.841	0.757	2.87E-12
DYNLL1	3.15E-13	0.258264	0.99	0.96	7.47E-09
EMX1	5.93E-11	0.259631	0.831	0.725	1.41E-06
AL136366.1	4.53E-14	0.259886	0.505	0.327	1.08E-09
RAP1B	3.13E-14	0.265808	0.834	0.746	7.44E-10

AURKAIP1	1.55E-12	0.266559	0.841	0.751	3.68E-08
SSTR2	1.27E-10	0.27421	0.859	0.768	3.02E-06
HSPB1	8.05E-16	0.275879	0.705	0.517	1.91E-11
HIST1H4C	1.94E-06	0.276468	0.921	0.876	0.046157
TM7SF2	1.71E-13	0.287456	0.746	0.628	4.07E-09
FEZF2	7.72E-13	0.289826	0.847	0.709	1.84E-08
GSTP1	1.33E-13	0.314187	0.869	0.801	3.16E-09
KHDRBS3	5.65E-15	0.340758	0.604	0.428	1.34E-10
CHD4	4.95E-19	0.358205	0.9	0.826	1.18E-14
NXPH4	2.28E-14	0.398659	0.744	0.59	5.43E-10
LMO4	2.23E-15	0.404641	0.836	0.7	5.30E-11
CALB2	3.51E-08	0.467437	0.56	0.422	0.0008342
LHX2	1.31E-22	0.56674	0.816	0.639	3.12E-18
CRABP1	6.02E-24	0.726833	0.569	0.309	1.43E-19

Table S5. Primer sequences for CRISPR-Cas9 target and off-target regions

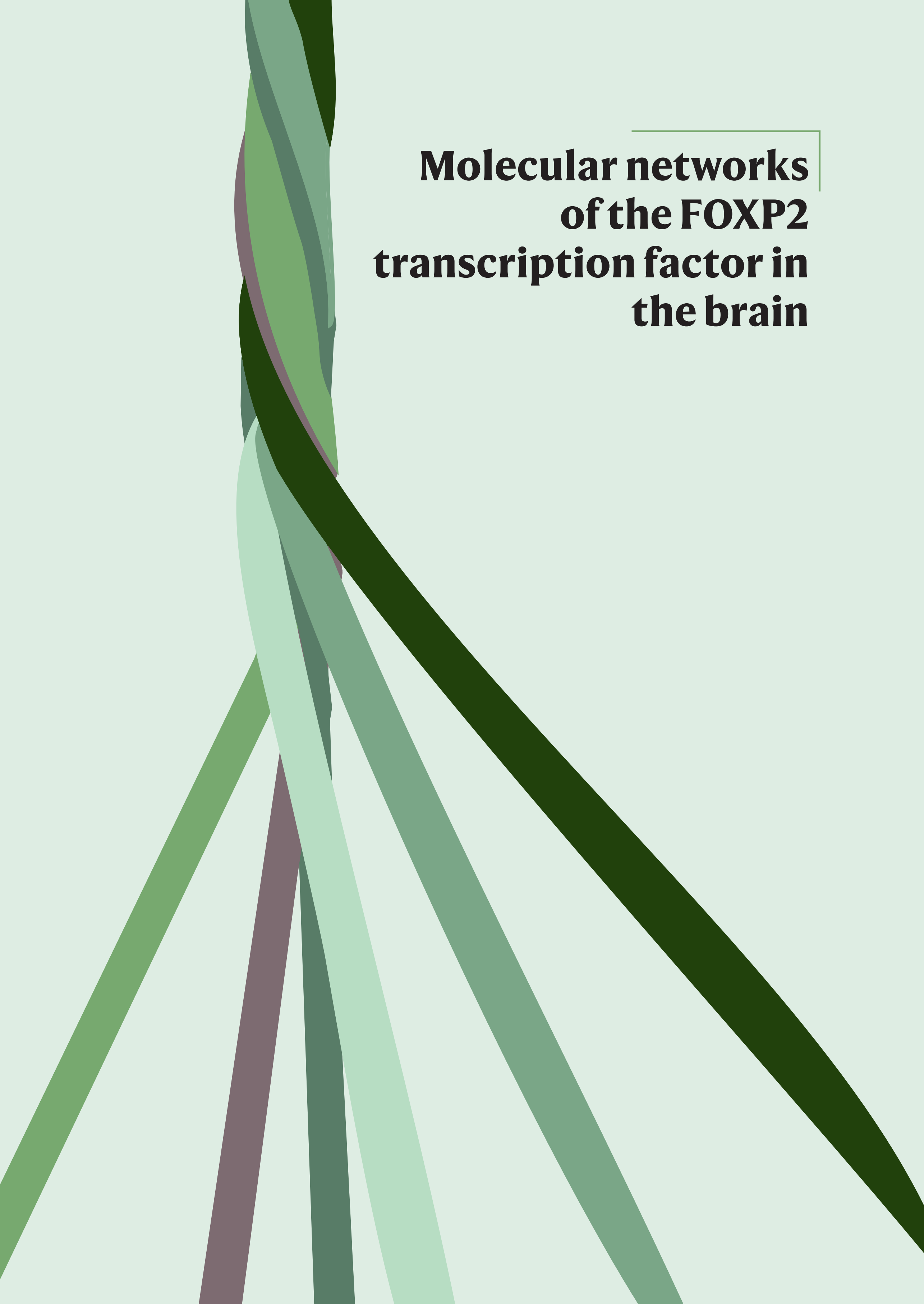
Target	Primer sequence
CHD3-target-F	5’-ATGTGCTGAGAACAGTTTCTGG-3’,
CHD3-target-R	5’-CATGCTCCACATCCTCCAGG-3’
Off-target1-F	5’-TGTGTAAAGGACGGCTGTGG-3’
Off-target1-R	5’-TCCTAAGGGCACAAGCAAGG-3’
Off-target2-F	5’-GTGGAAACCCAGAGGCTTGA-3’
Off-target2-R	5’-TTTGCAGCCCCATACCAGAG-3’
Off-target3-F	5’-ACTCGGAGGCTGAGAACAAA-3’
Off-target3-R	5’-GCTGATCCTGACAGCTCCTC-3’
Off-target4-F	5’-AAACAAATGCGGCACAGGAC-3’
Off-target4-R	5’-CCTTCTTGGTCCAGAGGGGA-3’
Off-target5-F	5’-ACTTCATGACGCTCCGGTTT-3’
Off-target5-R	5’-CTCAAAAGCCCCTCTCCCTG-3’

Table S6. qPCR primer sequences.

Target	Primer sequence
CHD3-F	5'-AGGAAGACCAAGACAACCAGTCAG-3'
CHD3-R	5'-TGACTGTCTACGCCCTTCAGGA-3'
CHD4-F	5'-AGTGCTGCAACCATCCATACCTCT-3'
CHD4-R	5'-ATGCCCACCCTCCTTAAGGTTCTT-3'
CHD5-F	5'-TGCTTAAAGGAGCCCAAGTC-3'
CHD5-R	5'-TTGGTCAGCGTGTGGTAATC-3'
TBP-F	5'-GGGCACCACTCCACTGTATC-3'
TBP-R	5'-CGAAGTGCAATGGTCTTTAGG-3'
PPIA-F	5'-TATCTGCACTGCCAAGACTGAGTG-3'
PPIA-R	5'-CTTCTTGCTGGTCTTGCCATTCC-3'

6

This chapter has been published as:
Den Hoed, J., Devaraju, K., & Fisher, S. E. (2021). Molecular networks of the FOXP2 transcription factor in the brain. EMBO reports, e52803.



Molecular networks of the FOXP2 transcription factor in the brain

Abstract

The discovery of the FOXP2 transcription factor, and its implication in a rare severe human speech and language disorder, has led to two decades of empirical studies focused on uncovering its roles in the brain using a range of *in vitro* and *in vivo* methods. Here, we discuss what we have learned about the regulation of FOXP2, its downstream effectors and its modes of action as a transcription factor in brain development and function, providing an integrated overview of what is currently known about the critical molecular networks.

Introduction

FOXP2 was the first gene to be clearly linked to speech and language development. The initial finding was made through studies of a large multi-generational family (the KE family) with a severe dominantly-inherited developmental speech and language disorder (MIM #602081)[1]. All fifteen affected family members carried a heterozygous missense mutation (p.R553H) disrupting *FOXP2*. In the two decades since then, additional cases of FOXP2-related speech and language disorders have been discovered, both inherited and *de novo*[2-4], with childhood apraxia of speech (also called developmental verbal dyspraxia) as a core phenotypic feature, characterised by difficulties in coordinating sequences of articulatory movements underlying proficient speech. In a subset of individuals, broader phenotypes are observed including oral motor deficits, global developmental delays, and/or autism spectrum disorder[5]. Beyond the well-documented consequences of rare highly penetrant genetic disruptions, studies have investigated contributions of common variation in *FOXP2* to genetically complex traits. For example, some studies of small samples proposed that single nucleotide polymorphisms (SNPs) in the FOXP2 gene are associated with schizophrenia risk[6-8], but there is little evidence of replication[9]. Large-scale systematic genome-wide association studies have identified significant associations of intronic *FOXP2* SNPs with several traits, including attention-deficit/hyperactivity disorder (ADHD)[10], and risk-taking behaviours[11]. Although rare disruptions in *FOXP2* have been associated with changes in brain activity[12] and structure[13-15], common variation could not be linked to task-based neural activations on language tasks[16] or neuroanatomical differences between individuals[17].

FOXP2 belongs to the Forkhead box/winged-helix (FOX) family of proteins, a large group of transcription factors that share a highly conserved DNA-binding domain of ~80-100 amino acids, called the Forkhead box[18; 19] (following nomenclature guidelines, we use FOXP2 for humans, Foxp2 for mice and FoxP2 for other species). There are 19 subclasses of FOX proteins, from FOXA to FOXS[19; 20], with important roles in various biological processes, including cell differentiation, proliferation and development[19; 21]. Although they all share a characteristic DNA-binding domain, different FOX proteins have distinct expression patterns and are involved in diverse mechanisms[22].

The FOXP subclass comprises four members, FOXP1-4[23; 24]. As well as the DNA-binding domain, FOXP proteins share a zinc finger and leucine zipper motif (Figure 1A)[24; 25]. Moreover, FOXP1, FOXP2 and FOXP4 contain long N-terminal glutamine-rich regions of unknown function[24; 25]. A unique feature of the FOXP subclass is that they form homo- and heterodimers via the conserved leucine zipper, which appears essential for DNA binding and transcription regulation[24]. They may even form oligomer complexes, as detected for FoxP1, FoxP2 and FoxP4 in studies of zebra finch brain[26]. Formation of FOXP homo- and heterodimers in any particular tissue/cell-type is likely mediated by expression and availability of the different FOXP proteins, providing potential for more complex regulation of downstream pathways.

While FOXP3 expression and function is largely limited to the immune system[27], FOXP1, FOXP2 and FOXP4 are expressed in various tissues throughout the body, including the brain, where they show distinctive, yet partially overlapping, expression patterns (human foetal and postnatal expression of *FOXP1*, *FOXP2* and *FOXP4* based on BrainSpan expression data: Figure 1B and 1C. For a detailed review on the expression patterns of FOXP genes in the brain, see Ref. 28). FOXP1 expression is enriched in layers III-IV of the cerebral cortex[29; 30], as well as the thalamus, striatum, and CA1 sub region of the hippocampus[30]. Main sites of FOXP2 expression include layers IV-VI of the cerebral cortex[29-32], the striatum[30-33], the posterior and lateral thalamic nuclei[30-32], the Purkinje cells in the cerebellum[31; 32], and the inferior olive[30-32]. FOXP4 has been less well studied than the other FOXP proteins, but is expressed in the subventricular zone, throughout the cortical plate and in the striatum during embryonic development[34], and in Purkinje cells[35].

The roles of *FOXP2* have been investigated by studying its orthologues in an array of animal models. Mice that lack both alleles of *Foxp2* have severe motor impairments, developmental delays, and typically die by postnatal day 21[36], while heterozygous animals show no obvious differences compared to wild-type littermates, but display some altered vocal behaviours[37]. Mice that are heterozygous for the mutation originally identified in the KE family display reduced motor-skill learning[38] and produce shorter sequences of ultrasonic vocalisations with less complex syntax[39], as compared to wild-type littermates. *Foxp2* expression in the mouse cortex, striatum and cerebellum modulates different aspects of motor function, as demonstrated by conditional homozygous knock-outs targeting these structures[40]. However, selective deletion of the gene in each of these brain regions does not significantly alter production of ultrasonic vocalisations[41]. Interestingly, while selective deletion of *Foxp2* in the mouse cortex does not appear to impact development of cortical structures during embryogenesis[42; 43], cortical-specific knockouts are reported to nonetheless show altered social behaviours[42; 44]. When mouse *Foxp2* is constitutively replaced by a partially humanised version, medium spiny neurons in the striatum show increases in dendrite length and synaptic plasticity[45], consistent with multiple studies implicating the gene in development and function of corticostriatal circuitry[40; 46-50]. Moreover, knockdown and overexpression studies in the brains of zebra finches suggest that avian *FoxP2* is important not only in auditory-guided vocal learning during development, but also for maintenance of vocal behaviours in adulthood [51-55].

Notably, in humans, heterozygous disruptions of *FOXP1* and *FOXP4* have also been linked to neurodevelopmental disorders: an intellectual disability syndrome, frequently accompanied with autistic features and language impairment (MIM #613670)[56-60], and a milder developmental disorder with speech/language delays and congenital abnormalities[61], respectively. Some of the aetiological variants affect equivalent residues in the DNA-binding domain of these genes[61; 62]. While differences in the associated phenotypes are likely explained by the distinct expression patterns of the FOXP proteins, there are also regions of overlap where they can potentially form heterodimers. More thorough phenotypic comparison studies between these distinct neurodevelopmental disorders and functional follow-up would be required to uncover whether equivalent variants in *FOXP1* and *FOXP4* directly impact speech and language, or if they have an indirect effect on the function of FOXP2.

In depth studies of the functions of FOXP2 and its orthologues in brain development have involved not only mice and zebra finches (as noted above), but also other models

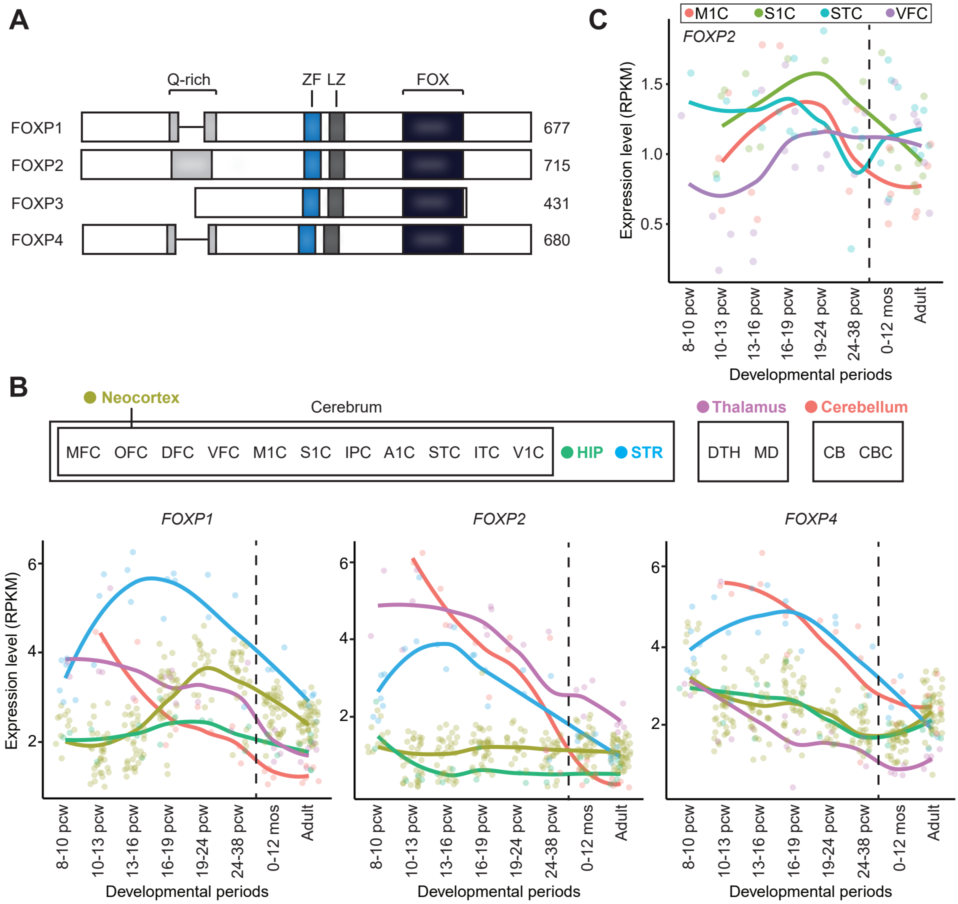


Figure 1. FOXP expression in the brain. A) Schematic representation of the FOXP family of proteins. The polyglutamine rich region is shaded in light grey (Q-rich), the zinc finger domain in blue (ZF), the leucine zipper in dark grey (LZ) and the Forkhead domain in black (FOX). **B**) Expression patterns of FOXP1, FOXP2 and FOXP4 in the brain, based on the developmental human RNA sequencing data set of BrainSpan (http://www.brainspan.org/). **C**) Expression patterns of FOXP2 in a selection of cortical regions. These regions were selected based on structural MRI studies with KE family members carrying a FOXP2 mutation[13; 175; 176]: grey matter differences were found in the cortical motor-related areas, the inferior frontal gyrus and the superior temporal gyrus, among other regions. While the expression in the primary motor cortex (M1C) and the primary sensory cortex (S1C) peaks during development, the expression of FOXP2 in the superior temporal cortex (STC) and the ventromedial prefrontal cortex (VFC) seems to be maintained during adulthood. **B-C**) Each individual dot represents a brain sample and the lines are loess curves fitted through the data points. The dashed vertical line represents time of birth. Abbreviations for the analysed brain regions are MFC: medial frontal cortex, OFC: orbitofrontal, DFC: dorsolateral prefrontal cortex, VFC: ventromedial prefrontal cortex, M1C: primary motor cortex, S1C: primary sensory cortex, IPC: inferior parietal cortex, A1C: primary auditory cortex, STC: superior temporal cortex, ITC: inferior temporal cortex, V1C: primary visual cortex, HIP: hippocampus, STR: striatum, DTH: dorsal thalamic nucleus, MD: mediodorsal thalamic nucleus, CB: cerebellum, CBC: cerebellar cortex. Pcw: post conception week, mos: months.

such as zebrafish and cell-based systems. These investigations have uncovered upstream regulators of its expression, downstream targets that it regulates, and protein interactions that modulate its functions. Here, we give an up-to-date overview of the molecular networks of FOXP2 in the brain, highlighting how this information promises to deliver novel insights into roles of the gene in cognition and behaviour.

Regulation of *FOXP2* expression

Although the specific spatiotemporal expression patterns of FOXP2 in the brain imply tight regulation, little is known about the upstream mechanisms involved. Only a few transcription factors have been shown to bind to the genomic locus and/or to directly regulate its expression.

Tbr1 activates *Foxp2* expression in the developing cortex

TBR1 is a neural transcription factor with high expression in deep layers of the cortex, where it promotes a layer-VI identity, largely via repression of layer-V genes[63; 64]. In adult mice, almost 70% of FOXP2-positive cells in layer VI express TBR1[44], and cell-based assays have demonstrated that TBR1, in complex with its co-regulator CASK, can activate *FOXP2* expression (Figure 2A)[65; 66]. Conditional deletion of *Tbr1* in layer-VI neurons of mice leads to reduced *Foxp2* expression in these neurons, which shift to a layer-V-like identity[66]. Although the role of FOXP2 in cortical lamination is limited, based on studies with cortical-specific knockout mice[43], the gene may be part of the regulatory programme involved in formation, maintenance and connectivity of corticothalamic neurons in layer-VI[67], under control of TBR1. People with heterozygous *FOXP2* disruptions have been reported to show subtle differences in grey matter density in several parts of the cortex[13], based on voxel-based morphometry of MRI scans, although it is not known whether this involves altered connectivity and/or function of layer VI neurons in those regions. Recurrent *de novo* mutations of *TBR1* have been linked to a neurodevelopmental syndrome involving intellectual disability and/or autism spectrum disorder, and sometimes language impairments (MIM #606053), suggesting some phenotypic overlaps with FOXP2-related disorder[68].

Regulation of *FOXP2* by the canonical WNT/β-catenin signalling pathway

The genomic region upstream of the *FOXP2* locus contains six highly conserved binding regions for TCF/LEF transcription factors[69; 70], regulatory proteins that are activated by canonical WNT/β-catenin signalling, and involved in proliferation and direction of cell fate[69]. Binding of WNT to its receptor, Frizzled, leads to inhibition of GSK3β and accumulation of β-catenin, which translocates to the nucleus and activates transcription via TCF/LEF transcription factors[71]. One such TCF/LEF transcription factor is LEF1. *FoxP2* and *Lef1* are co-expressed in the developing zebrafish brain, where knockdown of *Lef1* expression yields loss of *FoxP2* expression[69]. Chromatin immunoprecipitation (ChIP) against Lef1 showed enrichment of the predicted Tcf/Lef binding regions upstream of *FoxP2*, suggesting that Lef1 directly binds to these enhancers to activate *FoxP2* expression[69].

The *FOXP2* locus also includes multiple highly conserved binding sites for PAX6, a key regulator of central nervous system development[72]. Knockdown of *Pax6* in developing zebrafish embryos disrupts *FoxP2* expression, while for knockout mice lacking *Pax6*, expression of Foxp2 is absent in the dorsolateral telencephalon[72]. ChIP against Pax6 in zebrafish embryos showed enrichment of binding sites in the *FoxP2* locus, confirming it as a direct target[72]. In the developing neocortex, PAX6 is

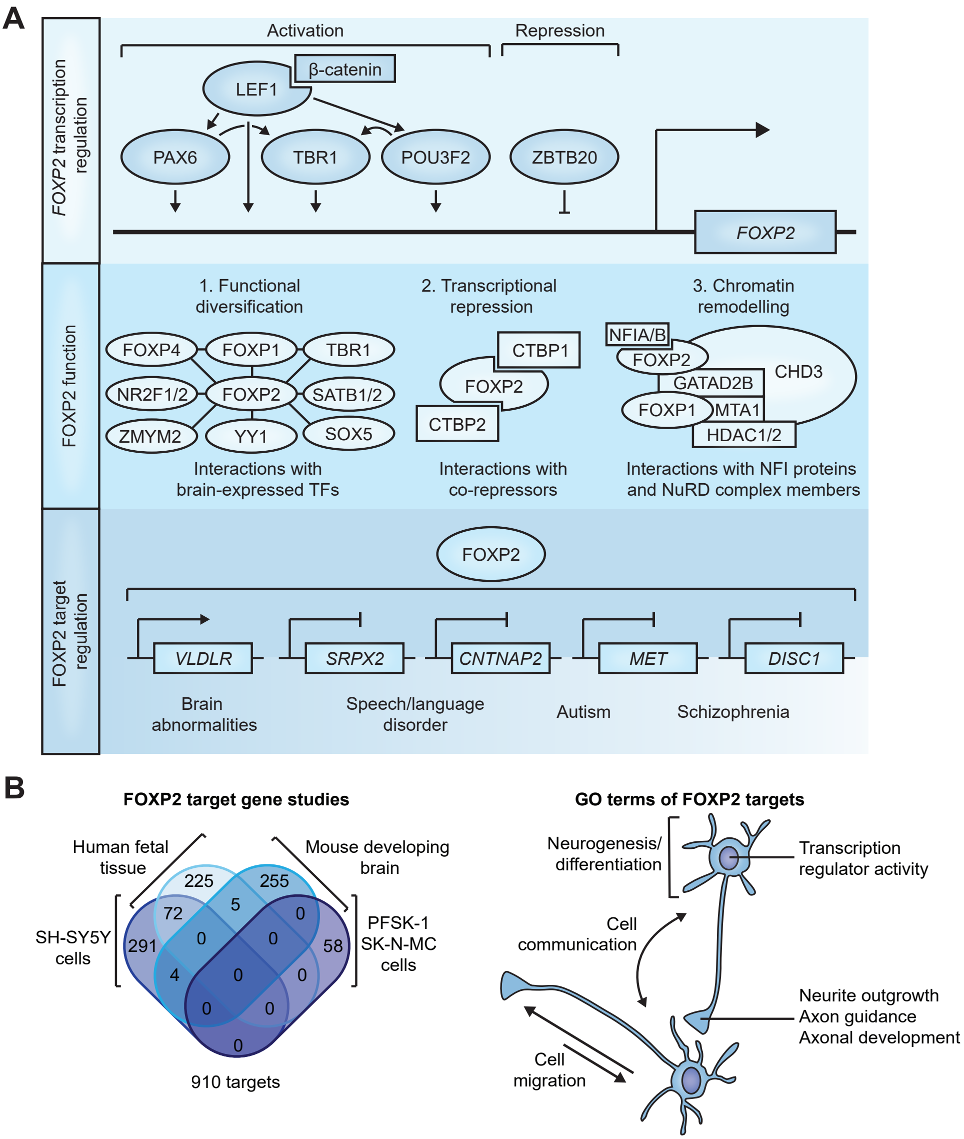


Figure 2. FOXP2 molecular networks. A) An overview of FOXP2 molecular networks in the brain, at the level of transcription regulation, function and target regulation. This overview represents results from a selection of separate studies using different types of model systems. TFs: transcription factors. **B**) Left, a Venn diagram showing the overlap between FOXP2 target genes identified in four FOXP2 ChIP-chip/seq studies. SH-SY5Y and SK-N-MC are human neuroblastoma cell lines and PFSK-1 is a neuroectodermal tumour cell line. Right, a schematic with a selection of gene ontology (GO) terms that are associated with the identified FOXP2 target genes.

expressed in neural progenitor cells in the ventricular zone, regulating the cell cycle and differentiation[73], while *FOXP2* is expressed at low levels in progenitor cells[33; 74] but at higher levels in neurons in the cortical plate[31; 33] and (as noted above) later in deep cortical layers[29]. Under control of WNT3, secreted by thalamic axons that grow into the developing neocortex, FOXP2 mRNA has been shown to be actively translated, driving differentiation of early neurons into deep layer neurons[75]. Activation of *FOXP2* by PAX6 might therefore be one of the steps that lead to differentiation of neural progenitor cells into neurons, fine-tuning their activity and connectivity.

The middle of the *FOXP2* locus contains an intronic regulatory element with a binding site for POU3F2, a well-known neural transcription factor[76]. This element drew the attention of molecular anthropologists studying the evolution of *FOXP2*, because the POU3F2-binding site contains a DNA variant that arose specifically on the human lineage after splitting from our common ancestor with Neanderthals/Denisovans. However, the site is not fixed in modern human populations; analysis of next-generation sequencing data from around the world shows that it remains polymorphic in southern Africa, casting doubt on the significance of this variant for human evolution[77] (see Ref. 78 for a recent account of how views of the relevance of *FOXP2* for human evolution have shifted with the availability of comprehensive genome-wide sequencing datasets and enhanced methods for assessing signals of selection). Based on reporter gene assays with the intronic enhancer it has been suggested that binding of POU3F2 to this site may lead to increased *FOXP2* expression[76], although this finding has not been confirmed in a more physiologically-relevant model and it is possible that the element instead regulates the expression of a different gene in the vicinity. Pou3f2 plays important roles in the formation and radial migration of upper layer cortical neurons[79; 80] and is known to drive expression of *Ngn2*, *Tbr2* and *Tbr1*, facilitating the differentiation of glutamatergic neurons[81].

PAX6 and POU3F2 are, like FOXP2, direct downstream targets of LEF1[82-84]. The LEF1-β-catenin/PAX6 signalling pathway is involved in self-renewal of neural progenitors and neurogenesis during neocortical development, initiating the PAX6/NGN2/TBR2/NEUROD/TBR1 cascade[82]. LEF1-β-catenin/POU3F2 signalling has been found to contribute to expansion of cortical neural progenitors and neurogenesis via the POU3F2/TBR2 and POU3F2/TBR1 cascades[81; 84]. We speculate that FOXP2 and its transcriptional regulators LEF1, PAX6 and POU3F2 may all be downstream effectors of WNT/β-catenin signalling (Figure 2A), a suggestion that could be tested in future with targeted experiments. Intriguingly, ectopic activation of Wnt signalling in the chicken optic cup has been shown to lead to upregulation of *FoxP2* expression[85].

FOXP2 has been reported to regulate the transcription of WNT pathway genes and to directly interact with β-catenin[86]. Moreover, the FOXP2-regulator TBR1 promotes maturation of layer-VI cortical neurons by enhancing WNT signalling[87]. As both an upstream and downstream player of this pathway, FOXP2 may potentially fulfil a central role in WNT/β-catenin signalling in the brain, a hypothesis that would be interesting to explore with future studies.

Zbtb20 represses Foxp2 expression in the hippocampus

To our knowledge, the only well-characterised repressor of *FOXP2* identified through animal models is ZBTB20 (Figure 2A)[88], a transcription factor expressed in hippocampal projection neurons, cerebellar granular cells, and gliogenic progenitors[89]. Zbtb20 was found to bind to and repress cortical layer marker genes, including *Foxp2*, in the developing mouse hippocampus, thereby directing a hippocampal fate while repressing other neuronal identities[88]. Consistently, transgenic expression of *Zbtb20* in mice results in reduced *Foxp2* expression[88]. Mouse Zbtb20 and human ZBTB20 proteins are highly conserved, with similar neural expression patterns[88], suggesting that the human orthologue may be important for *FOXP2* repression in the human hippocampus.

Downstream effectors of FOXP2

Multiple studies have sought downstream neural targets of FOXP2, yielding insights into pathways that it regulates in the context of brain development, function and disease.

FOXP2 targets are important for neurite outgrowth and cell migration

In early work on identifying targets of FOXP2, three studies performed ChIP-chip experiments on human foetal tissue[90], human neuroblastoma cells[91] and embryonic mouse brain tissue[47]. Although no identified targets were common to all three studies, they are enriched for genes associated with similar gene ontology categories, namely cell communication/migration and nervous system development including neurogenesis, neurite development and axon guidance[47; 90; 91] (Figure 2B). A ChIP-sequencing study of FOXP2 in neuroectodermal tumour cells and neuroblastoma cells, identified 58 targets near high-confidence ChIP peaks from a merged data set, that were mostly enriched for genes linked to transcriptional (regulatory) activity[92].

Follow-up experiments confirmed that Foxp2 promotes neurite outgrowth in both mouse neuroblastoma cells and mouse striatal primary neurons[47]. Indeed, genetic manipulations of *Foxp2* in an array of mouse models have been found to have effects on dendrite length. Specifically, introducing a partially humanised version of *Foxp2* into mice results in increased dendrite length of medium spiny neurons[45], while a loss-of-function mutation of the gene is reported to lead to decreased dendrite length of layer-VI excitatory neurons in the cortex[67]. The roles of *Foxp2* in neuronal migration are less clear-cut; although *in vitro* studies support effects of the gene on cell migration phenotypes[93], *in vivo* data from different mouse models are somewhat inconsistent with each other. For example, studies in which *Foxp2* expression was knocked down during embryonic development identified changes in cortical neurogenesis[74] and in migration of neural progenitors out of the subventricular zone[33], but selective deletion of the gene was not found to have such effects[43].

FOXP2 target genes are implicated in neurodevelopmental disorders

Out of the hundreds of putative targets of FOXP2, a small subset have received special attention through validation and follow-up in animal or cell-based models. One of the first targets to be studied in this way was *CNTNAP2*, which encodes CASPR2, a neurexin transmembrane protein expressed widely in the brain, with roles in nerve conduction, neuronal migration, neurite outgrowth and connectivity[95]. FOXP2 directly binds to regulatory regions of the *CNTNAP2* locus to repress expression[26; 96]. This is consistent with complementary expression patterns reported for the two genes in human foetal cerebral cortex[96] and increased *Cntnap2* expression in the cerebellum of a Foxp2-R552H mouse model (based on the human KE-family mutation)[97]. However, *CNTNAP2* expression changes temporally[98] and expression patterns of these genes could potentially show different relationships at distinct stages of development and/or in different brain regions. Interestingly, a cluster of SNPs in CNTNAP2 has been associated with reduced performance on a nonsense-word repetition task in a cohort of children with developmental language disorders[96] and with a measure of early communicative behaviour in a general population sample[99]. Furthermore, homozygous and compound heterozygous loss-of-function variants cause a severe neurodevelopmental disorder with epilepsy and intellectual disability (MIM #610042)[100-102]. Although in prior work both common and rare *CNTNAP2* variation has been linked to a range of brain-related phenotypes (Figure 2A), including autism (MIM #612100)[103; 104] and schizophrenia[105; 106], data from a recent large-scale study argue that the contributions of this gene to risk of these psychiatric disorders have been overstated[107].

Other genes that are repressed by FOXP2, and where links have been investigated in follow-up studies, include *SRPX2*[108], *MET*[109] and *DISC1*[90; 92; 110]. FOXP2 overexpression in cell-based assays lowers the expression of *SRPX2*[108], *MET*[109] and *DISC1*[110], and FOXP2 directly binds to regulatory sequences in *MET* and *SRPX2*[108; 109]. Cell-based assays additionally suggest that the FOXP2-R553H mutation disrupts regulation of *SRPX2* and *DISC1*[108; 110]. *SRPX2* variants have been identified in people with epilepsy of the rolandic speech area, speech apraxia, polymicrogyria and intellectual disability (MIM #300643)[108; 111; 112], although their aetiological relevance is uncertain given subsequent discovery of *GRIN2A* disruptions in the affected individuals[113]. Common variation in *MET* has been associated with autism spectrum disorder (MIM %611015)[114; 115] and schizophrenia[116], and post-mortem brain studies have shown altered *MET* expression in individuals with autism[117]. The *DISC1* gene has been linked to schizophrenia (MIM #604906)[118-120].

Beyond its effects as a transcriptional repressor, noted above, FOXP2 has been reported to be a direct activator of *VLDLR* expression[26; 90; 91; 121]. VLDLR is a receptor for RELN, expressed in the apical processes of migrating neurons in the developing cortex, with roles in neuronal migration, dendrite and spine development, and synaptic function[122]. Studies of zebra finch brain have found that FoxP2 protein directly binds to regulatory sequences of the *Vldlr* locus and that knockdown of the former reduces expression of the latter[121]. Homozygous disruptions of the human *VLDLR* gene have been discovered in patients with cerebellar hypoplasia, mild cerebral gyral simplification and intellectual disability (MIM #224050)[123-125].

Based on data thus far collected on downstream pathways, FOXP2 and its targets belong to molecular networks that are crucial for aspects of brain function, and that are implicated in a range of neurodevelopmental disorders with partially overlapping phenotypes, raising the possibility that aetiological variants of these genes affect shared mechanisms (Figure 2A).

FOXP2 transcriptional regulation

Although studies of FOXP2 have probed its expression patterns, regulation and transcriptional targets, the molecular mechanisms by which this regulatory protein acts as a transcription factor have been much less explored.

FOXP2 interacts with the CTBP transcriptional co-repressors

The first proteins to be identified as putative binding partners of FOXP2 were CTBP1 and CTBP2[24], transcriptional co-repressors that also interact with FOXP1 via a consensus binding site, which is lacking in FOXP4[24; 126]. *Drosophila* CtBP enhances repression by directly blocking the transcription initiation complex or inhibiting adjacent transcriptional activators[127]. Moreover, CTBP1 and CTBP2 were identified in a core protein complex that contained DNA-binding proteins, histone-modifying enzymes, histone methyltransferases and chromodomain-containing proteins[128], and may thereby aid FOXP2 in its transcriptional repressive functions (Figure 2A). Indeed, in cell-based assays, CTBP1 is able to increase FOXP1 and FOXP2 repression of reporter constructs[24]. The FOXP2-R553H protein, which harbours an aetiological substitution disrupting the DNA-binding domain[129], retains its ability to bind to CTBP1 and CTBP2, suggesting that DNA-binding of FOXP2 is not essential for the CTBP-FOXP2 interaction[126]. Since CTBP proteins depend on their interaction partners to be recruited to DNA, and FOXP2-R553H is unable to bind to DNA, it is unlikely that this complex represses target genes.

SUMOylation of FOXP2 modulates its function

Post-translational modifications are another way to dynamically regulate transcription factor activity. One such modification is SUMOylation, the reversible coupling of small ubiquitin-like modifiers (SUMOs), which are ubiquitously-expressed polypeptides, to specific sites in proteins. FOXP2 has a SUMOylation site at position K674, which is SUMOylated by SUMO1/2/3 via interaction with PIAS1/3[130; 131]. K674 SUMOylation is not critical for FOXP2 protein stability, dimerisation and subcellular localisation in human cell-lines[130; 132], but may alter its transcriptional activity[132]. Although one study did not detect changes in transcriptional repression of a non-SUMOylated FOXP2 K674R mutant[130], another found this mutant to be less effective in repressing target promoters compared to wild-type protein[132]. Disrupting the equivalent SUMOylation site in FOXP1 (K670) abolishes FOXP1 repression, while K670 SUMOylation in wild-type FOXP1 enhances binding to the CTBP1 co-repressor[133]. Studies of mice suggest that FOXP2 SUMOylation in the cerebellum is important for Purkinje cell development and motor functions[131]. In cell-based studies, ubiquitination, another form of post-translational modification, has been found for an alternatively-spliced short isoform of unknown significance (FOXP2.10+), while the canonical isoform was not ubiquitinated[129]. Whether other post-translational modifications beyond SUMOylation and ubiquitination, such as phosphorylation and acetylation, significantly contribute to regulation of FOXP2 functions has yet to be elucidated.

FOXP2 interacts with other brain-expressed transcription factors

A mass spectrometry study to characterise the FOXP2 interactome identified multiple transcription factors binding to FOXP2 in HEK293 cells, including NR2F1, NR2F2, SATB1, SATB2, SOX5, YY1 and ZMYM2[134]. Foxp2 is co-expressed with Sox5, Satb1, Satb2 and Nr2f1 in a subset of neurons in the mouse cerebral cortex, and with Nr2f2 in Purkinje cells[134]. The interactions were validated in cell-lines using bioluminescence resonance energy transfer (BRET) assays[134]. Additionally, the cortical transcription factor TBR1 was identified as a putative FOXP2 interactor in a yeast-two-hybrid assay[135], and confirmed with BRET[68]. The aetiological FOXP2 p.R553H mutation disrupts the interactions with these brain-expressed transcription factors[68; 134]. The functional importance of these interactions for *in vivo* brain development have not yet been studied, but may contribute to diversification of FOXP2 activity, guiding the protein to specific transcriptional complexes, changing its affinity for certain targets, and/or helping to recruit transcriptional co-factors (Figure 2A).

FOXP2 regulatory activity may be mediated via the NuRD chromatin remodelling complex

FOXP1, FOXP2 and FOXP4 all interact with the nucleosome remodelling and histone deacetylase (NuRD) complex[136], a multiprotein complex that couples two independent chromatin-regulatory functions, (1) ATP-dependent histone remodelling and (2) histone deacetylation[137; 138]. The complex, involved in both activation and repression of genes[139], is the most abundant form of deacetylase in mammals[140] and is linked to fundamental biological processes, including cell cycle progression, genomic integrity[141] and differentiation of embryonic stem cells[139; 140]. FOXP1 interacts with NuRD complex members HDAC1/2, GATAD2B and MTA1[136], FOXP4 with HDAC1/2 and GATAD2B[136], and FOXP2 with GATAD2B[136] and CHD3[130]. For FOXP1 and FOXP4 these interactions further reduce target gene expression in cell-based reporter assays, suggesting that these NuRD complex interaction partners act as co-repressors. For the FOXP2-GATAD2B interaction however, assays found no evidence of synergistic repression[136].

Interestingly, the NuRD complex plays an important role in the developing brain, apparent from the links of multiple of the core NuRD complex members with neurodevelopmental disorders that are characterised by features that partly overlap with the FOXP2-associated phenotypes. Mutations in the CHD4 gene result in an intellectual disability syndrome that includes global developmental delay and in some cases macrocephaly (MIM #617159)[142; 143]. A mutation in CHD3 was first discovered in a child with childhood apraxia of speech[144], whereafter additional aetiological variants were found in a number of patients that displayed intellectual disability, accompanied by speech/language problems and brain abnormalities including both macrocephaly and microcephaly (MIM #618205)[145]. Furthermore, GATAD2B disruptions have been identified in patients with intellectual disability and limited speech (MIM #615074)[146-148].

In addition to the direct interactions of FOXPs with NuRD complex members, there are multiple indirect links. FOXP2 and the HDAC1/2 proteins share at least three common interaction partners, the cortical transcription factors YY1[134; 149; 150], SATB1[134; 151] and SATB2[134; 152]. In layer-IV neurons of the cortex, Satb2 has been shown to assemble the NuRD complex upstream of Bcl11b, resulting in Bcl11b repression, via the Satb2-Hdac1 interaction[153]. Repression of BCL11B in SATB2 positive neurons is an essential mechanism in cortical lamination, resulting in upper-layer neuron specification[153]. In humans, YY1 (MIM #617557), SATB1 (MIM # 619228 and #619229) and SATB2 (MIM #612313) are all implicated in neurodevelopmental disorders[154-156]. Notably, SATB2 mutations cause severe language impairments[157]. Furthermore, CTBP2, a direct FOXP2 interactor and co-repressor[126], interacts with several NuRD complex members, namely HDAC2, MTA2, GATAD2B and CHD4[158]. Whether these FOXP2 interactors interact with FOXP2 and the NuRD complex simultaneously has not been studied.

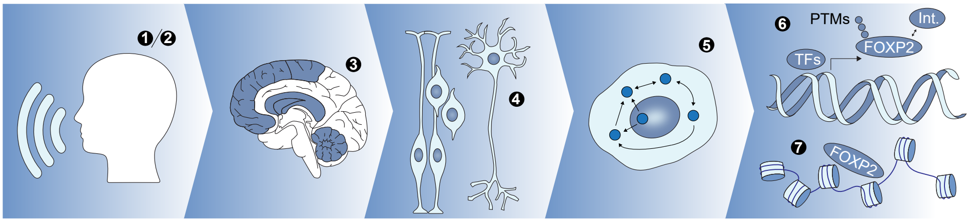
Most FOXP-NuRD complex interactions have only been characterised in cell-lines, or in the context of lung function (another tissue where FOXP proteins are expressed)[136], and the importance of such interactions for brain development remains to be uncovered. The NuRD complex plays major roles in the proliferation, migration and differentiation of neurons[159] and interactions with cortical transcription factors, such as SATB2, seem to recruit it to specific targets[153]. Hence, the FOXP proteins (as homo/heterodimers or together with other co-factors) may guide the NuRD complex to the DNA, to repress or activate target sequences via chromatin remodelling (Figure 2A). FOXP2 mutations may disrupt this mechanism by abolishing either DNA binding or interaction with NuRD complex members, resulting in abnormal regulation of downstream targets. Mutations in NuRD complex members may result in similar transcriptional regulatory defects, contributing to partial overlaps in the neurodevelopmental phenotypes that are associated with FOXP2, GATAD2B and SATB2 mutations.

In addition to potential chromatin remodelling functions via interactions with the NuRD complex, FOXP2 has been reported to mediate chromatin accessibility by interacting with transcriptional cofactors NFIA and NFIB in neuronal cell-based models[160]. Direct interactions of FOXP2 with DNA were found to yield repression of proliferation-promoting genes, while FOXP2-NFI complexes activated expression of genes driving neuronal differentiation via chromatin alterations[160]. Although FOXP2-R553H in complex with NFIA was still able to open chromatin, it did not activate gene expression. Thus, these data suggest the existence of distinct FOXP2 regulatory modes that together mediate target gene expression.

Future perspectives

Two decades of molecular studies on the functions of FOXP2 have shown that it belongs to an extensive molecular network with brain-expressed transcription factors and co-regulators, mediating neuronal differentiation, neurite outgrowth and cell migration in human cell-based assays, and shaping the development, plasticity and maturation of corticostriatal and corticocerebellar circuits important for behavioural phenotypes in animal models. Despite the attention FOXP2 has received over the years, much remains to be learned regarding its regulatory capabilities, position in molecular pathways, roles in cellular functions, and ultimately its effects on brain development and human speech and language capacities (Figure 3).

New and more sophisticated models may hold special promise for furthering our understanding of FOXP2 functions, particularly in light of links to speech and language. Human brain organoids grown from stem cells can model early stages of development of various parts of the nervous system[161; 162], and overcome species-specific developmental programmes[163], providing the opportunity to study the human transcriptome during brain development. Long-term[164] and slice cultures[165; 166] of these brain organoids result in maturation up to late foetal and early postnatal stages, while merging of region-specific organoids make it possible to model early establishment of brain circuitries[167; 168]. Genetic manipulation of *FOXP2* in such model systems could reveal human-specific functions that have been unable to be studied in traditional *in*



❶ Why do *FOXP2* variants have disproportionate effects on speech/language skills in humans, as compared to other cognitive functions?
❷ What explains similarities and differences between phenotypes associated with FOXP1, FOXP2 and FOXP4 dysfunction?
❸ How, and at which stages, is FOXP2 crucial in brain development?
❹ Which aspects of cell differentiation and function are dependent on FOXP2?
❺ And which molecular pathways are mediated by FOXP2?
❻ How is FOXP2 expression and function modulated at the molecular level in diverse tissues and at different developmental time points?
❼ How does FOXP2 control chromatin remodelling and target gene expression?
Can newly available tools help to dissect out downstream pathways?

Figure 3. Open questions on the molecular aspects of FOXP2 in the brain. Schematic with different levels of FOXP2 functioning. For each level questions are included that have remained largely unanswered, and should be focus of future studies. The brain areas in the schematic of Question 3, shaded in blue, represent regions of expression of FOXP2 that have been main focus in current literature. TFs: transcription factors, Int.: protein-interactors, PTMs: post-translational modifications.

vitro settings so far.

For studying FOXP2 functions *in vivo*, more relevant and non-traditional animal models are also being explored[169]. In addition to zebra finches, other species of birds display auditory-guided vocal learning[170], as well as bats[171; 172] and ocean mammals[173]. The latter two are evolutionarily closer to us, with brain structures and circuitries more similar to human brains. Indeed, analyses of FoxP expression patterns in the brains of bat species are already proving informative[174]. Although the genetic tools in such species are not yet as well-established as in the traditional animal models, optimisation and validation of these in the coming years will open up exciting new avenues for investigations of FOXP2 and its orthologues, placing the critical molecular networks in their broader evolutionary context.

Acknowledgements

We would like to thank dr. Cleo J.L.M. Smeets (L'Institut du Cerveau et de la Moelle Épinière, Paris Brain Institute) for her suggestions and feedback and dr. Else Eising (Max Planck Institute for Psycholinguistics) for help with analysing BrainSpan data. This work was supported by the Max Planck Society.

References

1. Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 519-523.
2. Reuter, M.S., Riess, A., Moog, U., Briggs, T.A., Chandler, K.E., Rauch, A., Stampfer, M., Steindl, K., Glaser, D., Joset, P., et al. (2017). FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. Journal of medical genetics 54, 64-72.
3. MacDermot, K.D., Bonora, E., Sykes, N., Coupe, A.M., Lai, C.S., Vernes, S.C., Vargha-Khadem, F., McKenzie, F., Smith, R.L., Monaco, A.P., et al. (2005). Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. American journal of human

genetics 76, 1074-1080.
4. Feuk, L., Kalervo, A., Lipsanen-Nyman, M., Skaug, J., Nakabayashi, K., Finucane, B., Hartung, D., Innes, M., Kerem, B., Nowaczyk, M.J., et al. (2006). Absence of a Paternally Inherited FOXP2 Gene in Developmental Verbal Dyspraxia. American journal of human genetics 79, 965-972.
5. Morgan, A., Fisher, S., Scheffer, I., and Hildebrand, M. (2016). FOXP2 Related Speech and Language Disorders. Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJH et al, editors Gene reviews(R) Seattle (WA): University of Washington.
6. Spaniel, F., Horacek, J., Tintera, J., Ibrahim, I., Novak, T., Cermak, J., Klirova, M., and Hoschl, C. (2011). Genetic variation in FOXP2 alters grey matter concentrations in schizophrenia patients. Neuroscience letters 493, 131-135.
7. Li, T., Zeng, Z., Zhao, Q., Wang, T., Huang, K., Li, J., Li, Y., Liu, J., Wei, Z., Wang, Y., et al. (2013). FoxP2 is significantly associated with schizophrenia and major depression in the Chinese Han population. The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry 14, 146-150.
8. Rao, W., Du, X., Zhang, Y., Yu, Q., Hui, L., Yu, Y., Kou, C., Yin, G., Zhu, X., Man, L., et al. (2017). Association between forkhead-box P2 gene polymorphism and clinical symptoms in chronic schizophrenia in a Chinese population. Journal of neural transmission (Vienna, Austria : 1996) 124, 891-897.
9. Yin, J., Jia, N., Liu, Y., Jin, C., Zhang, F., Yu, S., Wang, J., and Yuan, J. (2018).No association between FOXP2 rs10447760 and schizophrenia in a replication study of the Chinese Han population. Psychiatric genetics 28, 19-23.
10. Demontis, D., Walters, R.K., Martin, J., Mattheisen, M., Als, T.D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Baekvad-Hansen, M., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nature genetics 51, 63-75.
11. Clifton, E.A.D., Perry, J.R.B., Imamura, F., Lotta, L.A., Brage, S., Forouhi, N.G., Griffin, S.J., Wareham, N.J., Ong, K.K., and Day, F.R. (2018). Genome–wide association study for risk taking propensity indicates shared pathways with body mass index. Communications Biology 1, 36.
12. Liégeois, F., Baldeweg, T., Connelly, A., Gadian, D.G., Mishkin, M., and Vargha-Khadem, F. (2003). Language fMRI abnormalities associated with FOXP2 gene mutation. Nature neuroscience 6, 1230-1237.
13. Watkins, K.E., Vargha-Khadem, F., Ashburner, J., Passingham, R.E., Connelly, A., Friston, K.J., Frackowiak, R.S., Mishkin, M., and Gadian, D.G. (2002). MRI analysis of an inherited speech and language disorder: structural brain abnormalities. Brain : a journal of neurology 125, 465-478.
14. Liégeois, F.J., Hildebrand, M.S., Bonthrone, A., Turner, S.J., Scheffer, I.E., Bahlo, M., Connelly, A., and Morgan, A.T. (2016). Early neuroimaging markers of FOXP2 intragenic deletion. Scientific reports 6, 35192.
15. Argyropoulos, G.P.D., Watkins, K.E., Belton-Pagnamenta, E., Liégeois, F., Saleem, K.S., Mishkin, M., and Vargha-Khadem, F. (2019). Neocerebellar Crus I Abnormalities Associated with a Speech and Language Disorder Due to a Mutation in FOXP2. The Cerebellum 18, 309-319.
16. Uddén, J., Hultén, A., Bendtz, K., Mineroff, Z., Kucera, K.S., Vino, A., Fedorenko, E., Hagoort, P., and Fisher, S.E. (2019). Toward Robust Functional Neuroimaging Genetics of Cognition. The Journal of Neuroscience 39, 8778-8787.
17. Hoogman, M., Guadalupe, T., Zwiers, M.P., Klarenbeek, P., Francks, C., and Fisher, S.E. (2014). Assessing the effects of common variation in the FOXP2 gene on human brain structure. Frontiers in Human Neuroscience 8, 473.
18. Weigel, D., and Jackle, H. (1990). The fork head domain: a novel DNA binding motif of eukaryotic transcription factors? Cell 63, 455-456.
19. Hannenhalli, S., and Kaestner, K.H. (2009). The evolution of Fox genes and their role in development and disease. Nature reviews Genetics 10, 233-240.
20. Kaestner, K.H., Knochel, W., and Martinez, D.E. (2000). Unified nomenclature for the winged helix/forkhead transcription factors. Genes & development 14, 142-146.
21. Zhang, W., Duan, N., Song, T., Li, Z., Zhang, C., and Chen, X. (2017). The Emerging Roles of Forkhead Box (FOX) Proteins in Osteosarcoma. Journal of Cancer 8, 1619-1628.
22. Benayoun, B.A., Caburet, S., and Veitia, R.A. (2011). Forkhead transcription factors: key players in health and disease. Trends in Genetics 27, 224-232.
23. Shu, W., Yang, H., Zhang, L., Lu, M.M., and Morrisey, E.E. (2001). Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as

transcriptional repressors. The Journal of biological chemistry 276, 27488-27497.
24. Li, S., Weidenfeld, J., and Morrisey, E.E. (2004). Transcriptional and DNA Binding Activity of the Foxp1/2/4 Family Is Modulated by Heterotypic and Homotypic Protein Interactions. Molecular and Cellular Biology 24, 809-822.
25. Wang, B., Lin, D., Li, C., and Tucker, P. (2003). Multiple domains define the expression and regulatory properties of Foxp1 forkhead transcriptional repressors. The Journal of biological chemistry 278, 24259-24268.
26. Mendoza, E., and Scharff, C. (2017). Protein-Protein Interaction Among the FoxP Family Members and their Regulation of Two Target Genes, VLDLR and CNTNAP2 in the Zebra Finch Song System. Frontiers in molecular neuroscience 10, 112.
27. Fontenot, J.D., Gavin, M.A., and Rudensky, A.Y. (2003). Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. Nature immunology 4, 330-336.
28. Co, M., Anderson, A.G., and Konopka, G. (2020). FOXP transcription factors in vertebrate brain development, function, and disorders. WIREs Developmental Biology 9, e375.
29. Hisaoka, T., Nakamura, Y., Senba, E., and Morikawa, Y. (2010). The forkhead transcription factors, Foxp1 and Foxp2, identify different subpopulations of projection neurons in the mouse cerebral cortex. Neuroscience 166, 551-563.
30. Ferland, R.J., Cherry, T.J., Preware, P.O., Morrisey, E.E., and Walsh, C.A. (2003). Characterization of Foxp2 and Foxp1 mRNA and protein in the developing and mature brain. The Journal of comparative neurology 460, 266-279.
31. Lai, C.S., Gerrelli, D., Monaco, A.P., Fisher, S.E., and Copp, A.J. (2003). FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. Brain : a journal of neurology 126, 2455-2462.
32. Campbell, P., Reep, R.L., Stoll, M.L., Ophir, A.G., and Phelps, S.M. (2009). Conservation and diversity of Foxp2 expression in muroid rodents: functional implications. The Journal of comparative neurology 512, 84-100.
33. Garcia-Calero, E., Botella-Lopez, A., Bahamonde, O., Perez-Balaguer, A., and Martinez, S. (2016). FoxP2 protein levels regulate cell morphology changes and migration patterns in the vertebrate developing telencephalon. Brain structure & function 221, 2905-2917.
34. Takahashi, K., Liu, F.C., Hirokawa, K., and Takahashi, H. (2008). Expression of Foxp4 in the developing and adult rat forebrain. Journal of neuroscience research 86, 3106-3116.
35. Tam, W.Y., Leung, C.K., Tong, K.K., and Kwan, K.M. (2011). Foxp4 is essential in maintenance of Purkinje cell dendritic arborization in the mouse cerebellum. Neuroscience 172, 562-571.
36. Shu, W., Cho, J.Y., Jiang, Y., Zhang, M., Weisz, D., Elder, G.A., Schmeidler, J., De Gasperi, R., Sosa, M.A., Rabidou, D., et al. (2005). Altered ultrasonic vocalization in mice with a disruption in the Foxp2 gene. Proceedings of the National Academy of Sciences of the United States of America 102, 9643-9648.
37. Castellucci, G.A., McGinley, M.J., and McCormick, D.A. (2016). Knockout of Foxp2 disrupts vocal development in mice. Scientific reports 6, 23305.
38. Groszer, M., Keays, D.A., Deacon, R.M., de Bono, J.P., Prasad-Mulcare, S., Gaub, S., Baum, M.G., French, C.A., Nicod, J., Coventry, J.A., et al. (2008). Impaired synaptic plasticity and motor learning in mice with a point mutation implicated in human speech deficits. Current biology: CB 18, 354-362.
39. Chabout, J., Sarkar, A., Patel, S.R., Radden, T., Dunson, D.B., Fisher, S.E., and Jarvis, E.D. (2016). A Foxp2 Mutation Implicated in Human Speech Deficits Alters Sequencing of Ultrasonic Vocalizations in Adult Male Mice. Frontiers in behavioral neuroscience 10, 197.
40. French, C.A., Vinueza Veloz, M.F., Zhou, K., Peter, S., Fisher, S.E., Costa, R.M., and De Zeeuw, C.I. (2019). Differential effects of Foxp2 disruption in distinct motor circuits. Mol Psychiatry 24, 447-462.
41. Urbanus, B.H.A., Peter, S., Fisher, S.E., and De Zeeuw, C.I. (2020). Region-specific Foxp2 deletions in cortex, striatum or cerebellum cannot explain vocalization deficits observed in spontaneous global knockouts. Scientific reports 10, 21631.
42. Co, M., Hickey, S.L., Kulkarni, A., Harper, M., and Konopka, G. (2019). Cortical Foxp2 Supports Behavioral Flexibility and Developmental Dopamine D1 Receptor Expression. Cerebral Cortex 30, 1855-1870.
43. Kast, R.J., Lanjewar, A.L., Smith, C.D., and Levitt, P. (2019). FOXP2 exhibits projection neuron class specific expression, but is not required for multiple aspects of cortical histogenesis. eLife 8, e42012.
44. Medvedeva, V.P., Rieger, M.A., Vieth, B., Mombereau, C., Ziegenhain, C., Ghosh, T., Cressant, A., Enard, W., Granon, S., Dougherty, J.D., et al. (2019). Altered social behavior in mice carrying

a cortical Foxp2 deletion. Human molecular genetics 28, 701-717.
45. Enard, W., Gehre, S., Hammerschmidt, K., Holter, S.M., Blass, T., Somel, M., Bruckner, M.K., Schreiweis, C., Winter, C., Sohr, R., et al. (2009). A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell 137, 961-971.
46. Chen, Y.C., Kuo, H.Y., Bornschein, U., Takahashi, H., Chen, S.Y., Lu, K.M., Yang, H.Y., Chen, G.M., Lin, J.R., Lee, Y.H., et al. (2016). Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by opposing Mef2c. Nature neuroscience 19, 1513-1522.
47. Vernes, S.C., Oliver, P.L., Spiteri, E., Lockstone, H.E., Puliyadi, R., Taylor, J.M., Ho, J., Mombereau, C., Brewer, A., Lowy, E., et al. (2011). Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain. PLoS genetics 7, e1002145.
48. French, C.A., Jin, X., Campbell, T.G., Gerfen, E., Groszer, M., Fisher, S.E., and Costa, R.M. (2012). An aetiological Foxp2 mutation causes aberrant striatal activity and alters plasticity during skill learning. Molecular psychiatry 17, 1077-1085.
49. van Rhijn, J.-R., Fisher, S.E., Vernes, S.C., and Nadif Kasri, N. (2018). Foxp2 loss of function increases striatal direct pathway inhibition via increased GABA release. Brain structure & function 223, 4211-4226.
50. Hachigian, L.J., Carmona, V., Fenster, R.J., Kulicke, R., Heilbut, A., Sittler, A., Pereira de Almeida, L., Mesirov, J.P., Gao, F., Kolaczyk, E.D., et al. (2017). Control of Huntington's Disease-Associated Phenotypes by the Striatum-Enriched Transcription Factor Foxp2. Cell reports 21, 2688-2695.
51. Haesler, S., Wada, K., Nshdejan, A., Morrisey, E.E., Lints, T., Jarvis, E.D., and Scharff, C. (2004). FoxP2 expression in avian vocal learners and non-learners. The Journal of neuroscience : the official journal of the Society for Neuroscience 24, 3164-3175.
52. Heston, J.B., and White, S.A. (2015). Behavior-linked FoxP2 regulation enables zebra finch vocal learning. The Journal of neuroscience : the official journal of the Society for Neuroscience 35, 2885-2894.
53. Day, N.F., Hobbs, T.G., Heston, J.B., and White, S.A. (2019). Beyond Critical Period Learning: Striatal FoxP2 Affects the Active Maintenance of Learned Vocalizations in Adulthood. eNeuro 6.
54. Norton, P., Barschke, P., Scharff, C., and Mendoza, E. (2019). Differential Song Deficits after Lentivirus-Mediated Knockdown of FoxP1, FoxP2, or FoxP4 in Area X of Juvenile Zebra Finches. The Journal of Neuroscience 39, 9782-9796.
55. Xiao, L., Merullo, D.P., Koch, T.M.I., Cao, M., Co, M., Kulkarni, A., Konopka, G., and Roberts, T.F. (2021). Expression of FoxP2 in the basal ganglia regulates vocal motor sequences in the adult songbird. Nature Communications 12, 2617.
56. Hamdan, F.F., Daoud, H., Rochefort, D., Piton, A., Gauthier, J., Langlois, M., Foomani, G., Dobrzeniecka, S., Krebs, M.O., Joober, R., et al. (2010). De novo mutations in FOXP1 in cases with intellectual disability, autism, and language impairment. American journal of human genetics 87, 671-678.
57. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics 43, 585-589.
58. Srivastava, S., Cohen, J.S., Vernon, H., Baranano, K., McClellan, R., Jamal, L., Naidu, S., and Fatemi, A. (2014). Clinical whole exome sequencing in child neurology practice. Annals of neurology 76, 473-483.
59. Lozano, R., Vino, A., Lozano, C., Fisher, S.E., and Deriziotis, P. (2015). A de novo FOXP1 variant in a patient with autism, intellectual disability and severe speech and language impairment. European journal of human genetics : EJHG 23, 1702-1707.
60. Sollis, E., Graham, S.A., Vino, A., Froehlich, H., Vreeburg, M., Dimitropoulou, D., Gilissen, C., Pfundt, R., Rappold, G.A., Brunner, H.G., et al. (2016). Identification and functional characterization of de novo FOXP1 variants provides novel insights into the etiology of neurodevelopmental disorder. Human molecular genetics 25, 546-557.
61. Snijders Blok, L., Vino, A., den Hoed, J., Underhill, H.R., Monteil, D., Li, H., Reynoso Santos, F.J., Chung, W.K., Amaral, M.D., Schnur, R.E., et al. (2021). Heterozygous variants that disturb the transcriptional repressor activity of FOXP4 cause a developmental disorder with speech/language delays and multiple congenital abnormalities. Genet Med 23, 534–542.
62. Sollis, E., Deriziotis, P., Saitsu, H., Miyake, N., Matsumoto, N., Hoffer, M.J.V., Ruivenkamp, C.A.L., Alders, M., Okamoto, N., Bijlsma, E.K., et al. (2017). Equivalent missense variant in the FOXP2 and FOXP1 transcription factors causes distinct neurodevelopmental disorders. Human mutation 38, 1542-1554.
63. McKenna, W.L., Betancourt, J., Larkin, K.A., Abrams, B., Guo, C., Rubenstein, J.L., and Chen,

B. (2011). Tbr1 and Fezf2 regulate alternate corticofugal neuronal identities during neocortical development. The Journal of neuroscience : the official journal of the Society for Neuroscience 31, 549-564.

64. Han, W., Kwan, K.Y., Shim, S., Lam, M.M., Shin, Y., Xu, X., Zhu, Y., Li, M., and Sestan, N. (2011). TBR1 directly represses Fezf2 to control the laminar origin and development of the corticospinal tract. Proceedings of the National Academy of Sciences of the United States of America 108, 3041-3046.
65. Becker, M., Devanna, P., Fisher, S.E., and Vernes, S.C. (2018). Mapping of Human FOXP2 Enhancers Reveals Complex Regulation. Frontiers in molecular neuroscience 11, 47.
66. Fazel Darbandi, S., Robinson Schwartz, S.E., Qi, Q., Catta-Preta, R., Pai, E.L.-L., Mandell, J.D., Everitt, A., Rubin, A., Krasnoff, R.A., Katzman, S., et al. (2018). Neonatal Tbr1 Dosage Controls Cortical Layer 6 Connectivity. Neuron 100, 831-845.e837.
67. Druart, M., Groszer, M., and Le Magueresse, C. (2020). An Etiological Foxp2 Mutation Impairs Neuronal Gain in Layer VI Cortico-Thalamic Cells through Increased GABAB/GIRK Signaling. The Journal of Neuroscience 40, 8543.
68. Deriziotis, P., O'Roak, B.J., Graham, S.A., Estruch, S.B., Dimitropoulou, D., Bernier, R.A., Gerdts, J., Shendure, J., Eichler, E.E., and Fisher, S.E. (2014). De novo TBR1 mutations in sporadic autism disrupt protein functions. Nature Communications 5, 4954.
69. Bonkowsky, J.L., Wang, X., Fujimoto, E., Lee, J.E., Chien, C.B., and Dorsky, R.I. (2008). Domain-specific regulation of foxP2 CNS expression by lef1. BMC developmental biology 8, 103.
70. Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124, 47-59.
71. Ciani, L., and Salinas, P.C. (2005). WNTS in the vertebrate nervous system: from patterning to neuronal connectivity. Nature Reviews Neuroscience 6, 351.
72. Coutinho, P., Pavlou, S., Bhatia, S., Chalmers, K.J., Kleinjan, D.A., and van Heyningen, V. (2011). Discovery and assessment of conserved Pax6 target genes and enhancers. Genome research 21, 1349-1359.
73. Gotz, M., Stoykova, A., and Gruss, P. (1998). Pax6 controls radial glia differentiation in the cerebral cortex. Neuron 21, 1031-1044.
74. Tsui, D., Vessey, J.P., Tomita, H., Kaplan, D.R., and Miller, F.D. (2013). FoxP2 regulates neurogenesis during embryonic cortical development. The Journal of neuroscience : the official journal of the Society for Neuroscience 33, 244-258.
75. Kraushar, M.L., Viljetic, B., Wijeratne, H.R.S., Thompson, K., Jiao, X., Pike, J.W., Medvedeva, V., Groszer, M., Kiledjian, M., Hart, R.P., et al. (2015). Thalamic WNT3 Secretion Spatiotemporally Regulates the Neocortical Ribosome Signature and mRNA Translation to Specify Neocortical Cell Subtypes. The Journal of neuroscience : the official journal of the Society for Neuroscience 35, 10911-10926.
76. Maricic, T., Gunther, V., Georgiev, O., Gehre, S., Curlin, M., Schreiweis, C., Naumann, R., Burbano, H.A., Meyer, M., Lalueza-Fox, C., et al. (2013). A recent evolutionary change affects a regulatory element in the human FOXP2 gene. Molecular biology and evolution 30, 844-852.
77. Atkinson, E.G., Audesse, A.J., Palacios, J.A., Bobo, D.M., Webb, A.E., Ramachandran, S., and Henn, B.M. (2018). No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. Cell 174, 1424-1435 e1415.
78. Fisher, S.E. (2019). Human Genetics: The Evolving Story of FOXP2. Current Biology 29, R65-R67.
79. McEvilly, R.J., de Diaz, M.O., Schonemann, M.D., Hooshmand, F., and Rosenfeld, M.G. (2002). Transcriptional regulation of cortical neuron migration by POU domain factors. Science 295, 1528-1532.
80. Sugitani, Y., Nakai, S., Minowa, O., Nishi, M., Jishage, K., Kawano, H., Mori, K., Ogawa, M., and Noda, T. (2002). Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons. Genes & development 16, 1760-1765.
81. Dominguez, M.H., Ayoub, A.E., and Rakic, P. (2013). POU-III Transcription Factors (Brn1, Brn2, and Oct6) Influence Neurogenesis, Molecular Identity, and Migratory Destination of Upper-Layer Cells of the Cerebral Cortex. Cerebral Cortex 23, 2632-2643.
82. Gan, Q., Lee, A., Suzuki, R., Yamagami, T., Stokes, A., Nguyen, B.C., Pleasure, D., Wang, J., Chen, H.W., and Zhou, C.J. (2014). Pax6 Mediates β-Catenin Signaling for Self-Renewal and Neurogenesis by Neocortical Radial Glial Stem Cells. Stem cells 32, 45-58.
83. Goodall, J., Martinozzi, S., Dexter, T.J., Champeval, D., Carreira, S., Larue, L., and Goding, C.R.

(2004). Brn-2 Expression Controls Melanoma Proliferation and Is Directly Regulated by β-Catenin. Molecular and Cellular Biology 24, 2915-2922.

84. Belinson, H., Nakatani, J., Babineau, B.A., Birnbaum, R.Y., Ellegood, J., Bershteyn, M., McEvilly, R.J., Long, J.M., Willert, K., Klein, O.D., et al. (2016). Prenatal β-catenin/Brn2/Tbr2 transcriptional cascade regulates adult social and stereotypic behaviors. Molecular psychiatry 21, 1417-1433.
85. Trimarchi, J.M., Cho, S.-H., and Cepko, C.L. (2009). Identification of genes expressed preferentially in the developing peripheral margin of the optic cup. Developmental dynamics : an official publication of the American Association of Anatomists 238, 2327-2329.
86. Richter, G., Gui, T., Bourgeois, B., Koyani, C.N., Ulz, P., Heitzer, E., von Lewinski, D., Burgering, B.M.T., Malle, E., and Madl, T. (2021). beta-catenin regulates FOXP2 transcriptional activity via multiple binding sites. The FEBS journal 288, 3261-3284.
87. Fazel Darbandi, S., Robinson Schwartz, S.E., Pai, E.L.-L., Everitt, A., Turner, M.L., Cheyette, B.N.R., Willsey, A.J., State, M.W., Sohal, V.S., and Rubenstein, J.L.R. (2020). Enhancing WNT Signaling Restores Cortical Neuronal Spine Maturation and Synaptogenesis in Tbr1 Mutants. Cell reports 31, 107495.
88. Nielsen, J.V., Thomassen, M., Mollgard, K., Noraberg, J., and Jensen, N.A. (2014). Zbtb20 defines a hippocampal neuronal identity through direct repression of genes that control projection neuron development in the isocortex. Cerebral cortex (New York, NY : 1991) 24, 1216-1229.
89. Mitchelmore, C., Kjaerulff, K.M., Pedersen, H.C., Nielsen, J.V., Rasmussen, T.E., Fisker, M.F., Finsen, B., Pedersen, K.M., and Jensen, N.A. (2002). Characterization of two novel nuclear BTB/POZ domain zinc finger isoforms. Association with differentiation of hippocampal neurons, cerebellar granule cells, and macroglia. The Journal of biological chemistry 277, 7598-7609.
90. Spiteri, E., Konopka, G., Coppola, G., Bomar, J., Oldham, M., Ou, J., Vernes, S.C., Fisher, S.E., Ren, B., and Geschwind, D.H. (2007). Identification of the transcriptional targets of FOXP2, a gene linked to speech and language, in developing human brain. American journal of human genetics 81, 1144-1157.
91. Vernes, S.C., Spiteri, E., Nicod, J., Groszer, M., Taylor, J.M., Davies, K.E., Geschwind, D.H., and Fisher, S.E. (2007). High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. American journal of human genetics 81, 1232-1250.
92. Nelson, C.S., Fuller, C.K., Fordyce, P.M., Greninger, A.L., Li, H., and DeRisi, J.L. (2013). Microfluidic affinity and ChIP-seq analyses converge on a conserved FOXP2-binding motif in chimp and human, which enables the detection of evolutionarily novel targets. Nucleic acids research 41, 5991-6004.
93. Devanna, P., Middelbeek, J., and Vernes, S.C. (2014). FOXP2 drives neuronal differentiation by interacting with retinoic acid signaling pathways. Frontiers in cellular neuroscience 8, 305.
94. Rhinn, M., and Dolle, P. (2012). Retinoic acid signalling during development. Development 139, 843-858.
95. Rodenas-Cuadrado, P., Ho, J., and Vernes, S.C. (2014). Shining a light on CNTNAP2: complex functions to complex disorders. European Journal of Human Genetics 22, 171-178.
96. Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcon, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., et al. (2008). A functional genetic link between distinct developmental language disorders. The New England journal of medicine 359, 2337-2345.
97. Fujita, E., Tanabe, Y., Momoi, M.Y., and Momoi, T. (2012). Cntnap2 expression in the cerebellum of Foxp2(R552H) mice, with a mutation related to speech-language disorder. Neuroscience letters 506, 277-280.
98. Gordon, A., Salomon, D., Barak, N., Pen, Y., Tsoory, M., Kimchi, T., and Peles, E. (2016). Expression of Cntnap2 (Caspr2) in multiple levels of sensory systems. Molecular and Cellular Neuroscience 70, 42-53.
99. Whitehouse, A.J.O., Bishop, D.V.M., Ang, Q.W., Pennell, C.E., and Fisher, S.E. (2011). CNTNAP2 variants affect early language development in the general population. Genes Brain Behav 10, 451-456.
100. Strauss, K.A., Puffenberger, E.G., Huentelman, M.J., Gottlieb, S., Dobrin, S.E., Parod, J.M., Stephan, D.A., and Morton, D.H. (2006). Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. The New England journal of medicine 354, 1370-1377.
101. Zweier, C., de Jong, E.K., Zweier, M., Orrico, A., Ousager, L.B., Collins, A.L., Bijlsma, E.K., Oortveld, M.A., Ekici, A.B., Reis, A., et al. (2009). CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. American journal of human genetics 85, 655-666.

102. Smogavec, M., Cleall, A., Hoyer, J., Lederer, D., Nassogne, M.C., Palmer, E.E., Deprez, M., Benoit, V., Maystadt, I., Noakes, C., et al. (2016). Eight further individuals with intellectual disability and epilepsy carrying bi-allelic CNTNAP2 aberrations allow delineation of the mutational and phenotypic spectrum. Journal of medical genetics 53, 820-827.
103. Alarcon, M., Abrahams, B.S., Stone, J.L., Duvall, J.A., Perederiy, J.V., Bomar, J.M., Sebat, J., Wigler, M., Martin, C.L., Ledbetter, D.H., et al. (2008). Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. American journal of human genetics 82, 150-159.
104. Arking, D.E., Cutler, D.J., Brune, C.W., Teslovich, T.M., West, K., Ikeda, M., Rea, A., Guy, M., Lin, S., Cook, E.H., et al. (2008). A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. American journal of human genetics 82, 160-164.
105. Friedman, J.I., Vrijenhoek, T., Markx, S., Janssen, I.M., van der Vliet, W.A., Faas, B.H., Knoers, N.V., Cahn, W., Kahn, R.S., Edelmann, L., et al. (2008). CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy. Mol Psychiatry 13, 261-266.
106. Ji, W., Li, T., Pan, Y., Tao, H., Ju, K., Wen, Z., Fu, Y., An, Z., Zhao, Q., Wang, T., et al. (2013). CNTNAP2 is significantly associated with schizophrenia and major depression in the Han Chinese population. Psychiatry research 207, 225-228.
107. Toma, C., Pierce, K.D., Shaw, A.D., Heath, A., Mitchell, P.B., Schofield, P.R., and Fullerton, J.M. (2018). Comprehensive cross-disorder analyses of CNTNAP2 suggest it is unlikely to be a primary risk gene for psychiatric disorders. PLoS genetics 14, e1007535-e1007535.
108. Roll, P., Vernes, S.C., Bruneau, N., Cillario, J., Ponsole-Lenfant, M., Massacrier, A., Rudolf, G., Khalife, M., Hirsch, E., Fisher, S.E., et al. (2010). Molecular networks implicated in speech-related disorders: FOXP2 regulates the SRPX2/uPAR complex. Human molecular genetics 19, 4848-4860.
109. Mukamel, Z., Konopka, G., Wexler, E., Osborn, G.E., Dong, H., Bergman, M.Y., Levitt, P., and Geschwind, D.H. (2011). Regulation of MET by FOXP2, genes implicated in higher cognitive dysfunction and autism risk. The Journal of neuroscience : the official journal of the Society for Neuroscience 31, 11437-11442.
110. Walker, R.M., Hill, A.E., Newman, A.C., Hamilton, G., Torrance, H.S., Anderson, S.M., Ogawa, F., Derizioti, P., Nicod, J., Vernes, S.C., et al. (2012). The DISC1 promoter: characterization and regulation by FOXP2. Human molecular genetics 21, 2862-2872.
111. Roll, P., Rudolf, G., Pereira, S., Royer, B., Scheffer, I.E., Massacrier, A., Valenti, M.P., Roeckel-Trevisiol, N., Jamali, S., Beclin, C., et al. (2006). SRPX2 mutations in disorders of language cortex and cognition. Human molecular genetics 15, 1195-1207.
112. Chen, X.S., Reader, R.H., Hoischen, A., Veltman, J.A., Simpson, N.H., Francks, C., Newbury, D.F., and Fisher, S.E. (2017). Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. Scientific reports 7, 46105.
113. Lesca, G., Rudolf, G., Bruneau, N., Lozovaya, N., Labalme, A., Boutry-Kryza, N., Salmi, M., Tsintsadze, T., Addis, L., Motte, J., et al. (2013). GRIN2A mutations in acquired epileptic aphasia and related childhood focal epilepsies and encephalopathies with speech and language dysfunction. Nature genetics 45, 1061-1066.
114. Campbell, D.B., Sutcliffe, J.S., Ebert, P.J., Militerni, R., Bravaccio, C., Trillo, S., Elia, M., Schneider, C., Melmed, R., Sacco, R., et al. (2006). A genetic variant that disrupts MET transcription is associated with autism. Proceedings of the National Academy of Sciences of the United States of America 103, 16834-16839.
115. Thanseem, I., Nakamura, K., Miyachi, T., Toyota, T., Yamada, S., Tsujii, M., Tsuchiya, K.J., Anitha, A., Iwayama, Y., Yamada, K., et al. (2010). Further evidence for the role of MET in autism susceptibility. Neuroscience research 68, 137-141.
116. Burdick, K.E., DeRosse, P., Kane, J.M., Lencz, T., and Malhotra, A.K. (2010). Association of genetic variation in the MET proto-oncogene with schizophrenia and general cognitive ability. The American journal of psychiatry 167, 436-443.
117. Campbell, D.B., D'Oronzio, R., Garbett, K., Ebert, P.J., Mirnics, K., Levitt, P., and Persico, A.M. (2007). Disruption of cerebral cortex MET signaling in autism spectrum disorder. Annals of neurology 62, 243-250.
118. Hennah, W., Varilo, T., Kestila, M., Paunio, T., Arajarvi, R., Haukka, J., Parker, A., Martin, R., Levitzky, S., Partonen, T., et al. (2003). Haplotype transmission analysis provides evidence of association for DISC1 to schizophrenia and suggests sex-dependent effects. Human molecular genetics 12, 3151-3159.
119. Hodgkinson, C.A., Goldman, D., Jaeger, J., Persaud, S., Kane, J.M., Lipsky, R.H., and Malhotra, A.K. (2004). Disrupted in schizophrenia 1 (DISC1): association with schizophrenia,

schizoaffective disorder, and bipolar disorder. American journal of human genetics 75, 862-872.
120. Schumacher, J., Laje, G., Abou Jamra, R., Becker, T., Muhleisen, T.W., Vasilescu, C., Mattheisen, M., Herms, S., Hoffmann, P., Hillmer, A.M., et al. (2009). The DISC locus and schizophrenia: evidence from an association study in a central European sample and from a meta-analysis across different European populations. Human molecular genetics 18, 2719-2727.
121. Adam, I., Mendoza, E., Kobalz, U., Wohlgemuth, S., and Scharff, C. (2016). FoxP2 directly regulates the reelin receptor VLDLR developmentally and by singing. Molecular and cellular neurosciences 74, 96-105.
122. Lee, G.H., and D'Arcangelo, G. (2016). New Insights into Reelin-Mediated Signaling Pathways. Frontiers in cellular neuroscience 10, 122.
123. Boycott, K.M., Flavelle, S., Bureau, A., Glass, H.C., Fujiwara, T.M., Wirrell, E., Davey, K., Chudley, A.E., Scott, J.N., McLeod, D.R., et al. (2005). Homozygous Deletion of the Very Low Density Lipoprotein Receptor Gene Causes Autosomal Recessive Cerebellar Hypoplasia with Cerebral Gyral Simplification. American journal of human genetics 77, 477-483.
124. Ozcelik, T., Akarsu, N., Uz, E., Caglayan, S., Gulsuner, S., Onat, O.E., Tan, M., and Tan, U. (2008). Mutations in the very low-density lipoprotein receptor VLDLR cause cerebellar hypoplasia and quadrupedal locomotion in humans. Proceedings of the National Academy of Sciences of the United States of America 105, 4232-4236.
125. Dixon-Salazar, T.J., Silhavy, J.L., Udpa, N., Schroth, J., Bielas, S., Schaffer, A.E., Olvera, J., Bafna, V., Zaki, M.S., Abdel-Salam, G.H., et al. (2012). Exome sequencing can improve diagnosis and alter patient management. Sci Transl Med 4, 138ra178.
126. Estruch, S.B., Graham, S.A., Chinnappa, S.M., Deriziotis, P., and Fisher, S.E. (2016). Functional characterization of rare FOXP2 variants in neurodevelopmental disorder. Journal of neurodevelopmental disorders 8, 44.
127. Nibu, Y., Senger, K., and Levine, M. (2003). CtBP-Independent Repression in the Drosophila Embryo. Molecular and Cellular Biology 23, 3990-3999.
128. Shi, Y., Sawada, J.-i., Sui, G., Affar, E.B., Whetstine, J.R., Lan, F., Ogawa, H., Po-Shan Luke, M., Nakatani, Y., and Shi, Y. (2003). Coordinated histone modifications mediated by a CtBP co-repressor complex. Nature 422, 735.
129. Vernes, S.C., Nicod, J., Elahi, F.M., Coventry, J.A., Kenny, N., Coupe, A.M., Bird, L.E., Davies, K.E., and Fisher, S.E. (2006). Functional genetic analysis of mutations implicated in a human speech and language disorder. Human molecular genetics 15, 3154-3167.
130. Estruch, S.B., Graham, S.A., Deriziotis, P., and Fisher, S.E. (2016). The language-related transcription factor FOXP2 is post-translationally modified with small ubiquitin-like modifiers. Scientific reports 6, 20911.
131. Usui, N., Co, M., Harper, M., Rieger, M.A., Dougherty, J.D., and Konopka, G. (2017). Sumoylation of FOXP2 Regulates Motor Function and Vocal Communication Through Purkinje Cell Development. Biological psychiatry 81, 220-230.
132. Meredith, L.J., Wang, C.M., Nascimento, L., Liu, R., Wang, L., and Yang, W.H. (2016). The Key Regulator for Language and Speech Development, FOXP2, is a Novel Substrate for SUMOylation. Journal of cellular biochemistry 117, 426-438.
133. Rocca, D.L., Wilkinson, K.A., and Henley, J.M. (2017). SUMOylation of FOXP1 regulates transcriptional repression via CtBP1 to drive dendritic morphogenesis. Scientific reports 7, 877.
134. Estruch, S.B., Graham, S.A., Quevedo, M., Vino, A., Dekkers, D.H.W., Deriziotis, P., Sollis, E., Demmers, J., Poot, R.A., and Fisher, S.E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Human molecular genetics 27, 1212-1227.
135. Sakai, Y., Shaw, C.A., Dawson, B.C., Dugas, D.V., Al-Mohtaseb, Z., Hill, D.E., and Zoghbi, H.Y. (2011). Protein interactome reveals converging molecular pathways among autism disorders. Science translational medicine 3, 86ra49-86ra49.
136. Chokas, A.L., Trivedi, C.M., Lu, M.M., Tucker, P.W., Li, S., Epstein, J.A., and Morrisey, E.E. (2010). Foxp1/2/4-NuRD interactions regulate gene expression and epithelial injury response in the lung via regulation of interleukin-6. The Journal of biological chemistry 285, 13304-13313.
137. Tong, J.K., Hassig, C.A., Schnitzler, G.R., Kingston, R.E., and Schreiber, S.L. (1998). Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. Nature 395, 917-921.
138. Xue, Y., Wong, J., Moreno, G.T., Young, M.K., Cote, J., and Wang, W. (1998). NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. Molecular cell 2, 851-861.
139. Basta, J., and Rauchman, M. (2015). The nucleosome remodeling and deacetylase complex in development and disease. Translational research: the journal of laboratory and clinical medicine

165, 36-47.
140. Torchy, M.P., Hamiche, A., and Klaholz, B.P. (2015). Structure and function insights into the NuRD chromatin remodeling complex. Cellular and Molecular Life Sciences 72, 2491-2507.
141. Lai, A.Y., and Wade, P.A. (2011). Cancer biology and NuRD: a multifaceted chromatin remodelling complex. Nature reviews Cancer 11, 588-596.
142. Sifrim, A., Hitz, M.P., Wilsdon, A., Breckpot, J., Turki, S.H., Thienpont, B., McRae, J., Fitzgerald, T.W., Singh, T., Swaminathan, G.J., et al. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nature genetics 48, 1060-1065.
143. Weiss, K., Terhal, P.A., Cohen, L., Bruccoleri, M., Irving, M., Martinez, A.F., Rosenfeld, J.A., Machol, K., Yang, Y., Liu, P., et al. (2016). De Novo Mutations in CHD4, an ATP-Dependent Chromatin Remodeler Gene, Cause an Intellectual Disability Syndrome with Distinctive Dysmorphisms. American journal of human genetics 99, 934-941.
144. Eising, E., Carrion-Castillo, A., Vino, A., Strand, E.A., Jakielski, K.J., Scerri, T.S., Hildebrand, M.S., Webster, R., Ma, A., Mazoyer, B., et al. (2019). A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. Mol Psychiatry 24, 1065-1078.
145. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nat Commun 9, 4619.
146. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. The New England journal of medicine 367, 1921-1929.
147. Willemsen, M.H., Nijhof, B., Fenckova, M., Nillesen, W.M., Bongers, E.M., Castells-Nobau, A., Asztalos, L., Viragh, E., van Bon, B.W., Tezel, E., et al. (2013). GATAD2B loss-of-function mutations cause a recognisable syndrome with intellectual disability and are associated with learning deficits and synaptic undergrowth in Drosophila. Journal of medical genetics 50, 507-514.
148. Shieh, C., Jones, N., Vanle, B., Au, M., Huang, A.Y., Silva, A.P.G., Lee, H., Douine, E.D., Otero, M.G., Choi, A., et al. (2020). GATAD2B-associatedneurodevelopmental disorder (GAND): clinical and molecular insights into a NuRD-related disorder. Genetics in Medicine 22, 878-888.
149. Yang, W.-M., Inouye, C., Zeng, Y., Bearss, D., and Seto, E. (1996). Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. Proceedings of the National Academy of Sciences 93, 12845-12850.
150. Yao, Y.L., Yang, W.M., and Seto, E. (2001). Regulation of transcription factor YY1 by acetylation and deacetylation. Mol Cell Biol 21, 5979-5991.
151. Yasui, D., Miyano, M., Cai, S., Varga-Weisz, P., and Kohwi-Shigematsu, T. (2002). SATB1 targets chromatin remodelling to regulate genes over long distances. Nature 419, 641-645.
152. Gyorgy, A.B., Szemes, M., De Juan Romero, C., Tarabykin, V., and Agoston, D.V. (2008). SATB2 interacts with chromatin-remodeling molecules in differentiating cortical neurons. European Journal of Neuroscience 27, 865-873.
153. Britanova, O., de Juan Romero, C., Cheung, A., Kwan, K.Y., Schwark, M., Gyorgy, A., Vogel, T., Akopov, S., Mitkovski, M., Agoston, D., et al. (2008). Satb2 Is a Postmitotic Determinant for Upper-Layer Neuron Specification in the Neocortex. Neuron 57, 378-392.
154. Gabriele, M., Vulto-van Silfhout, A.T., Germain, P.L., Vitriolo, A., Kumar, R., Douglas, E., Haan, E., Kosaki, K., Takenouchi, T., Rauch, A., et al. (2017). YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. American journal of human genetics 100, 907-925.
155. den Hoed, J., de Boer, E., Voisin, N., Dingemans, A.J.M., Guex, N., Wiel, L., Nellaker, C., Amudhavalli, S.M., Banka, S., Bena, F.S., et al. (2021). Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction. American journal of human genetics 108, 346-356.
156. Bengani, H., Handley, M., Alvi, M., Ibitoye, R., Lees, M., Lynch, S.A., Lam, W., Fannemel, M., Nordgren, A., Malmgren, H., et al. (2017). Clinical and molecular consequences of disease-associated de novo mutations in SATB2. Genetics in Medicine 19, 900-908.
157. Zarate, Y.A., and Fish, J.L. (2017). SATB2-associated syndrome: Mechanisms, phenotype, and practical recommendations. American journal of medical genetics Part A 173, 327-337.
158. Zhao, L.-J., Subramanian, T., Vijayalingam, S., and Chinnadurai, G. (2014). CtBP2 proteome:

Role of CtBP in E2F7-mediated repression and cell proliferation. Genes & Cancer 5, 31-40.

159. Nitarska, J., Smith, J.G., Sherlock, W.T., Hillege, M.M., Nott, A., Barshop, W.D., Vashisht, A.A., Wohlschlegel, J.A., Mitter, R., and Riccio, A. (2016). A Functional Switch of NuRD Chromatin Remodeling Complex Subunits Regulates Mouse Cortical Development. Cell reports 17, 1683-1698.

160. Hickey, S.L., Berto, S., and Konopka, G. (2019). Chromatin Decondensation by FOXP2 Promotes Human Neuron Maturation and Expression of Neurodevelopmental Disease Genes. Cell reports 27, 1699-1711.e1699.

161. Kelava, I., and Lancaster, M.A. (2016). Dishing out mini-brains: Current progress and future prospects in brain organoid research. Dev Biol 420, 199-209.

162. Marton, R.M., and Pașca, S.P. (2020). Organoid and Assembloid Technologies for Investigating Cellular Crosstalk in Human Brain Development and Disease. Trends in Cell Biology 30, 133-143.

163. Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., et al. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature 574, 418-422.

164. Gordon, A., Yoon, S.J., Tran, S.S., Makinson, C.D., Park, J.Y., Andersen, J., Valencia, A.M., Horvath, S., Xiao, X., Huguenard, J.R., et al. (2021). Long-term maturation of human cortical organoids matches key early postnatal transitions. Nature neuroscience 24, 331-342.

165. Qian, X., Su, Y., Adam, C.D., Deutschmann, A.U., Pather, S.R., Goldberg, E.M., Su, K., Li, S., Lu, L., Jacob, F., et al. (2020). Sliced Human Cortical Organoids for Modeling Distinct Cortical Layer Formation. Cell Stem Cell 26, 766-781.e769.

166. Giandomenico, S.L., Mierau, S.B., Gibbons, G.M., Wenger, L.M.D., Masullo, L., Sit, T., Sutcliffe, M., Boulanger, J., Tripodi, M., Derivery, E., et al. (2019). Cerebral organoids at the air-liquid interface generate diverse nerve tracts with functional output. Nature neuroscience 22, 669-679.

167. Miura, Y., Li, M.Y., Birey, F., Ikeda, K., Revah, O., Thete, M.V., Park, J.Y., Puno, A., Lee, S.H., Porteus, M.H., et al. (2020). Generation of human striatal organoids and cortico-striatal assembloids from human pluripotent stem cells. Nature biotechnology 38, 1421-1430.

168. Andersen, J., Revah, O., Miura, Y., Thom, N., Amin, N.D., Kelley, K.W., Singh, M., Chen, X., Thete, M.V., Walczak, E.M., et al. (2020). Generation of Functional Human 3D Cortico-Motor Assembloids. Cell 183, 1913-1929.e1926.

169. Lattenkamp, E.Z., and Vernes, S.C. (2018). Vocal learning: a language-relevant trait in need of a broad cross-species approach. Current Opinion in Behavioral Sciences 21, 209-215.

170. Pfenning, A.R., Hara, E., Whitney, O., Rivas, M.V., Wang, R., Roulhac, P.L., Howard, J.T., Wirthlin, M., Lovell, P.V., Ganapathy, G., et al. (2014). Convergent transcriptional specializations in the brains of humans and song-learning birds. Science 346, 1256846.

171. Vernes, S.C. (2017). What bats have to say about speech and language. Psychonomic Bulletin & Review 24, 111-117.

172. Knörnschild, M. (2014). Vocal production learning in bats. Current Opinion in Neurobiology 28, 80-85.

173. Ravignani, A., Fitch, W.T., Hanke, F.D., Heinrich, T., Hurgitsch, B., Kotz, S.A., Scharff, C., Stoeger, A.S., and de Boer, B. (2016). What Pinnipeds Have to Say about Human Speech, Music, and the Evolution of Rhythm. Front Neurosci 10, 274.

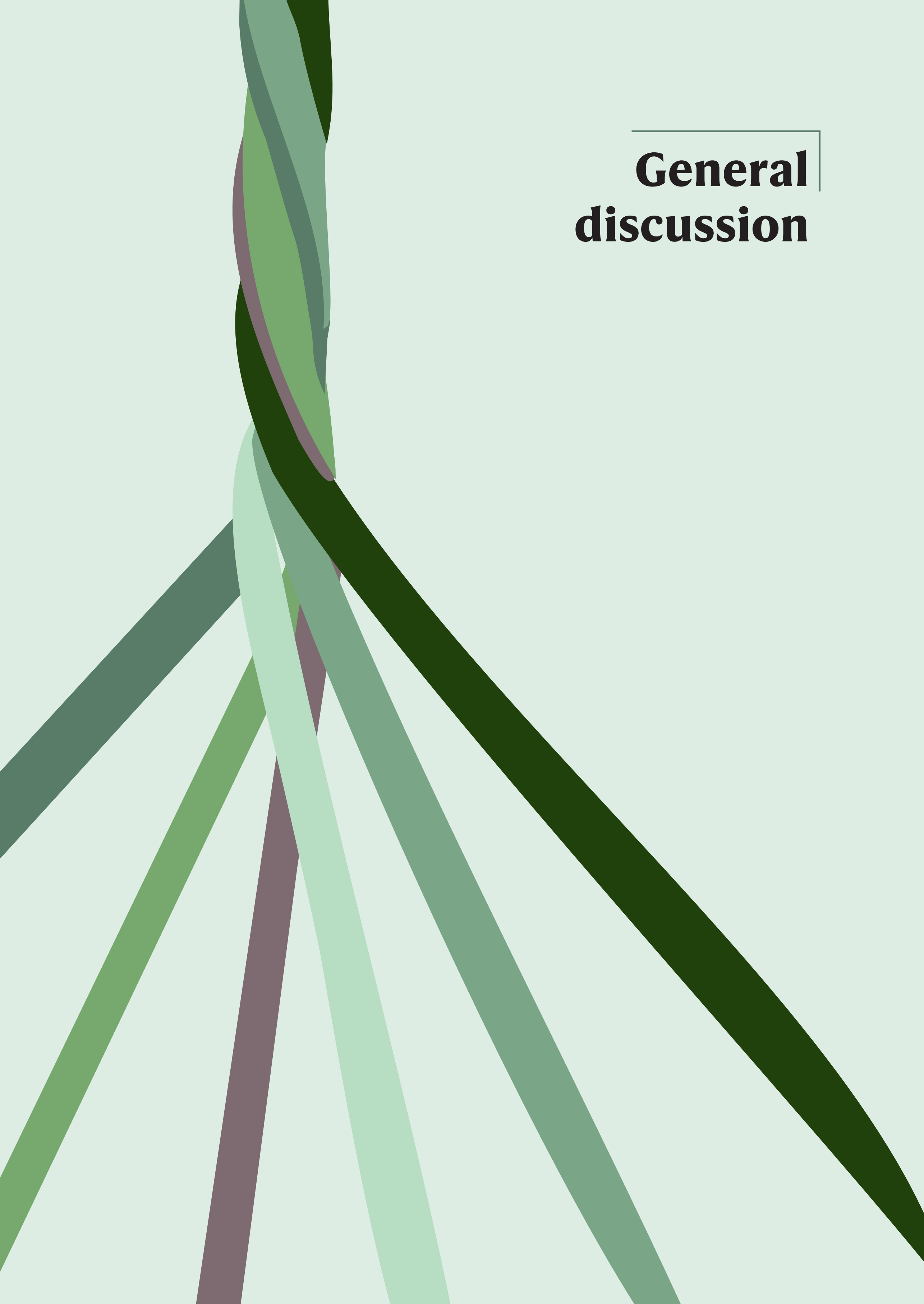
174. Rodenas-Cuadrado, P.M., Mengede, J., Baas, L., Devanna, P., Schmid, T.A., Yartsev, M., Firzlaff, U., and Vernes, S.C. (2018). Mapping the distribution of language related genes FoxP1, FoxP2, and CntnaP2 in the brains of vocal learning bat species. The Journal of comparative neurology 526, 1235-1266.

175. Belton, E., Salmond, C.H., Watkins, K.E., Vargha-Khadem, F., and Gadian, D.G. (2003). Bilateral brain abnormalities associated with dominantly inherited verbal and orofacial dyspraxia. Human Brain Mapping 18, 194-200.

176. Vargha-Khadem, F., Watkins, K.E., Price, C.J., Ashburner, J., Alcock, K.J., Connelly, A., Frackowiak, R.S., Friston, K.J., Pembrey, M.E., Mishkin, M., et al. (1998). Neural basis of an inherited speech and language disorder. Proceedings of the National Academy of Sciences of the United States of America 95, 12695-12700.

7

General discussion



Summary of the results

The studies in this thesis converge on the coupling of clinical phenotype data to functional read-outs from cell-based techniques to examine the molecular underpinnings of phenotypic variability in speech disorders and neurodevelopmental disorders (NDDs; Figure 1; **Chapter 2**). Based on the scope of the individual studies, we used different types of cell systems, covering immortalised cell lines to screen for effects of multiple putative pathogenic variants (**Chapter 3**), patient-derived cells to examine the effects of variants on endogenous gene expression levels (**Chapter 4**) and stem cell-based models to study gene function in relevant contexts (**Chapter 5**).

In **Chapter 3**, we focused on the gene *SATB1* using a combination of clinical data with functional read-outs from cell-based assays. After we identified an individual with a *SATB1* protein-truncating variant with developmental delay and severely affected speech, we assembled a cohort of individuals with (*de novo*) *SATB1* variants with a wide range of clinical features. To establish a genotype-phenotype correlation in *SATB1*-related disorder, we used an overexpression-based approach in immortalised cells to test for the functional consequences of nine *SATB1* putative pathogenic variants (identified in fifteen unrelated affected individuals), three rare *SATB1* variants present in healthy individuals, as well as three *SATB1* variants to study post-translational modification, and an aetiological *SATB2* variant. By assessing the effects of this relatively large number of genetic variants, in combination with an objectified clustering analysis based on standardised clinical data, we were able to establish that different variant types in *SATB1* are related to distinct clinical entities, with missense variants in the DNA-binding domains causing a severe NDD including hypotonia and epilepsy, and protein-truncating variants resulting in a milder clinical outcome.

For the second genotype-phenotype study of the thesis, through close collaboration with clinicians, we collected twenty-one families with a proband with phenotypic features strongly overlapping with the clinical spectrum of a previously described disorder involving *de novo* variants of the *CHD3* gene (**Chapter 4**). In all these families, an inherited *CHD3* variant was identified, with carrier parents who were healthy, or only mildly affected, hinting towards variable phenotypic expressivity. Using objectified analyses of standardised clinical data and facial photographs, we confirmed strong overlap between the phenotypes of individuals with *CHD3*-related disorder carrying a *de novo CHD3* variant, and probands with inherited *CHD3* variants. Using patient-derived immortalised cell lines from a family with a protein truncating *CHD3* variant, we found no differences between *CHD3* transcript and CHD3 protein expression levels between the affected proband, two healthy carrier relatives and one non-carrier family member, as well as five unrelated controls. Thus, compensation of *CHD3* expression from the wildtype allele was not a likely explanation for the variability in phenotypes in the family. This study shows how objectified phenotype data combined with patient-derived cell lines can be used to confirm or exclude potential molecular mechanisms underlying phenotypic variability.

In **Chapter 5** we further extended our investigations of *CHD3* using cell-based functional assays. We created heterozygous and homozygous loss-of-function variants in an induced pluripotent stem cell (iPSC) line using CRISPR-Cas9 gene editing. These genetically modified cell lines were differentiated into cerebral organoids to examine the effects of loss of *CHD3* on neurodevelopment. Analysing the organoids with an array of methods including immunohistochemistry, and bulk and single-cell transcriptomics, we found *CHD3* to be important for the balance between neural progenitor cells and mature neurons. In particular, a lack of *CHD3* resulted in an increase in expression of genes regulating neural progenitor maintenance and a decrease in genes important for neuronal differentiation. These experiments show the potential of stem cell-based model systems to uncover novel gene functions in the relevant tissue type, and to provide new hypotheses for functional characterisation of genetic variants identified in patients.

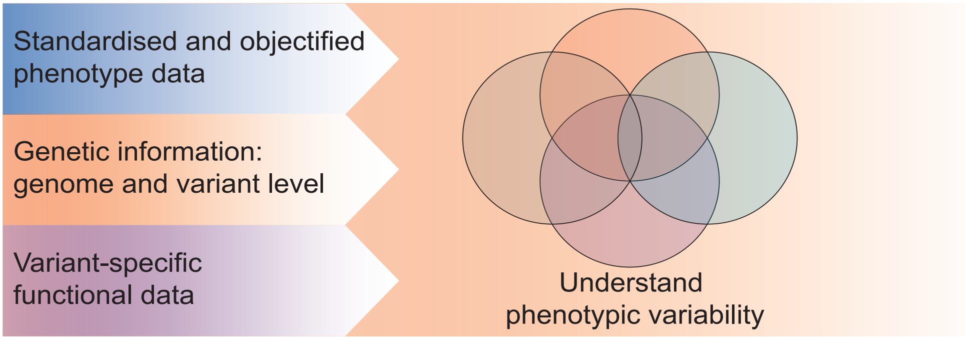

Standardised and objectified phenotype data
Genetic information: genome and variant level
Variant-specific functional data
Understand phenotypic variability

Figure 1. Integration of clinical and functional data.

Overall, the research in this thesis demonstrates the importance of functional assays for establishing pathogenicity of newly identified variants and drawing genotype-phenotype correlations. While *in silico* predictions of the impact of variants and/or clinical data by themselves are often unable to draw firm conclusions or to identify subtle differences, integrating their output with read-outs derived from cell-based functional assays could provide a more complete understanding of the variability of phenotypes associated with speech disorders and NDDs (Figure 1). Advances in the field of cell-based functional testing to allow for higher-throughput and more in-depth functional read-outs combined with advances in registering phenotype data in standardised ways will be crucial to map the molecular space underlying phenotypic variability in human disorder.

Contributions to the field

In this thesis we performed in-depth review of the literature on monogenic speech phenotypes and *FOXP2* (**Chapter 2** and **Chapter 6**), a gene associated with a Mendelian form of severe speech disorder[1]. Moving beyond well-studied genes like *FOXP2*, an individual with severely affected speech and a *de novo SATB1* variant prompted us to further investigate *SATB1* and its link to human disease (**Chapter 3**). We found SATB1 of particular interest given its protein-protein interactions with FOXP2 and SATB2[2], another protein associated with an NDD characterised by severely affected speech[3]. By using a gene-driven approach to identify more individuals with putative pathogenic *SATB1* variants, combined with clinical and cell-based functional analyses, we identified that missense variants were associated with severe symptoms, including intellectual disability and early-onset epilepsy, while protein-truncating variants cause a milder NDD phenotype. These two clinically and molecularly distinct disorders are now listed as Mendelian NDDs in the Online Mendelian Inheritance in Man (OMIM) database: ‘Kohlschutter-Tonz syndrome-like’ (MIM #619229) and ‘Developmental delay with dysmorphic facies and dental anomalies’ (MIM #619228) respectively. Furthermore, *SATB1* has been added to the ‘Genetic epilepsy syndromes’ and ‘Intellectual disability’ gene panels for genetic diagnostic testing in routine care of intellectual disability and epilepsy, including the Genomics England PanelApp, an evidence-based curated collection of gene panels widely used in the field of genetics of rare diseases[4].

In 2018, a *de novo* missense variant in the *CHD3* gene was reported in a child with childhood apraxia of speech[5]. A gene-driven follow up study identified thirty-five individuals with *de novo CHD3* missense variants clustering in one of the functional domains of the protein in individuals with a broad NDD syndrome ranging from mild to more severe features[6]. We further expanded the clinical and genetic spectrum of *CHD3*-related disorder by describing twenty-one families with inherited *CHD3* missense and protein-truncating variants (**Chapter 4**). While the phenotypes of the affected probands strongly overlapped with the described NDD associated with *de novo CHD3* variants, an objectified clustering analysis showed that the phenotypes of the probands were significantly different from the mildly affected or healthy carrier parents. Hence, we demonstrated variable expressivity for these inherited *CHD3* variants, in contrast to the previously described highly penetrant *de novo* variants. This is an important finding for the diagnostics of *CHD3*-related disorder, indicating that rare inherited variants in *CHD3* should not be excluded as possibly causative. Cases with a likely pathogenic inherited variant in *CHD3* and the transmitting parents should be closely monitored for phenotypic features linked to *CHD3*-related disorder. Our example of *CHD3* suggests that variable expressivity/reduced penetrance, also in genes already implicated in dominant forms of NDD, could potentially explain some of the so far unexplained NDD cases, and functional studies could help to better understand what underlies the phenotypic variability (for examples see Ref. 7 and 8),

To increase understanding of the functions of *CHD3* during early brain development, we generated heterozygous and homozygous knockout stem cell lines that we differentiated into cerebral organoids (**Chapter 5**). Aetiological *CHD3* missense variants disrupt the chromatin remodelling function of the protein[6]. One of the main phenotypic features of *CHD3*-related disorder is macrocephaly[6; 9] and therefore we hypothesized that dysfunction of CHD3 could result in a dysregulation of neural progenitor self-renewal and neuronal differentiation underlying brain growth. We indeed found that a loss of CHD3 resulted in a shift in cell composition in brain organoids, with relatively more cells with a neural progenitor cell identity in organoids lacking *CHD3* expression. These findings were consistent with the increase in expression of genes promoting neural stem cell identity and a decrease in markers of neuronal differentiation. These results show that cerebral organoids from genetically modified or patient-derived cells should provide suitable model systems to functionally characterise pathogenic *CHD3* variants. Moreover, the results from the knockout cell lines have uncovered self-renewal of neural progenitor cells and the switch to differentiation as a particular pathway that could be the focus of future studies on characterisation of patient variants using high-throughput functional assays.

Functional testing of genetic variants

As next-generation sequencing generates large amounts of data, interpretation of genetic variants in a clinical context remains challenging. Filtering for *de novo* or bi-allelic variants is powerful for identifying disease variants[10; 11], but the majority of individuals with NDDs remain undiagnosed. While *de novo* loss-of-function variants in genes associated with haploinsufficiency disorders and intolerance to loss-of-function variation are mostly detected in the current computational variant filtering workflows, rare inherited loss-of-function variants and in particular (both *de novo* and inherited) missense variants are challenging to classify. At present, such variants are often reported as variants of unknown significance, and keep the affected individuals and their families, as well as treating physicians, in a situation of uncertainty. Functional assays can aid to establish pathogenicity, and therefore, gene-driven cohort studies that describe rare genetic diseases often include the functional characterisation of (a selection) of putative pathogenic variants, as we have done in the studies presented in **Chapter 3** and **Chapter 4**. The different model systems that are commonly used have their own advantages and disadvantages (Figure 2).

Using animal models for functional characterisation of genes

Animal models are widely used, and can inform about both molecular pathogenic mechanisms and effects on behavioural aspects (Figure 2). In **Chapter 6** of this thesis, we discuss in detail what we have learned from studies with mice about the cellular functions of FOXP2 and its contributions to behaviour. For example, mice with a conditional cortex-specific knock-out of *Foxp2* have shown that the expression of the gene is important for social behaviour and cognitive flexibility[12; 13], while a mouse model carrying a heterozygous variant equivalent to the variant identified in a family with severe speech disorder[1] displays reduced motor-skill learning and abnormalities in ultrasonic vocalisations compared to wildtype animals[14; 15]. Although an increasing number of different *FOXP2* variants are described in individuals with speech disorder in literature[16], only two mouse models exist with genetic variants that match those identified in humans[14]. As the generation of mice with a genetic modification is time-consuming and costly, and performing the subsequent experiments is limited by ethical guidelines, researchers often work with a knock-out model or a small selection of gene disruptions. When investigating the putative pathogenic mechanisms of a monogenic haploinsufficiency disorder, this may be sufficient, but such an approach does not cover all aspects when multiple modes of pathogenicity are associated with the gene (such as what we describe for *SATB1* in **Chapter 3**). Moreover, variants with similar molecular characteristics do not always affect gene functions in a comparable way, as we reported for pathogenic missense variants in *SATB2* (**Chapter 3**), and therefore an animal model with a specific variant may not be representative for other variants of the same type. Thus, animal models can provide in-depth information about gene function and the associations with behaviour, but are not suitable to screen for a large number of putative pathogenic variants (Figure 2). Furthermore, certain disorders are caused by human-specific genes, such as *NOTCH2NL* copy number variants associated with macrocephaly and microcephaly[17], genes can have

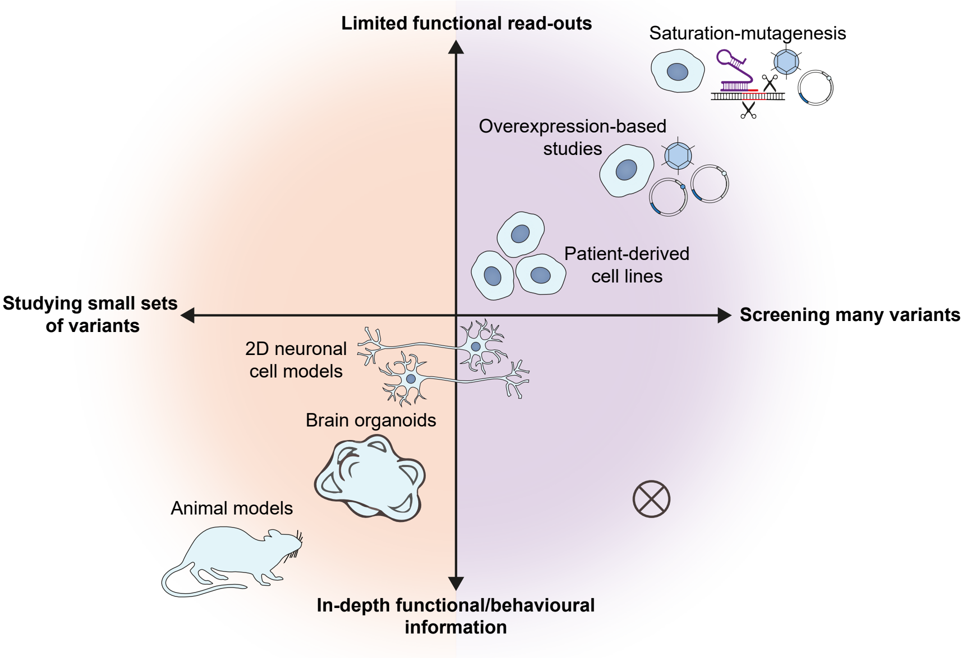

Limited functional read-outs
Saturation-mutagenesis
Overexpression-based studies
Patient-derived cell lines
Studying small sets of variants
Screening many variants
2D neuronal cell models
Brain organoids
Animal models
In-depth functional/behavioural information

Figure 2. Plot of strengths and weaknesses of model systems used for functional characterisation of genetic disruptions and variants. The cross represents where combinations of novel and/or future advances in high-throughput functional assays and automated culturing of physiological cell-based model systems could potentially position in this strengths-weaknesses plot.

human-specific functions (e.g. transcriptional regulation of FOXP2[18]), or the genetic background of the affected individuals can have a modifying role, as we hypothesized for inherited *CHD3* variants described in **Chapter 4**. Such conditions cannot easily be investigated in animal models.

Using cell-based models for functional characterisation

To circumvent the limitations of animal models in relation to human-specific mechanism or scalability, genetic studies often include functional assays that are performed in human cells. Different human cellular models exist, from patient-derived stem cell-based neurons[19; 20], immortalised patient-derived cells such as skin fibroblasts or lymphoblastoid cells (which we used in **Chapter 4**), to immortalised tumour cell lines combined with overexpression approaches (see **Chapter 3**). Although the breadth of scalability depends on the type of human cellular model (Figure 2), in general, these cell-based approaches allow for the functional characterisation of multiple putative pathogenic variants in parallel given that a suitable functional read-out can be identified. Patient-derived (stem cell-based) methods have the advantage of studying genetic variants in endogenously expressed genes in the genetic background of the affected individual, but they are limited by the availability of patient material and the requirement of patient consent. An overexpression approach circumvents the issue of availability, but is more limited in terms of the complexity of the functional read-outs, as it studies the gene in a non-physiological state. Overall, stem cell-based models of specific cell types are suitable to screen for effects of a selection of genetic variants on a variety of cell type-specific cellular and molecular mechanisms, while overexpression-based methods make it possible to screen for the effects of larger numbers of genetic variants on one or a small number of specific protein functions (Figure 2).

Recent developments in cell-based model systems

Advances in the field of human cellular models are changing how we study gene function and the effects of genetic variants. With the emergence of protocols to culture brain organoids[21], we can now study human-specific developmental processes and neuronal circuitry in a cellular model containing multiple types of cells[22] and gene expression profiles comparable to the developing foetal brain[23]. Although these methods require more complex culture media, long-time culturing (> 1 month) and specific skills, limiting their scalability, they can be grown from patient-derived cells, or cell lines that are genetically modified. This makes brain organoids a suitable model for testing the effects of selections of putative pathogenic variants on neurodevelopmental processes, such as differentiation, migration and neuronal functions.

In **Chapter 5**, we used CRISPR-Cas9 gene editing to generate stem cell lines carrying homozygous and heterozygous *CHD3* knock-out alleles, which we differentiated into cerebral organoids. Although these knock-out lines can teach us about the essential functions of *CHD3* during early stages of brain development, we did not directly model the patient-condition, as haploinsufficiency is unlikely to be the underlying pathogenic mechanism in typical forms of *CHD3*-related disorder[6], and almost all *CHD3* protein-truncating variants described have shown variable expressivity (**Chapter 4**). Future studies could use patient-derived cell lines, allowing to study the variant of interest in the genetic background of the patient, while a genetic rescue using gene editing could increase the understanding of specific contributions of the pathogenic allele. However, patient-derived cells require optimisation of the brain organoid protocol for each cell line, and complicate the experimental design with the need of age- and sex-matched controls. Gene-edited lines circumvent problems with availability, patient consent, appropriate controls, and allow researchers to model any variant of interest. Moreover, developments with the CRISPR-Cas9 toolkit have led to high efficiency in creating specific single nucleotide variants in a gene of interest[24].

Advances are made on the other end of spectrum of cellular models as well. Although using simple non-physiological expression systems, such as immortalised cells, yeast or bacterial phages, deep mutational scanning approaches using saturation mutagenesis can screen for the effects of all possible variants in a gene on a specific protein function (Figure 2)[25]. These methods are powerful tools to create functional variant effect maps, revealing which regions of the gene are tolerant or intolerant to variation, but also which variant types and residue changes are particularly detrimental to particular gene functions. In **Chapter 3**, we used an overexpression-based system to characterise *SATB1* missense variants located in the DNA-binding domains of the protein. We focused specifically on three protein functions related to these domains (nuclear localisation, transcriptional activity and protein mobility), and found all tested missense variants to result in stronger DNA-binding and increased transcriptional activity. We did not assess thethe effects of any missense variants located outside these domains. As we identified one of the pathogenic mechanisms of *SATB1*-related disorders to be haploinsufficiency, it is possible that missense variants in other regions of the protein disrupt the function of SATB1, causing haploinsufficiency. Therefore, performing a deep mutational scan on *SATB1* could help in the future to better understand where and which types of missense variation could lead to *SATB1* haploinsufficiency. Interpreting the role of missense variants remains challenging in other disease genes as well. So far, these developments in functional testing have only been used for a handful of genes[26-28], and have not been fully exploited yet.

Integration of clinical and functional data

Standardising clinical data

NDDs and speech disorders are both genetically and phenotypically heterogeneous. While genetic variation in many different genes can result in overlapping phenotypic features, individuals with putative pathogenic variants in the same gene can present with variable clinical outcomes[29]. In order to consistently compare and analyse disease phenotypes between and within NDDs and speech disorders, clinical data need to be recorded in a comprehensive, standardised and accessible format. However, small cohort studies often rely on clinical data with variable completeness and level of detail. Moreover, in many cases extra attention is given to a specific phenotypic feature that falls within the expertise of the medical specialist or facility that diagnosed the individual, while other aspects may be missed. Another complication is the use of different clinical terminologies, with various terms in use for overlapping phenotypic features, sometimes with unclear criteria, making the data challenging to process and interpret.

In order to better define clinically distinct phenotypes (see **Chapter 3**), or study the phenotypic spectrum, variable expressivity or reduced penetrance of a syndrome (see **Chapter 4**), it is important that human phenotypic data collections are made suitable for downstream (computational) analyses and comparisons. However, for that, clinical data need a computational data structure with a controlled vocabulary. In 2008, the Human Phenotype Ontology (HPO) was launched[33]. The HPO is a data model that includes all clinical entries that are observed in human monogenic diseases, derived from the OMIM database, curated by clinical experts. With recent extensions[34; 35], the HPO has become a widely used method for exchange and analysis of phenotypic data[35]. For example, the HPO is used in the variant prioritisation tool Exomiser, by matching the phenotypic data of carriers with clinical descriptions associated with disease-genes[36]. In this thesis, in **Chapter 3** and **Chapter 4**, we made use of the HPO-terminology to store our clinical data in a standardised way, and allow for objectified analyses on phenotypic clusters within our cohorts. Using the standardised registration of clinical data, we were able to objectively identify two distinct clinical entities associated with *SATB1* loss-of-function and missense variants, and we could show that individuals with *de novo* and inherited *CHD3* variants had phenotypes that were not significantly different. Instead of conducting these analyses with a selection of the most observed features, we were able to assess phenotypic clustering computationally on the complete data set, taking all available information into account. Hence, the examples in this thesis show the importance of standardised and consistent phenotype reporting.

In the field of speech and language disorders, the terminology and classification has been the topic of confusion and ongoing debate. The lack of clear criteria for identification and classification of communication phenotypes has hampered diagnosis and therapy, but has also made it challenging to compare cohort data from different genetic and/or phenotypic studies[30]. In 2016 and 2017, an international multidisciplinary consortium (CATALISE) made recommendations about the classification of speech and language disorders[30; 31]. However, aspects of these recommendations remain a topic of discussion, including the 'absence of a biomedical condition' as a criterion for developmental language disorder[32]. A consensus on the criteria of speech and language disorders and systematic reporting of clinical data could facilitate recruitment of larger and more consistent cohorts to study the genetic underpinnings of these phenotypes.

Standardising functional data

So far, functional read-outs in studies focusing on rare genetic disorders suffer the same issues as clinical data in terms of comprehensive, accessible and systematic registration. Consortia focusing on animal models have made great progress in consistently characterising phenotypes and registering functional data on knockouts, providing an incredibly rich resource for both fundamental and translational research[37]. In contrast, small cohort studies on rare diseases using cell-based assays to screen for effects of a selection of gene disruptions and aetiological variants (such as used in **Chapter 3**, **4** and **5** of this thesis) often report on different aspects of gene function with different level of detail, and/or use different methods, and are therefore more difficult to compare. With the advances in physiologically-relevant model systems to identify important novel gene functions[38], and the developments in deep mutational screening approaches to test for the effects of all possible genetic variants for specific functions[25], future efforts should focus on building a resource of comprehensive functional variant effect maps (with the Atlas of Variant Effects Alliance as a recent initiative to achieve this). Families of genes with comparable functions, or associated with overlapping genetic pathogenic mechanisms, could follow similar saturation mutagenesis workflows. For example, genes intolerant for loss-of-function variation and in which protein-truncating variants are associated with disease could be screened using a gene-essentiality approach[28]. Alternatively, genes that have been shown to be important for cell proliferation could be screened for growth capabilities[27], while those relying on post-translational modifications to perform their functions can be tested with respect to a specific type of modification[26].

Such data could potentially be combined with standardised (HPO-based) phenotypic data, to provide a valuable resource for variant interpretation, genetic diagnosis, and understanding of the molecular mechanisms of phenotypic variability (Figure 1).

Future perspectives

The emergence of next-generation sequencing in a clinical setting has accelerated gene discovery and drastically increased the diagnostic yield for NDDs[39]. For speech disorders, these developments have so far lagged behind, but with better consensus on the classification of speech disorders[30; 31], the efforts to assemble thoroughly phenotyped cohorts for genetic follow-ups[5; 32; 40], and the decreasing costs for exome and genome sequencing, the field of genetics of speech disorders could rise to a comparable level in the coming years.

Functional cell-based screens are a powerful tool to help interpret likely pathogenic variants that emerge from next-generation sequencing strategies. At the moment, there still exists a trade-off for using different types of cell-based model systems. Immortalised cell lines, combined with large plasmid-based (over)expression or gene-editing libraries, are compatible with high-throughput testing of genetic variants, but can only do this for a selected number of protein functions, in a non-physiological context. In contrast, physiologically-relevant systems make it possible to decipher the effects of aetiological variants on complex mechanisms, but long-term and costly culture conditions limit the number of variants that can be tested at the same time (Figure 2). Moreover, although gene editing in isogenic lines offers a promising way to circumvent difficulties regarding patient material availability, consent for the use of the cells and appropriate experimental controls, patient-derived lines will remain essential when the genetic backgrounds of the affected individuals need to be taken into account.

With the current techniques, physiologically-relevant model systems, such as the cerebral organoids, seem appropriate models for uncovering novel gene functions. This is, in particular, valuable for genes that have largely remained unstudied in the relevant tissues, and therefore their functions are unknown. Studies using knockout models and small sets of patient-derived cell lines and/or lines carrying aetiological variants, followed by sophisticated cell culturing protocols to generate cell types of interest, could generate hypotheses about specific protein functions affected and can steer the design of (high-throughput) follow-up experiments. For example, the protein localisation and mobility assays that we used in immortalised cell lines to assess the effects of novel putative *SATB1* missense variants, were based on earlier discoveries in thymocytes and T_H cells[41; 42], cell types relevant to the expression and function of SATB1. Conversely, for *CHD3*, using cerebral organoids, we identified roles in neural cell proliferation and differentiation, that could potentially be used in functional characterisation studies of putative pathogenic variants in the gene. All in all, the array of currently available cell model systems serves different goals.

In closing, it is worth nothing that the developments in this field tend to be fast. While stem cell-derived organoid models are being adapted for (automated) larger-scale culturing[43; 44], the organisation of stem cell facilities working together with (local) biobanks overcomes issues with availability of patient material[45]. Moreover, with current functional genome-wide screening protocols using CRISPR libraries it is already possible to test the essentiality of every single gene in human neuronal cells[46; 47], and such studies hold promise for deep mutational scans to screen for variant effects in more physiologically-relevant cell types on a multitude of protein functions. Combinations of these technical advances may result in studies that screen large numbers of genetic variants in physiologically-relevant cell model systems (Figure 2) that will fill our current gaps in knowledge required to improve variant interpretation. Ultimately, these endeavours should help to better understand phenotype variability in speech disorders and NDDs, providing crucial new insights into human brain development and the myriad ways in which this may go awry.

References

1. Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 519-523.
2. Estruch, S.B., Graham, S.A., Quevedo, M., Vino, A., Dekkers, D.H.W., Deriziotis, P., Sollis, E., Demmers, J., Poot, R.A., and Fisher, S.E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. Human molecular genetics 27, 1212-1227.
3. Snijders Blok, L., Goosen, Y.M., van Haaften, L., van Hulst, K., Fisher, S.E., Brunner, H.G., Egger, J.I.M., and Kleefstra, T. (2021). Speech-language profiles in the context of cognitive and adaptive functioning in SATB2-associated syndrome. Genes, Brain and Behavior 20, e12761.
4. Stark, Z., Foulger, R.E., Williams, E., Thompson, B.A., Patel, C., Lunke, S., Snow, C., Leong, I.U.S., Puzriakova, A., Daugherty, L.C., et al. (2021). Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution. The American Journal of Human Genetics 108, 1551-1557.
5. Eising, E., Carrion-Castillo, A., Vino, A., Strand, E.A., Jakielski, K.J., Scerri, T.S., Hildebrand, M.S., Webster, R., Ma, A., Mazoyer, B., et al. (2019). A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. Molecular Psychiatry 24, 1065-1078.
6. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al. (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nature Communications 9, 4619.
7. Raraigh, K.S., Han, S.T., Davis, E., Evans, T.A., Pellicore, M.J., McCague, A.F., Joynt, A.T., Lu, Z., Atalar, M., Sharma, N., et al. (2018). Functional Assays Are Essential for Interpretation of Missense Variants Associated with Variable Expressivity. The American Journal of Human Genetics 102, 1062-1077.
8. Jensen, M., Tyryshkina, A., Pizzo, L., Smolen, C., Das, M., Huber, E., Krishnan, A., and Girirajan, S. (2021). Combinatorial patterns of gene expression changes contribute to variable expressivity of the developmental delay-associated 16p12.1 deletion. Genome Medicine 13, 163.
9. Drivas, T.G., Li, D., Nair, D., Alaimo, J.T., Alders, M., Altmüller, J., Barakat, T.S., Bebin, E.M., Bertsch, N.L., Blackburn, P.R., et al. (2020). A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome. European journal of human genetics: EJHG 28, 1422-1431.
10. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433-438.
11. Martin, H.C., Jones, W.D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson, J.D., Jones, C.P., Handsaker, J., Gallone, G., Bruntraeger, M., et al. (2018). Quantifying the contribution of recessive coding variation to developmental disorders. Science 362, 1161-1164.
12. Co, M., Hickey, S.L., Kulkarni, A., Harper, M., and Konopka, G. (2020). Cortical Foxp2 Supports Behavioral Flexibility and Developmental Dopamine D1 Receptor Expression. Cerebral Cortex 30, 1855-1870.
13. Medvedeva, V.P., Rieger, M.A., Vieth, B., Mombereau, C., Ziegenhain, C., Ghosh, T., Cressant, A., Enard, W., Granon, S., Dougherty, J.D., et al. (2019). Altered social behavior in mice carrying a cortical Foxp2 deletion. Human molecular genetics 28, 701-717.
14. Groszer, M., Keays, D.A., Deacon, R.M., De Bono, J.P., Prasad-Mulcare, S., Gaub, S., Baum, M.G., French, C.A., Nicod, J., and Coventry, J.A. (2008). Impaired synaptic plasticity and motor learning in mice with a point mutation implicated in human speech deficits. Current Biology 18, 354-362.
15. Chabout, J., Sarkar, A., Patel, S.R., Radden, T., Dunson, D.B., Fisher, S.E., and Jarvis, E.D. (2016). A Foxp2 mutation implicated in human speech deficits alters sequencing of ultrasonic vocalizations in adult male mice. Frontiers in behavioral neuroscience 10, 197.
16. Reuter, M.S., Riess, A., Moog, U., Briggs, T.A., Chandler, K.E., Rauch, A., Stampfer, M., Steindl, K., Gläser, D., and Joset, P. (2017). FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. Journal of medical genetics 54, 64-72.
17. Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M., van den Bout, A., Bishara, A., Rosenkrantz, J.L., et al. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 173, 1356-1369.e1322.

18. Konopka, G., Bomar, J.M., Winden, K., Coppola, G., Jonsson, Z.O., Gao, F., Peng, S., Preuss, T.M., Wohlschlegel, J.A., and Geschwind, D.H. (2009). Human-specific transcriptional regulation of CNS development genes by FOXP2. Nature 462, 213-217.
19. Shi, Y., Kirwan, P., and Livesey, F.J. (2012). Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. Nature Protocols 7, 1836-1846.
20. Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., et al. (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron 78, 785-798.
21. Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature 501, 373-379.
22. Velasco, S., Kedaigle, A.J., Simmons, S.K., Nash, A., Rocha, M., Quadrato, G., Paulsen, B., Nguyen, L., Adiconis, X., Regev, A., et al. (2019). Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. Nature 570, 523-527.
23. Camp, J.G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M., et al. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proceedings of the National Academy of Sciences 112, 15672.
24. Anzalone, A.V., Koblan, L.W., and Liu, D.R. (2020). Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. Nature Biotechnology 38, 824-844.
25. Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature Methods 11, 801-807.
26. Leung, I., Dekel, A., Shifman, J.M., and Sidhu, S.S. (2016). Saturation scanning of ubiquitin variants reveals a common hot spot for binding to USP2 and USP21. Proc Natl Acad Sci U S A 113, 8705-8710.
27. Mighell, T.L., Evans-Dutson, S., and O'Roak, B.J. (2018). A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. American journal of human genetics 102, 943-955.
28. Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature 562, 217-222.
29. Parenti, I., Rabaneda, L.G., Schoen, H., and Novarino, G. (2020). Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends in Neurosciences 43, 608-621.
30. Bishop, D.V.M., Snowling, M.J., Thompson, P.A., Greenhalgh, T., and consortium, C. (2016). CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. PLOS ONE 11, e0158753.
31. Bishop, D.V., Snowling, M.J., Thompson, P.A., Greenhalgh, T., Consortium, C., Adams, C., Archibald, L., Baird, G., Bauer, A., and Bellair, J. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. Journal of Child Psychology and Psychiatry 58, 1068-1080.
32. Snijders Blok, L. (2021). Let the genes speak! De novo variants in developmental disorders with speech and language impairment.
33. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics 83, 610-615.
34. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., and Campbell, J. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research 42, D966-D974.
35. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al. (2021). The Human Phenotype Ontology in 2021. Nucleic Acids Research 49, D1207-D1217.
36. Smedley, D., Jacobsen, J.O.B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. Nature Protocols 10, 2004-2015.
37. Brown, S.D.M. (2021). Advances in mouse genetics for the study of human disease. Human molecular genetics 30, R274-R284
38. Amin, N.D., and Paşca, S.P. (2018). Building Models of Brain Disorders with Three-Dimensional Organoids. Neuron 100, 389-405.
39. Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability

and related disorders. Nature Reviews Genetics 17, 9-18.
40. Hildebrand, M.S., Jackson, V.E., Scerri, T.S., Van Reyk, O., Coleman, M., Braden, R.O., Turner, S., Rigbye, K.A., Boys, A., Barton, S., et al. (2020). Severe childhood speech disorder: Gene discovery highlights transcriptional dysregulation. Neurology 94, e2148-e2167.
41. Cai, S., Lee, C.C., and Kohwi-Shigematsu, T. (2006). SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. Nature Genetics 38, 1278-1288.
42. Ghosh, R.P., Shi, Q., Yang, L., Reddick, M.P., Nikitina, T., Zhurkin, V.B., Fordyce, P., Stasevich, T.J., Chang, H.Y., Greenleaf, W.J., et al. (2019). Satb1 integrates DNA binding site geometry and torsional stress to differentially target nucleosome-dense regions. Nature communications 10, 3221-3221.
43. Renner, H., Grabos, M., Becker, K.J., Kagermeier, T.E., Wu, J., Otto, M., Peischard, S., Zeuschner, D., TsyTsyura, Y., Disse, P., et al. (2020). A fully automated high-throughput workflow for 3D-based chemical screening in human midbrain organoids. eLife 9, e52904.
44. Tristan, C.A., Ormanoglu, P., Slamecka, J., Malley, C., Chu, P.-H., Jovanovic, V.M., Gedik, Y., Jethmalani, Y., Bonney, C., Barnaeva, E., et al. (2021). Robotic high-throughput biomanufacturing and functional differentiation of human pluripotent stem cells. Stem Cell Reports 16, 3076-3092.
45. Steeg, R., Mueller, S.C., Mah, N., Holst, B., Cabrera-Socorro, A., Stacey, G.N., De Sousa, P.A., Courtney, A., and Zimmermann, H. (2021). EBiSC best practice: How to ensure optimal generation, qualification, and distribution of iPSC lines. Stem Cell Reports 16, 1853-1867.
46. Tian, R., Abarientos, A., Hong, J., Hashemi, S.H., Yan, R., Dräger, N., Leng, K., Nalls, M.A., Singleton, A.B., Xu, K., et al. (2021). Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. Nature neuroscience 24, 1020-1034.
47. Tian, R., Gachechiladze, M.A., Ludwig, C.H., Laurie, M.T., Hong, J.Y., Nathaniel, D., Prabhu, A.V., Fernandopulle, M.S., Patel, R., Abshari, M., et al. (2019). CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. Neuron 104, 239-255.e212.

Appendices

English summary

Dutch summary

Curriculum vitae

List of publications

English summary

Speech and language are complex abilities that children typically acquire with little effort and without the need to be formally taught. However, when children have persistent problems with speech and language, they can be diagnosed with a developmental disorder. While in many cases the genetic architecture of these speech and language disorders is complex, some are monogenic and are caused by rare high-penetrant variants in a single gene.

FOXP2 was the first gene that could be linked to a monogenic speech disorder, by studying a large multigenerational family of which half of the members presented with developmental speech difficulties. The affected family members all carried a *FOXP2* missense variant that disrupted the function of the gene (described in **Chapter 1**, **2** and **6**). Since that discovery, more genetic disruptions have been found in *FOXP2* in individuals with severe speech problems. Subsequent studies have used various *in vitro* and *in vivo* model systems to further characterise aetiological *FOXP2* variants and to uncover FOXP2's functions in the brain, reviewed in detail in **Chapter 6**.

While next-generation sequencing approaches combined with strategies prioritising *de novo* and/or bi-allelic variants have increased the genetic diagnostic yield for Mendelian neurodevelopmental disorders, such as intellectual disability and autism spectrum disorder, employing these methods to diagnose and study monogenic forms of speech and language disorders is still at an early stage. In **Chapter 1** and **2** of this thesis, we provide an overview of the current state of the field of the genetics of human speech disorders. Although it remains challenging to assemble well-phenotyped cohorts of individuals with speech disorders, due to the lack of consensus on the criteria for classifying communication disorders among other things, a number of phenotype-driven studies have used next-generation sequencing strategies to identify causative variants with a large effect. Interestingly, genes associated with these Mendelian forms of speech disorder converge on regulatory genes expressed during early stages of human brain development.

When genes associated with developmental speech disorders are used as a starting point for gene-driven follow up studies, analyses of additional putative pathogenic variants often establish a link to a broader spectrum of neurodevelopmental phenotypes (**Chapter 2**). For example, a *de novo* missense variant in *CHD3* was initially found in a child with childhood apraxia of speech. When an international collaboration identified an additional 35 individuals with *de novo CHD3* variants, the *CHD3*-associated phenotype was found to range from mild developmental delay to severe intellectual disability, without a clear genotype-phenotype correlation. So far, it often remains unclear what underlies the phenotypic variability observed between individuals with pathogenic variants in the same gene.

This thesis investigated how we could use different types of cell-based functional assays to obtain information about the molecular consequences of genetic variants in neurodevelopmental and speech disorders, coupling this to objectified phenotypic analyses to learn more about the molecular landscape underlying phenotypic variability.

In **Chapter 3**, we associated pathogenic variants in the *SATB1* gene with two distinct clinical outcomes using a combination of phenotypic and cell-based functional studies. *SATB1* encodes a transcription factor that serves as a chromatin organiser, and that is part of the FOXP2 interactome. However, prior to our work, *SATB1* had not yet been linked to disease. A child with intellectual disability and severely affected speech carrying a *de novo* protein-truncating *SATB1* variant prompted us to perform a gene-driven study to identify additional individuals with neurodevelopmental phenotypes carrying high-penetrant (*de novo*) variants in this gene. We identified 42 individuals with missense and protein-truncating variants in *SATB1*, and followed up by functionally characterising a significant subset of the identified variants with various molecular assays. Using immortalised cell lines we studied the effects of *SATB1* missense variants on co-localisation with the DNA, transcriptional activity and protein mobility. Coupling these functional data to objectified computational analyses of clinical features associated with *SATB1* missense and protein-truncating variants, we were able to identify a genotype-phenotype correlation linked to variant type. *SATB1* missense variants that cluster in the DNA-binding domains of the encoded protein, causing increased transcriptional activity, were associated with a severe neurodevelopmental syndrome including intellectual disability, epilepsy and in some cases premature death. In contrast, *SATB1* protein-truncating variants had a loss-of-function effect and caused a milder neurodevelopmental condition.

Next, we turned to the *CHD3* gene. So far, the large majority of individuals with *CHD3*-associated disorder that have been described in the literature carry *de novo* missense variants. In **Chapter 4**, we report the discovery of 21 families in which a proband with a neurodevelopmental disorder carries an inherited *CHD3* variant transmitted from a healthy or mildly affected parent, of which eight involve a protein-truncating variant. The probands in these families all had phenotypic features that overlapped with the phenotype described for *CHD3*-associated disorder. We used objectified computational analyses of clinical data to show that the phenotypes of probands with an inherited *CHD3* variant did not significantly differ from phenotypes associated with *de novo CHD3* variants. In contrast, the phenotypes of the probands with an inherited *CHD3* variant were found to significantly differ from phenotypes of their carrier parents, confirming that these inherited variants confer variable expressivity. To test the hypothesis that a loss-of-function allele in the carrier parent could potentially be compensated by increased expression of the wildtype allele, we cultured immortalised patient-derived cells from a proband, the carrier mother, and carrier grandmother with a *CHD3* protein-truncating variant, as well as from a non-carrier sister, and assayed the transcript and protein expression levels. We showed that the transcript carrying the *CHD3* variant was subjected to nonsense-mediated decay leading to decreased *CHD3* expression levels, but we did not find evidence for a rescue of expression in the unaffected mother and grandmother as a potential explanation for the variable expressivity.

In **Chapter 5** we further investigated the roles of *CHD3* during human brain development, towards a better understanding of the impact of gene variants on neurodevelopmental disorder. We generated induced pluripotent stem cell lines with heterozygous and homozygous *CHD3* loss-of-function alleles using CRISPR-Cas9 gene editing, and differentiated these into cerebral organoids. Cerebral organoids are a three-dimensional cell culture model resembling the early stages of human foetal brain development. Using publically available transcriptomic data sets and our cultured cerebral organoids, we showed that *CHD3* expression is low in neural progenitor cells at early stages of development, but increases over time, with high expression in mature neurons. We used a combination of immunohistochemistry, bulk and single-cell transcriptomics to examine the effects of loss of *CHD3*. We found that decreased *CHD3* expression resulted in an increase in expression of genes important for neural progenitor cell maintenance, and decreased expression of pro-neural genes. Consistently, we observed a shift in cell type composition of the organoids. Day-57 organoids carrying a *CHD3* knockout allele had relatively more neural progenitors, and a smaller pool of neurons. These results identify CHD3 as an upstream regulator of neurogenesis, controlling the balance between neural progenitor cells and neurons. Future studies investigating *CHD3*-associated disorder will further investigate how these functions are affected by aaetiological variants in the gene.

Taken together, **Chapters 3**, **4** and **5** show how different types of cell-based model systems can be used to establish genotype-phenotype correlations, explore mechanisms underlying variable expressivity/reduced penetrance and illuminate novel gene functions in a physiologically-relevant context. In these studies, cell-based functional assays were crucial in establishing pathogenicity of variants, understanding gene function and the precise functional effects of genetic variants.

Novel developments in cell-based culture methods, including deep mutational scans that can functionally characterise thousands of genetic variants in parallel, and automatised methods to culture three-dimensional stem cell-based culture systems, hold promise to combine high-throughput functional characterisation with physiologically-relevant cell models (**Chapter 7**). Such advances will make it possible to generate genetic variant maps for a multitude of protein functions, which could improve variant interpretation, and increase understanding about phenotypic variability at a molecular level.

Nederlandse samenvatting

Spraak en taal zijn complexe vaardigheden die kinderen opdoen zonder veel moeite, en zonder dat dit nadrukkelijk aan hen wordt onderwezen. Als kinderen echter aanhoudende problemen ondervinden met taal en spraak, kan er sprake zijn van een ontwikkelingsstoornis. Hoewel de onderliggende genetische oorzaken van spraak- en taalontwikkelingsstoornissen vaak complex zijn, is er in sommige gevallen een monogene oorzaak: één variant in een specifiek gen veroorzaakt de ontwikkelingsstoornis.

FOXP2 was het eerste gen dat geassocieerd werd met een monogene spraakontwikkelingsstoornis. In een grote familie had de helft van de familieleden, afkomstig uit verschillende generaties, problemen met spraakontwikkeling. De aangedane familieleden hadden allemaal een missense variant in *FOXP2*, die de functie van het gen verstoorde (uitgelegd in **Hoofdstuk 1**, **2** en **6**). Na deze ontdekking zijn meer individuen beschreven met ernstige spraakontwikkelingsproblemen en een genetische oorzaak in *FOXP2*. De studies die hierop volgden hebben verschillende *in vitro* en *in vivo* modelsystemen gebruikt om de effecten van etiologische varianten in *FOXP2* beter te begrijpen, en de functies van FOXP2 in relatie tot het brein bloot te leggen. Een overzicht van deze studies is beschreven in **Hoofdstuk 6**.

Hoewel een combinatie van *next-generation sequencing* methoden en prioriteren van *de novo* en/of bi-allelische varianten heeft geleid tot een sterke stijging in het aantal genetische diagnoses voor Mendeliaanse hersenontwikkelingsstoornissen, waaronder verstandelijke beperking en autisme, staat het gebruik van deze strategieën voor het diagnosticeren van spraak- en taalontwikkelingsstoornissen nog in de kinderschoenen. In **Hoofdstuk 1** en **2** van deze thesis beschrijven we de huidige stand van zaken in het veld van de genetica van spraakontwikkelingsstoornissen. Omdat er nog geen duidelijke consensus is over de criteria voor het classificeren van spraak- en taalontwikkelingsstoornissen, is het lastig om cohorten te vormen van individuen met een spraakontwikkelingsstoornis bij wie de klinische kenmerken nauwkeurig in kaart gebracht zijn. Desondanks hebben een aantal fenotype-gedreven studies *next-generation sequencing* strategieën gebruikt voor het identificeren van genetische varianten met een sterk effect. Opvallend is dat de genen die geassocieerd kunnen worden met Mendeliaanse spraakontwikkelingsstoornissen vaak regulatoire genen zijn die tot expressie komen tijdens de vroege stadia van de ontwikkeling van het menselijke brein.

Wanneer genen geassocieerd met spraakontwikkelingsstoornissen nader uitgewerkt worden in gen-specifieke studies door meer mogelijk-pathogene varianten te analyseren in deze genen, wordt er vaak een breder hersenontwikkelingsfenotype ontdekt (**Hoofdstuk 2**). In bijvoorbeeld *CHD3* werd in eerste instantie een *de novo missense* variant beschreven in een kind met een spraakontwikkelingsstoornis. Echter, toen een internationale samenwerking van onderzoekers 35 andere individuen identificeerden met een *de novo CHD3* variant, bleek het fenotype geassocieerd met *CHD3* te variëren van milde ontwikkelingsachterstand tot ernstige verstandelijke beperkingen, zonder een duidelijke genotype-fenotype correlatie. In veel gevallen blijft het onduidelijk wat de onderliggende oorzaak is van de variabiliteit in fenotypes die we zien tussen individuen met een pathogene variant in hetzelfde gen.

In deze thesis hebben we bekeken hoe we verschillende typen cellulaire functionele experimenten kunnen gebruiken om meer informatie te verzamelen over de moleculaire effecten van genetische varianten in spraak- en hersenontwikkelingsstoornissen. We hebben dit gecombineerd met objectieve fenotypische analyses, om zo het onderliggende moleculaire landschap van fenotypische variabiliteit beter te begrijpen.

In **Hoofdstuk 3** gebruikten we een combinatie van fenotypische en cellulaire functionele studies om pathogene varianten in *SATB1* te associëren met twee afzonderlijke klinische beelden. *SATB1* codeert voor een transcriptiefactor die de chromatine kan organiseren, en die deel uitmaakt van het *interactome* van FOXP2. Voorafgaand aan ons onderzoek was *SATB1* echter nog niet geassocieerd met een ziektebeeld. De ontdekking van een *de novo* eiwit-truncerende variant in *SATB1* in een kind met een verstandelijke beperking en ernstige spraakontwikkelingsproblemen, zette ons aan om een gen-specifieke studie te starten om meer individuen met hersentonwikkelingsfenotypes te identificeren met (*de novo*) varianten in dit gen. We vonden 42 individuen met een ontwikkelingsachterstand en *missense* en eiwit-truncerende varianten in *SATB1*. We brachten de functionele effecten van een significant aandeel van de gevonden varianten in kaart met moleculaire experimenten. Met continue cellijnen onderzochten we de effecten van *SATB1 missense* varianten op de co-lokalisatie met het DNA, de transcriptionele activiteit en de eiwitimmobiliteit. Door deze functionele data te koppelen aan objectieve computer-gebaseerde analyses van de klinische kenmerken die geassocieerd waren met *SATB1 missense* en eitwit-truncerende varianten, vonden we een genotype-fenotype correlatie op basis van het type variant. *SATB1 missense* varianten die clusteren in de domeinen belangrijk voor binding aan het DNA resulteren in sterkere transcriptionele activiteit, en leiden tot ernstige hersenontwikkelingsstoornissen met verstandelijke beperking, epilepsie, en in sommige gevallen voortijdig overlijden. In tegenstelling, eiwit-truncerende varianten in *SATB1* geven een *loss-of-function* effect en konden worden geassocieerd met een mildere hersenontwikkelingsstoornis.

Vervolgens bekeken we het *CHD3* gen. Tot zo ver hadden de meeste in de literatuur beschreven individuen met een *CHD3*-geassocieerde ontwikkelingsachterstand *de novo missense* varianten. In **Hoofdstuk 4** beschrijven we echter 21 families met een index met een hersenontwikkelingsstoornis en een overgeërfde *CHD3* variant afkomstig van een gezonde of mild aangedane ouder. Acht van deze gevallen betrof een eiwit-truncerende variant. De indexen in deze families hadden allemaal fenotypische eigenschappen die overlapten met het fenotypische beeld dat is beschreven voor de *CHD3*-geassocieerde ontwikkelingsstoornis. Met het gebruik van objectieve computer-gebaseerde analyses van klinische data toonden we aan dat de fenotypes van de indexen met een overgeërfde *CHD3* variant niet significant verschilden van de fenotypes geassocieerd met *de novo CHD3* varianten. Daarentegen waren de fenotypes van de indexen met een overgeërfde *CHD3* variant wel significant verschillend van de fenotypes van de ouders met deze variant. Deze resultaten laten zien dat de overgeërfde varianten variabel tot expressie komen. Om de hypothese te testen dat een *loss-of-function* variant in een gezonde of mild aangedane ouder mogelijk kan worden gecompenseerd door verhoogde expressie van het wildtype allel, groeiden we continue cellijnen afkomstig van een index, de moeder en een grootmoeder die de variant hebben, en een zus die de variant niet heeft. In deze cellen bekeken we de hoeveelheid *CHD3* transcript en CHD3 eiwit. Ondanks dat we konden bevestigen dat de *CHD3* variant leidde tot lagere expressie van het transcript en eiwit door *nonsense-mediated decay*, vonden we geen bewijs dat compensatie van de expressieniveaus een mogelijk onderliggend mechanisme kan zijn voor de variabele expressiviteit in deze familie.

In **Hoofdstuk 5** richtten we ons op de functies van *CHD3* tijdens de menselijk hersenontwikkeling, om zo de impact van genetische varianten in dit gen in hersenontwikkelingsstoornissen beter te begrijpen. Met behulp van CRISPR-Cas9 om wijzigingen aan te brengen in het DNA, groeiden we geïnduceerde pluripotente stamcellen met heterozygote en homozygote *CHD3 loss-of-function* allelen. Deze cellijnen differentieerden we vervolgens in cerebrale organoïden. Cerebrale organoïden zijn een driedimensionaal celkweek model dat gelijkenis vertoont met vroege stadia van de menselijke foetale hersenontwikkeling. Met het gebruik van openbare genexpressie data en onze gekweekte cerebrale organoïden, lieten we zien dat de expressie van *CHD3* laag is in neuronale *progenitor* (voorloper) cellen tijdens vroege stadia van ontwikkeling, maar dat de expressie toeneemt in de loop van de tijd, met hoge *CHD3* expressie in volwassen neuronen. We gebruikten een combinatie van immunohistochemie, *bulk* en *single-cell transcriptomics* om de effecten van een verlies van *CHD3* expressie te bestuderen. We ontdekten dat een verminderde *CHD3* expressie resulteert in een verhoogde expressie van genen die belangrijk zijn voor het onderhouden van neuronale progenitor cellen, terwijl genen belangrijk voor volwassen neuronale celtypen een verlaagde expressie vertonen. Consistent met deze bevinding, zagen we een verschuiving in de samenstelling van celtypen in de organoïden. Op dag 57 hadden organoïden met een *CHD3 loss-of-function* allel relatief meer neuronale *progenitors*, terwijl het aandeel neuronen kleiner was. Deze resultaten laten zien dat CHD3 een belangrijke regulator is van neurogenese, door het beheren van de balans tussen neuronal *progenitor* cellen en neuronen. Toekomstige studies die zich richten op de *CHD3*-geassocieerde ontwikkelingsstoornis zullen verder onderzoeken hoe deze functies verstoord worden door aetiologische varianten in dit gen.

Samengevat, **Hoofdstuk 3**, **4** en **5** laten zien hoe verschillende typen cellulaire modelsystemen gebruikt kunnen worden voor het vaststellen van genotype-fenotype correlaties, het onderzoeken van de onderliggend mechanismen van variabele expressiviteit/gereduceerde penetrantie en het ontdekken van nieuwe fysiologisch-relevante functies van genen. In deze studies waren cellulaire functionele experimenten cruciaal voor het bevestigen van de pathogeniteit van varianten, voor het begrijpen van de functies van genen, en de precieze effecten van genetische varianten op genfunctie.

Nieuwe ontwikkelingen in cel-gebaseerde kweekmethoden, waaronder *deep mutational scans* die duizenden genetisch varianten tegelijkertijd functioneel kunnen karakteriseren, en geautomatiseerde methoden voor het kweken van driedimensionale stamcel kweeksystemen, bieden perspectief op het combineren van *high-throughput* functionele karakterisatie met fysiologisch-relevante celmodellen (**Hoofdstuk 7**).

Zulke vooruitgangen zullen het mogelijk maken om de effecten van genetische varianten in een gen beter in kaart te brengen voor een veelheid aan eiwitfuncties. Zulke data zullen uiteindelijk de interpretatie van genetische varianten verbeteren, en de kennis over fenotypische variabiliteit op moleculair level verbreden.

Acknowledgements

The development of the brain is a complex process. While it starts out with a bunch of pretty undefined cells, already in a relatively short bit of time these cells divide, obtain specific functions, and, very importantly, make connections with each other, eventually giving rise to a fully functional organ. Although an analogy between brain development and doing a PhD may be a bit overstretched (yes, I actually agree), as much as cells are interdependent to make development a successful process, I, likewise, have been relying a lot on the people around me to be able to finish my PhD and wrap up my thesis – and all of them deserve a big thanks!

First of all, **Simon**, thanks for the incredible amount of freedom and trust you have given me during the last couple of years. I have been very lucky to do my PhD in such a well-equipped and well-funded place, and I feel you have always given me all the opportunities to let my creativity and curiosity lead the way. While in some early stages of my PhD that freedom could feel somewhat overwhelming, having had so much control over my own projects has made me, in the end, into an independent and confident researcher (I think). I really appreciate all your feedback on the conceptual aspects of my manuscripts, as well as your elegant ways to improve my writing without changing the messages I tried to get across. And thanks so much for hosting the New Year's parties and summer BBQs, they were a lot of fun!

Lisenka, thanks for adopting me (kind of). I really admire your scientific creativity and versatility, and I think you are a very supportive and motivating person. I remember your stimulating energy and enthusiasm during our first SATB1-update meeting very well, and when I was in need of an external supervisor to guide me through my PhD, you were the very first person that came to mind. Your postcard to celebrate the publication of Elke’s and my favourite project, for lack of any possibilities to go out for a drink, was such a nice surprise. I am very glad that we got to work together and I appreciate all your contributions.

My thesis was evaluated by the members of the Manuscript Committee **Hans**, **Silvia** and **Thomas**. I am very thankful that you accepted to be part of the committee, and for your time reading my chapters.

Without teaming up with our collaborators from the Human Genetics department in the Radboudumc, I would not have been able to do many of the research presented in this thesis. **Tjitske**, thanks for your co-supervision on some of my studies, and in particular for your input on the clinical side of things. **Christian**, thanks for brainstorming with Elke, Jet and me to find creative approaches to classify phenotypic differences and trying to understand genetic explanations for variable expressivity. **Lex**, I am very happy that you have been able to help out with some of the phenotype work of the projects – you have always been very helpful and so quick with your analyses.

Some of the collaborations crossed borders as well. **Wieland**, thanks for hosting me in your lab in Dresden. And **Michael**, thanks for spending so much time teaching me all the ins and outs of culturing brain organoids. **Alexandre**, it was very nice that we got to discuss the SATB1 project in person in Nijmegen, and your input on the manuscript was very valuable.

I would also like to thank all the people I have worked with in the past.

Buz, thanks that you gave me the opportunity to come all the way from across the Atlantic to make my first little steps in a research lab. **Christine** and **Anthony**, thanks for making me feel so at home in Atlanta. Lending me your bike allowed me to exhibit my most Dutch-like behaviour in the States, cycling every day to work (and I think the lab at Emory is the only place I ever encountered where I was allowed to bring my bike into the lab and park right next to my bench). I am happy that we are still in touch from time to time and hope that we see each other soon on one of our travels. **Diane**, you supervised my very first research project during that time, and I think you planted the seed for my enthusiasm and excitement for culturing cells (as well as for the beautiful Smokies). In the cell culture lab, I still do most things how you taught me to. Your supportive words and belief in my research abilities really gave me the motivation to continue with a research programme, and made me start thinking of doing a PhD.

Pela, during my next internship, with you as supervisor, I have learned so incredibly much. For sure one of the reasons to come back to the MPI was the amazing guidance I got from you, and I would have loved to continue working together. As my daily supervisor during that time, thanks **Elliot**, for your patience answering all my questions, and walking into the lab with me over and over again to look at my DNA gels and Western blots. Also, that internship would not have been the same with the other former Protein Group members that were around at that time. **Sara** and **Moritz**, you contributed to such a safe and open group feeling and atmosphere.

When I moved to Oxford for my second internship, I focused completely on animal models. Thanks **Pete**, for teaching me the secret tricks of tissue sectioning and preparing pretty fluorescent stainings – those skills were super useful during my PhD. **Dasha**, you made that time such an interesting experience, taking me to formal dinners and spending hours talking about food and the peculiar differences between Dutch and Russian culture.

Maggie and **Elke**, if I have to name people that pulled me through the difficult times of doing a PhD, it would be the two of you for sure.

Maggie (dr. Wong!), your systematic and structured approaches in the lab (sometimes considered a bit ‘strict’) have helped so much to get all the things up and running that I needed for my projects. For months, we worked together as a close team, sharing the workload and ticking off all our goals. With the 80s playlist on maximum volume we eventually managed to get our cell models working. We have had countless coffee breaks, talking about the very details of each other’s experiments, but also about completely random things (you are very good at coming up with the most unexpected topics of conversation). The trip to San Diego together, to present our results, was the best reward for all the hard work. While we have quite different working styles, we also think so much alike, making it very pleasant to have you as a colleague. But aside from that, you are also a good friend who always has time for a good chat.

Elke, je bent één van de beste onderzoekers die ik ken. Vol energie, oog voor detail, transparant, eerlijk en direct, zorgvuldig. Het is ontzettend prettig om met je samen te werken, en ik denk dat we heel erg op één lijn zitten wat betreft aanpak van projecten. Ons gezamenlijke onderzoek heeft me zeker de boost gegeven om door te zetten en vol te houden, ook op momenten waar ik soms niet helemaal zeker was of ik door wilde gaan met mijn PhD. Samen schrijven aan ons paper tijdens corona-tijd in jullie weelderige bostuin (niet alleen vanwege het mooie lenteweer, maar ook omdat je huis soms een dak of vloer miste) was super leuk en gezellig – en het was waarschijnlijk nooit zo gelopen als we niet allemaal verplicht thuis hadden moeten zitten. Ik ben blij dat we naast goede collega’s ook vrienden zijn geworden, en weet dat de zonnebloempitjes altijd klaar staan!

Jet, ons project was er één waar we begonnen met heel veel verschillende ideeën en plannen, maar waar we telkens opties moesten blijven wegstrepen omdat er toch niet helemaal uitkwam wat we ervan hadden gehoopt. Desondanks hebben we doorgezet, en ik denk ik dat we een hele mooie studie hebben neergezet! Ik bewonder hoe je nu de stap van klinisch onderzoek naar cellen en organoïden hebt gezet – absoluut uitdagend, maar ook met de mogelijkheid om je creatief uit te leven in het lab. Leuk dat we nu met meerdere groepen op de campus onze ervaringen kunnen delen, en wie weet zit er ooit nog weer eens een leuke samenwerking in.

Terwijl ik tijdens mijn Master stage jou de kneepjes van moleculair biologie bij mocht bijbrengen, heb jij mij in de afgelopen jaren de wereld van de genetica ingeleid. **Lot**, in al je bescheidenheid ben je het vast niet met me eens, maar dit boekje ligt er voor een heel groot deel ook dankzij jou. Onze overenthousiaste brainstormsessies over interessante nieuwe experimenten en projecten, en jouw netwerkskills, hebben er uiteindelijk voor gezorgd dat twee, in eerste instantie *side projects*, uiteindelijk uitgroeiden tot mooie publicaties en hoofdstukken in mijn thesis. Ik bewonder je enthousiasme voor wetenschap en het gemak waarmee je je in een interdisciplinair veld tussen kliniek, genetica en moleculaire biologie beweegt. Je bent echt een voorbeeld voor me, en hoop dat je je plek gaat vinden in de wetenschap.

Ezgi and **Willemijn**, you have been a little bit my guinea pigs during my PhD projects. Good supervision is not easy, and it has been a learning process. I hope I have not been too energetic, chaotic or unstructured, and I have given you the guidance you needed. But I am glad that both of you have found your way in science. Ezgi, it was a lot of fun working together with you. And just so you know, I am still using the mug you brought me from London every day! Willemijn, jouw energielevels zijn bewonderingswaardig. Zoals je in het lab het ene na het andere experiment (soms parallel) inplant, zo lijkt ook je leven buiten werk behoorlijk volgeboekt met sport, vakanties, etentjes – waar je dan ook graag met een kopje thee je ervaringen over deelt (ik denk dat ik nu bijna alles weet over klimmen). Altijd nieuwsgierig en kritisch, sta je vaak naast mijn bureau, en schroom je niet om mede te delen als je het niet eens bent met de voorgestelde aanpak. Je werk tijdens je stage heeft ontzettend bijgedragen aan mijn PhD, en ik vind het heel tof dat we ook na je stageperiode samen zijn blijven werken.

I also would like to thank all other members of the L&G department and the NVC group. **Arianna**, your organisation skills keep the lab running! **Gökberk**, thanks for all the good, sometimes deep, discussions. **Else**, je hebt altijd zoveel goede vragen en ideeën. **Dick**, bedankt voor al je hulp met het doorspitten van de UK biobank data. **Roos K.**, met de 'Pool-side Grooves'-playlist en dromen over een cocktail op het strand kwamen we die lange middagen in het lab lekker door. **Ine**, thanks for making me 'proper coffee' from time to time. **Lukas**, great to have a gym buddy! **Sabrina**, **Karthik**, **Cleo, Jelle**, **Midas**, **Paolo**, **Elpida**, **Janine**, **Nienke**, **Mubeen, Ellen**, **Mariska**, **Jitse**, **Clyde**, **Beate**, **Zhiqiang**, **Barbara**, **Fenja**, and everyone else who has been around in the last 4,5 years, thanks for the good times at the MPI.

Daarnaast zou mijn werk ook niet mogelijk zijn geweest zonder alle ondersteunende hulp van **Jasper**, **Jurgen** en **Martina**. Jasper, altijd een geruststellend gevoel dat er iemand is die ervoor zorgt dat het lab zo veilig mogelijk is. Jurgen, bedankt voor al die 'super urgente' bestellingen en verzendingen. Martina, jouw deur staat altijd open, en ontzettend fijn dat je zo vaak hebt meegedacht met het organiseren van stageprojecten, conferentiereizen, maar ook alles wat kwam kijken bij het plannen van mijn verdediging.

And not to forget, **Kevin**, in moments where I was unsure how to set up my supervisory checkpoint meetings, or doubting where to start with the thesis and defense, your IMPRS manuals and checklists were a great help!

Of course, my friends and family have played a very important supportive role in the last couple of years as well.

Cansu and **Gijs**, initially bonding over our shared enthusiasm for science, spending hours together in the library, furiously discussing our study materials and challenging each other – often interrupted by a dinner-break with our famous 'exam-pasta' – we have been taking very similar steps in life in the last few years. Together doing the Master's programme, going abroad for our internships, starting a PhD around the same time. Having friends that are also in academia makes it easier to share your experiences, worries and frustrations, as well as to celebrate the achievements. At the start of our PhDs, we may have been a bit too enthusiastic in that regard, organizing our 'journal club nights' to get updated about each other's projects. But even though those were not so successful in the end, we have always kept up-to-date with each other's progress, and I appreciate how we have supported each other. Now we are all finishing up our PhDs, our lives will start diverging more. Cansu, I really liked our relaxing weekend-walks along the Waal in Nijmegen, they were a great way to catch-up on each other's lifes and brainstorm about what the future holds. With a busy family life combined with work, it is a bit more challenging to meet up regularly now, but when we manage to, our coffee breaks and dinners are a lot of fun. Gijs, terwijl ik uiteindelijk terugkrabbel, ga jij het dan echt doen, een paar jaar naar de VS. Zet 'm op daar, en bewaar wat van je (weinige) vakantiedagen om me je nieuwe stek te laten zien.

Katie, thanks for all your support during the, sometimes difficult, times of my Bachelor and Master internships, and later on when I started with my PhD. You have always helped me to believe in myself as a scientist and pushed me to make the right choices.

Emma and **Samara**, after all the adventures in Oxford and the amazing summer trip in Italy, we stayed in contact when we moved back to the Netherlands, all three of us starting a PhD. Although only meeting up occasionally, it has always been a lot of fun hanging out together, and pretty relieving to share our PhD worries and struggles, realizing that we were all experiencing similar things. It is great that we can now celebrate the completion of each other's PhDs, and I wish you all the best in your new non-academic jobs. Let's keep in touch!

Roos S., **Sylvia** en **Arie**, zo nu en dan een leuke wandeling, een lange Anno-avond of een pannenkoekenbakmiddag, we zien elkaar dan niet zo vaak, maar onze weekendjes hebben altijd voor de welkome afleiding gezorgd.

Shai, onze gesprekken over studie en wetenschap hebben plaats gemaakt voor de belangrijkere thema's in het leven – en ik ben blij dat onze vriendschap al zo lang standhoudt.

Ook de rest van mijn vrienden in Nijmegen (of inmiddels niet meer), en familie in Zeeland, bedankt voor al de gezelligheid en steun de afgelopen jaren.

Mam en **pap**, bedankt voor het altijd hebben gekoesterd van mijn interesses en nieuwsgierigheid. Wandelende takken op mijn kamer, een kristallenverzameling, een (in mam's ogen mega-) aquarium op mijn slaapkamer, niks was te gek. Met mijn liefde voor de natuur en alles wat beweegt, was biologie dan ook de meest logische keuze als studie, en dat heeft uiteindelijk goed uitgepakt. Het is altijd fijn om thuis te komen en even uit te blazen tijdens een mooie strandwandeling of met een bak koffie en een goed gesprek in de tuin.

And dear **Lucía**, I am so happy we met. You make life light and easy. Although we both love science, and for sure chat about our work from time to time, we share so many values and interests that it feels like we can talk forever. I love coming home to our little jungle house every day and forget about all the stressy things in life. Thanks for keeping an eye on my work-life balance, thanks for listening to me and supporting me throughout, and thanks for making life so much fun!

Curriculum vitae

Joery den Hoed was born on the 17th February 1993 in Vlissingen, in the Netherlands. After finishing his pre-university secondary education in the neighbouring city of Middelburg, he moved to Nijmegen to study biology at the Radboud University in Nijmegen in 2012. He obtained his Bachelor of Science degree with a minor in ‘Medical Biology’ with the distinction ‘cum laude’ in 2015. During this period, he also participated in the Faculty of Science Honours programme, which provided him the opportunity to do a research internship in the laboratory of Prof. dr. H. A. Jinnah, at Emory University, Atlanta, GA, USA in 2015, focusing on cellular models of Lesch-Nyhan syndrome, a rare X-linked recessive condition. He then continued with the two-year Master of Science programme ‘Molecular Mechanisms of Disease’ at the Radboud University, which he completed in 2017 with the distinction ‘summa cum laude’. For this degree, he did an internship at the Max Planck Institute for Psycholinguistics in Nijmegen under dr. P. Derizioti and Prof. dr. S. E. Fisher in 2016, studying protein-protein interactions of transcription factors associated with developmental speech phenotypes. In late 2016, he moved to Oxford, UK, to work on his second internship project for his Master of Science programme in the laboratory of dr. P. L. Oliver at the University of Oxford, which involved the characterisation of mouse models of a family of proteins (TLDc proteins) associated with neurodegeneration and neurodevelopment. In September 2017, Joery then returned to Nijmegen, after he was awarded a four-year International Max Planck Research School (IMPRS) for Language Sciences PhD fellowship. There, he started his doctoral research in the Language and Genetics Department under the supervision of Prof. dr. S. E. Fisher, with the ambition to set up a new cellular model to study gene function in early human brain development. To learn this novel laboratory technique, he made a three-month visit to the group of Prof. dr. W. B. Huttner at the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden, Germany in 2018, with a Travelling Fellowship from the Company of Biologists, Cambridge, UK. In 2020, he was selected to attend the Human Brain Organogenesis workshop given by the laboratory of dr. Sergiu Pasça at Stanford University, CA, USA, to learn more about three-dimensional cell models of brain development. During his PhD work, he also closely collaborated with researchers from the Human Genetics Department at the RadboudUMC in Nijmegen. These trainings and collaborations led to the work that is part of this dissertation. Several of the studies included in this thesis were presented at international conferences, including the European Society for Human Genetics annual meeting in 2020 and the American Society for Human Genetics annual meeting in 2021. Joery continues his work on the functional characterisation of genetic variants in neurodevelopmental disorder, as a post-doctoral researcher in the Language and Genetics Department at the Max Planck Institute for Psycholinguistics.

List of publications

2022

Van der Spek, J.*, den Hoed, J.*, Snijders Blok, L., Dingemans, A. J., Schijven, D., Nellaker, C., ... & Kleefstra, T. (2021). Inherited variants in CHD3 demonstrate variable expressivity in Snijders Blok-Campeau syndrome. Genetics in Medicine (in press)

2021

Den Hoed, J., Devaraju, K., & Fisher, S. E. (2021). Molecular networks of the FOXP2 transcription factor in the brain. EMBO reports, 22(8), e52803.

Sutcliffe, D. J., Dinasarapu, A. R., Visser, J. E., den Hoed, J., Seifar, F., Joshi, P., ... & Jinnah, H. A. (2021). Induced pluripotent stem cells from subjects with Lesch-Nyhan disease. Scientific reports, 11(1), 1-15.

Castroflorio, E., den Hoed, J., Svistunova, D., Finelli, M. J., Cebrian-Serrano, A., Corrochano, S., ... & Oliver, P. L. (2021). The Ncoa7 locus regulates V-ATPase formation and function, neurodevelopment and behaviour. Cellular and Molecular Life Sciences, 78(7), 3503-3524.

Blok, L. S., Vino, A., den Hoed, J., Underhill, H. R., Monteil, D., Li, H., ... & Fisher, S. E. (2021). Heterozygous variants that disturb the transcriptional repressor activity of FOXP4 cause a developmental disorder with speech/language delays and multiple congenital abnormalities. Genetics in Medicine, 23(3), 534-542.

Den Hoed, J.*, de Boer, E.*, Voisin, N.*, Dingemans, A. J., Guex, N., Wiel, L., ... & Vissers, L. E. (2021). Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction. The American Journal of Human Genetics, 108(2), 346-356.

2020

Den Hoed, J., & Fisher, S. E. (2020). Genetic pathways involved in human speech disorders. Current Opinion in Genetics & Development, 65, 103-111.

2019

Tilot, A. K., Vino, A., Kucera, K. S., Carmichael, D. A., Van den Heuvel, L., den Hoed, J., ... & Fisher, S. E. (2019). Investigating genetic links between grapheme–colour synaesthesia and neuropsychiatric traits. Philosophical Transactions of the Royal Society B, 374(1787), 20190026.

2018

Den Hoed, J., Sollis, E., Venselaar, H., Estruch, S. B., Deriziotis, P., & Fisher, S. E. (2018). Functional characterization of TBR1 variants in neurodevelopmental disorder. Scientific reports, 8(1), 1-11.

* Contributed equally