Let the genes speak!

de novo variants in developmental disorders
with speech and language impairment

Lot Snijders Blok

Let the genes speak!
de novo variants in developmental disorders with speech and language impairment

The research presented in this thesis was carried out at the Human Genetics department of the Radboud University Medical Center in Nijmegen and the Language & Genetics Department of the Max Planck Institute for Psycholinguistics in Nijmegen, and within the Donders Graduate School for Cognitive Neuroscience (DGCN).

ISBN: 978-94-6284-308-0

Cover artwork: Lot Snijders Blok
Printed: ProefschriftMaken.nl

Let the genes speak!

de novo variants in developmental disorders
with speech and language impairment

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 vrijdag 15 oktober 2021
om 12.30 uur precies

door
Charlotte Snijders Blok
geboren op 24 april 1987
te Sneek

Promotoren:
Prof. dr. H.G. Brunner
Prof. dr. S.E. Fisher
Prof. dr. T. Kleefstra

Manuscriptcommissie:
Prof. dr. M.A.A.P. Willemsen (chair)
Prof. dr. H. van Esch (KU Leuven, België)
Dr. G.M. van Woerden (Erasmus MC)

Let the genes speak!

de novo variants in developmental disorders
with speech and language impairment

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 Friday, October 15, 2021
at 12.30 pm

by
Charlotte Snijders Blok
born on April 24, 1987
in Sneek, the Netherlands

Supervisors:
Prof. dr. H.G. Brunner
Prof. dr. S.E. Fisher
Prof. dr. T. Kleefstra

Manuscript Committee:
Prof. dr. M.A.A.P. Willemsen (chair)
Prof. dr. H. van Esch (KU Leuven, Belgium)
Dr. G.M. van Woerden (Erasmus MC)

Contents

1

Chapter 1

General Introduction

Developmental speech and language disorders

Normal and abnormal speech and language development

Most children acquire speech and language skills quickly in the first few years of life, in a way that seems almost effortless. For example, a typical two-and-a-half year old toddler can already speak over 1000 words and understand even more[1], without having received any formal teaching. Interestingly, the acquisition of spoken language appears to be specific to our species; no other living animal demonstrates this capacity[2]. In humans, interaction via language is crucial for socio-emotional development and academic performance[3]. The ability to acquire speech and language skills forms the basis for many other areas of development.

In some children, the development of speech and language skills is not that straightforward and uncomplicated. Such problems may be first recognized at toddler age, when a child starts talking later than other children or has difficulties understanding language. However, language impairments can also become apparent only at a later age. Language problems at a young age might be short-lived; a subset of so-called 'late talkers' or 'late bloomers' catch-up completely and do not have any difficulties in the long term[4]. Nonetheless, in a significant proportion of children, deficits do not resolve spontaneously[5]. If the problems persist and are not the obvious result of another primary cause, such as deafness, traumatic brain injury or insufficient language exposure, the child might have a developmental speech or language disorder, that is a neurobiological disorder characterized by impaired speech and/or language development.

Terminology and categorization

Although the existence of developmental speech and language disorders is not questioned, there are many different ideas on terminology and criteria used to define them. Especially for language impairments, several different names have been used over time, including *specific language impairment*, *developmental dysphasia* and *developmental language disorder*. Because language disorders are at the interface of different medical and educational professions, inconsistent naming and categorization can lead to confusion and inconsistency in clinical and research practices[6,7].

Two well-known diagnostic manuals contain definitions for developmental speech and language disorders: the Diagnostic and Statistical Manual of Mental disorders (DSM) and the International Classification of Diseases (ICD). Importantly, both classification systems do not always reflect the diagnoses given or the terminology used in practice by speech and language therapists[7]. The current version of the DSM, DSM-5, contains the category *communication disorders*, with the following subcategories: *language disorder, social (pragmatic) communication disorder, speech sound disorder* and *childhood-onset fluency disorder (stuttering)*[8]. In DSM-5, *verbal dyspraxia* is listed in the subcategory *speech sound disorder*. The ICD-10 specifies the category *specific developmental disorders of speech* *and language* which contains several subcategories (e.g. *phonological disorder, expressive language disorder, mixed receptive-expressive language disorder* and *apraxia*).

Although the terminology of communication disorders represents an ongoing topic of debate, results of a recent Delphi study, performed within the international and multidisciplinary CATALISE consortium, recommend the adoption of the consensus English term *developmental language disorder* (abbreviated as *DLD*)[9]. In the Netherlands, since 2014 the same label (in Dutch: *taalontwikkelingsstoornis*, commonly abbreviated as *TOS*) has been widely used by speech therapists, as well as all care and education organizations involved[10]. For developmental speech disorders, the term commonly used in the literature is *speech sound disorders* (in the Netherlands: *spraakontwikkelingsstoornis; developmental speech disorder).* These labels are used as umbrella terms for a heterogeneous category of children who have errors in speech production that significantly impact intelligibility.

In this thesis, the term *developmental speech and language disorders* is used to refer to children diagnosed with a DLD and/or a speech sound disorder. This encompasses all developmental speech and language disorders as referred to in DSM-V and ICD-10, with the exception of stuttering (fluency disorders). Figure 1 shows a schematic overview of the terms and classification as used in this thesis.

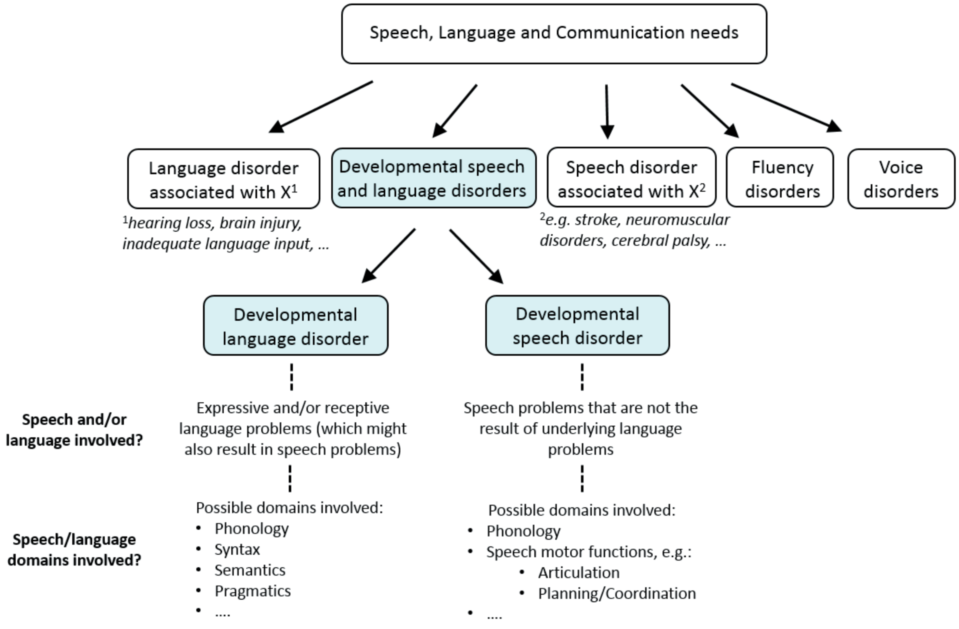

Speech, Language and Communication needs
Language disorder associated with X¹
Developmental speech and language disorders
Speech disorder associated with X²
Fluency disorders
Voice disorders
¹hearing loss, brain injury, inadequate language input, ...
²e.g. stroke, neuromuscular disorders, cerebral palsy, ...
Developmental language disorder
Developmental speech disorder
Speech and/or language involved?
Expressive and/or receptive language problems (which might also result in speech problems)
Speech problems that are not the result of underlying language problems
Speech/language domains involved?
Possible domains involved:
• Phonology
• Syntax
• Semantics
• Pragmatics
• ....
Possible domains involved:
• Phonology
• Speech motor functions, e.g.:
• Articulation
• Planning/Coordination
• ....

Figure 1: A framework for the classification of developmental speech and language disorders

Prevalence of developmental speech and language disorders

The prevalence of developmental speech and language disorders amongst children is not exactly known, and can vary widely depending on, for example, the age of the child and the diagnostic criteria that are used. The most cited study that estimated prevalence was a USA-based population study of monolingual children (i.e. children who were proficient in only one language) around five years of age, in which a DLD was identified in ~7% of the sample [11]. Another study in Sweden reported that the vast majority of the 5-6% of children who screened positive in a language screening at 2.5 years of age, later fulfil criteria for DLD when assessed by a speech therapist[5]. In the Netherlands the precise number of children with developmental speech and language disorders is unknown, as no systematic registration is in place. It is known that there we 84,000 children under ten years old with DLD who received speech therapy from a local speech therapist in 2017, while 4,900 such children received multidisciplinary treatment in a Cluster 2 setting[12]. Systematic terminology, clear diagnostic guidelines and diagnostic registration systems are needed in order to provide more accurate estimates of the prevalence of developmental speech and language disorders and of the various subtypes within this heterogeneous group of disorders.

Clinical features

Children with developmental speech and language disorders can have problems with various aspects of speech and/or language, with differing levels of severity. While the relationships between speech and language are complex, we here present a general framework that is helpful for approaching heterogeneity within speech and language disorders. In general, language is defined by the complex ability of using and understanding words and sentences (independent of the modality used to express those), speech is the process of shaping meaningful vocalisations.

According to this framework, on a language level, different facets and associated problems can be categorized[9,10], including but not limited to the following:

- *Phonology*: the organization of speech sounds into categories. An example of this is phonological awareness problems; when a child has problems with correctly identifying units of spoken language.
- *Syntax* and *morphosyntax*: the rules and principles that determine the structure of sentences and how words within a sentence are modified depending on the context. A child can have both expressive problems with syntax (e.g. use of short sentences only or problems with correct conjugation of verbs), but receptive language impairments affecting syntax can also occur (e.g. not understanding different sentence structures correctly).
- *Semantics*: the area concerned with the meaning of words and sentences. A child can for example have limited knowledge on the meaning of words, or word finding problems.

- *Pragmatics:* a topic concerning how language is used appropriately in a given context. Examples of pragmatic language problems include saying unrelated or inappropriate things in a certain situation, or trouble with understanding social communication cues. Of note, pragmatics and semantics interact in important ways; a child might misunderstand the intended meaning of a sentence because he or she lacks the necessary pragmatic abilities and come up with the wrong inference.

On a speech level, different types of speech problems and disorders can be characterized as well[4,13]. Using speech observation and specific speech tasks and tests, a speech therapist can distinguish between different underlying processes that lead to speech problems. Different classification systems of speech disorders are being used, based on e.g. etiological and/or linguistic approaches[13]. Examples of problems that can be defined on a speech level are:

- *Phonological problems:* These happen when a child does not understand and/or correctly use the rules for sounds of a language, leading to phonological error patterns, see also the description in previous paragraph (phonological problems can influence speech and language processes).
- *Phonetic articulation problems*: These include substitutions or distortions of the same sound in different situations (e.g. in isolation, verbs and sentences; during imitations and spontaneous speech). Articulation problems lead to mispronunciation (e.g. lisping).
- *Motor speech problems*: These include problems with planning, programming and execution. For example, childhood apraxia of speech (CAS) is a developmental speech disorder that is the result of planning and coordination difficulties.

There is no international consensus about inclusion criteria for developmental speech and language disorders and diagnostic test procedures, and a lot of variation exists not only between countries but also within countries[14]. Using a discrepancy criterion for IQ (referred to as *cognitive referencing)*, in which non-verbal IQ should be higher than the language scores, does not seem to contribute to better diagnosis[10]. In short, one can infer a developmental speech or language disorder if there are significant language and/or speech impairments, in the absence of a clear primary cause such as hearing loss, brain injuries, low non-verbal intelligence or language deprivation, and if problems endure into middle childhood and beyond, with a serious impact on everyday social interactions[9,10].

Specific language and speech problems can be present in distinct combinations, and with different problems being most prominent. Over the years, many attempts have been made to create separate profiles or subgroups, but no agreement has been reached on the definition of clear clusters or profiles[9,10]. All in all, it is most important to realize that even with consistent terminology and categorization, developmental speech and language disorders will always be heterogeneous disorders that include a wide range of speech and language-related problems, which may vary from one child to another, or at different points of development[5,13].

Diagnosis, treatment and prognosis

Systematic screening is important to identify children with possible speech and language impairments at a young age. Although some young children might catch up and not have a developmental speech or language disorder later in life, it is important to start monitoring and instigate treatment as early as possible. A wait-and-see approach can place young children at risk for life-long problems[15]; late talkers are generally thought to need at least watchful surveillance[16].

A developmental speech or language disorder diagnosis is made by a speech therapist, ideally in consultation with a multidisciplinary team, to not only assess speech/language capacities but also medical, psychological and environmental factors[10]. Testing of speech and language capacities is usually performed using a combination of observation, normalized speech and language tests, and spontaneous language analysis[10]. The testing procedure and tests chosen depend on many different factors, including but not limited to the age and capacities of the individual being tested, the language spoken and the test characteristics (e.g. the test reliability and validity). Examples of currently used tests in the Netherlands are the Schlichting tests for Language Comprehension[17] and Language Production-II[18], the Dutch version of the Peabody Picture Vocabulary test-III[19], the Clinical Evaluation of Language Fundamentals (CELF-4-NL)[20] and the Dutch Nonspeech Test (NNST)[21]. In addition to direct assessments, scoring checklists or scales are often used that can be filled in by parents and/or therapists such as ICS-NL and CCC. After thorough assessment, a subsequent therapeutic plan can be devised, depending on the profile of problems and their severity, as different types of therapies and treatment strategies exist[4]. In short, language therapy can be defined as specific intensive stimulation to improve aspects of language[22], in order to reach goals not just focused on the disorder but on improving communication and participation[10]. Most children receive treatment from a local speech therapist. In more severe cases specialized ambulatory support or education in a special education setting for children with communication difficulties might be necessary.

The long-term prognosis of a developmental speech or language disorder can vary. Some children do not have any speech/language problems at a later age[10], but for many children the problems persist later in life and can significantly affect social and societal participation[23,24].

Other neurodevelopmental disorders: clinical overlap and differences

Developmental speech and language disorders form a subgroup within the larger group of neurodevelopmental disorders with regard to categorization in the DSM-5. In fact, developmental speech and language disorders often co-occur with other developmental disorders or difficulties, including Attention Deficit Hyperactivity Disorder (ADHD), motor impairments, developmental dyslexia and autism spectrum disorder[7,25]. The presence of other developmental disorders might complicate diagnosis and intervention strategies for speech and language impairments[7].

On the other hand, problems with speech and/or language can also reflect a broader underlying neurodevelopmental disorder. The best-known example of this may be the co-occurrence of language problems in the context of autism spectrum disorder. Verbal children with autism spectrum disorder can have language impairments that appear similar to those in children with DLD[26]. Speech and language deficits can also be observed in childhood epilepsy disorders[27]. Specifically, Landau-Kleffner syndrome is a disorder of the epilepsy-aphasia spectrum in which children with previously normal development show language regression with a typical profile, together with nocturnal EEG abnormalities and rare seizures[27,28]. Finally, speech and language problems can also occur in the context of broader cognitive impairments, as in intellectual disability.

In the early literature on speech and language disorders, much emphasis was put on the idea of specificity, e.g. a person could not be diagnosed with a DLD if this individual was also diagnosed with an autism spectrum disorder. Using exclusionary factors rather than inclusion factors has led to major issues in categorization of these disorders, and this approach does not fit with knowledge on underlying pathogenic mechanisms, a topic further discussed in Chapter 9 of this thesis.

Genetics of developmental speech and language disorders

Heritability

It has been known for a long time that developmental speech and language disorders tend to run in families[29]; a positive family history is a clear risk factor for developing a DLD[9,10]. Yet, familial occurrence alone does not prove a genetic origin. Shared environment could also lead to familial clustering of DLD. It is known that environmental factors play an important role in normal speech and language development, because after all, a child will not learn a language without being exposed to it[30]. However, it is not entirely clear what role environmental factors exactly play in the development of developmental speech and language disorders.

Twin studies provide one means for disentangling contributions of genetic and environmental factors, and can give insights into the heritability of a trait. Heritability is defined as the proportion of phenotypic variance that can be attributed to genetic differences. As monozygotic twins have virtually identical genomes, while dizygotic twins on average share about 50% of polymorphic alleles, the co-occurrence of a disorder in monozygotic twins compared with that in dizygotic twins enables estimation of its heritability. For developmental speech and language disorders, several twin studies have shown that heritability is high (at least 0.5)[31-33], although such estimates might vary depending on inclusion criteria and cut-off values for speech or language impairments used[34].

Thus, it is clear from family and twin data that developmental speech and language disorders are strongly influenced by genetics. In addition, genetic factors also make a strong contribution to normal variation in speech and language development, as assessed with a range of quantitative tests[35]. Nevertheless, it is important to point out that the concept 'language' itself is not encoded in our genes. Rather, our DNA and the genetic variance that it harbours are important in shaping the human brain in such a way that it is able to use speech and language, sometimes referred to as the 'language-ready' brain.

Inheritance patterns: multifactorial and Mendelian inheritance

The concept of heritability is useful for demonstrating that there is a genetic influence on a trait, but it does not reveal anything about the likely mode(s) of inheritance. In the context of speech and language disorders, it is important to distinguish between multifactorial and Mendelian inheritance patterns.

Multifactorial inheritance means that a disorder or trait is the outcome of multiple genetic and environmental factors. On a genetic level, this means that many DNA variations contribute to the risk of developing the disorder, and that different combinations of different common variants can give rise to the same phenotype. These variants may have a range of sizes at the genomic level, as well as variable modes of influence, but they are generally common polymorphisms, each with a small effect size. Most recognised in this context are single nucleotide polymorphisms (SNPs); single base pair changes found at a frequency of >1% in the general population. Disorders with a multifactorial inheritance pattern are referred to as complex diseases. For complex disorders, it is not possible to perform a single DNA test on an individual level and determine the exact risk of developing the disorder. In this case, it is necessary to compare extremely large groups of people with and without the disorder of interest, to give adequate statistical power for tracking down specific genetic risk loci (e.g. SNPs). It has now become standard practice to analyse such large samples for association using millions of SNPs spread throughout the genome - these types of studies are called case-control 'genome wide association screens' (GWAS). Beyond case-control designs, similar approaches can be used to test SNPs for association with quantitative traits (i.e. continuously varying phenotypes) again requiring very large sample sizes to give sufficient power for detecting small effect sizes via a GWAS design.

Studies of developmental speech and language disorders indicate that much of the underlying genetic architecture is indeed complex and multifactorial. Several GWAS efforts have been undertaken, and some interesting genetic loci have been identified[36-39], but these studies have often been underpowered and not all of them used a replication cohort to confirm findings. Moreover, independent replication studies of the strongest findings are still scarce. Larger systematic studies are needed to better understand the complex genetic background of developmental speech and language disorders.

On the other hand, there is Mendelian inheritance. A Mendelian (also known as monogenic) disorder involves a single genetic locus, and thus displays Mendelian inheritance; a transmission pattern that follows Mendel's principles. Typically, Mendelian disorders are caused by rare genetic variants with a large effect size. Examples of Mendelian inheritance patterns are autosomal dominant (as seen in e.g. Marfan syndrome), autosomal recessive (e.g. cystic fibrosis), X-linked dominant (e.g. Rett syndrome) and X-linked recessive (e.g. red-green colour blindness).

Over the past decade, an increasing number of reports have identified Mendelian causes of developmental speech and language disorders. It is currently unknown what proportion of cases of developmental speech and language disorders may be explained by Mendelian causes, and what proportion is multifactorial. Nonetheless, it is well established that studying Mendelian forms of disorder is a powerful way for gaining insights into aetiology. This thesis focuses specifically on Mendelian causes of developmental speech and language disorders, and the prospects of using recent developments in DNA sequencing technologies for advancing our understanding of their biology.

FOXP2: the first Mendelian link to developmental speech and language disorders

Until the introduction of new genetic techniques in the last decade, identifying the gene responsible for a disease of interest was generally very labour- and time-intensive. Often, a method termed *positional cloning* was used, in which the segregation of the phenotype of interest was studied in a large family, or in multiple families with the same phenotype. It was this approach that led to the discovery of *FOXP2,* the first gene to be implicated in a Mendelian form of speech and language disorder.

FOXP2 was originally identified in 2001, through genetic studies of a large three-generational family with CAS as the core phenotype, and a complex suite of deficits in expressive and receptive language with varying degrees of severity[40,41]. While the affected individuals did not have intellectual disability, they had on average lower nonverbal IQ than their unaffected family members[41]. Fifteen (about half) of the family members in different generations were diagnosed with this disorder, clearly fitting a Mendelian (autosomal dominant) inheritance pattern, facilitating a positional cloning approach. After using linkage mapping to determine the most likely chromosomal location of the putative etiological variant (chromosome 7q31.2)[42], and identification of an unrelated individual with CAS and a balanced chromosomal translocation disrupting this same region, *FOXP2* was established as the gene responsible[40]. A single base pair substitution within *FOXP2* turned out to perfectly co-segregate with the disorder in the family of interest[42]. In the years following publication of these first cases of *FOXP2* mutation, more children with variants disrupting the gene have been identified, including different missense and truncating variants, as well as microdeletions encompassing the whole locus[43]. Although speech and language problems are still the most prominent feature of the *FOXP2*-associated disorder, other neurodevelopmental features reported in individuals with this disorder include mild cognitive delays, behavioural abnormalities and motor development delays[43,44].

FOXP2 has since been the focus of interest for many studies of genetic pathways involved in speech and language. The *FOXP2* gene encodes the FOXP2 protein, a member of the forkhead-box family of transcription factors with important roles in regulating the expression of other genes[45], and is highly expressed in several regions of the developing brain, including the cortex, striatum, thalamus and cerebellum[46]. For example, it has been shown that mice with cortex-specific disruptions of *Foxp2* show deficits in social behaviour and cognitive flexibility[47,48], and disruption of *FoxP2* in area X (a brain region involved in vocal learning) affects song development and song production in zebra finches[49]. All in all, *in vitro* and *in vivo* studies on FOXP2 function have led to important insights beyond the *FOXP2*-associated human disorder[50].

There is no doubt that rare disruptive *FOXP2* variants can cause a developmental speech and language disorder, with CAS as a prominent feature. Nonetheless, this does not mean that most children with CAS have a pathogenic *FOXP2* disruption. In fact, although the frequency of *FOXP2* variants in cohorts of children with CAS has not yet been systematically assessed, it is likely that pathogenic variants in *FOXP2* can only be found in a very small minority of children with CAS[43]. The reason for this is that there are likely many more genes in which pathogenic variants can cause CAS, resulting in genetic heterogeneity, a concept further explained later in this chapter. Until today *FOXP2* is the only gene recorded as associated with a specific speech and language disorder in the OMIM (Online Mendelian Inheritance in Man) database, where the phenotype associated with *FOXP2* variants is indexed as *Speech-language disorder-1* (MIM 602081).

Chromosomal abnormalities in developmental speech and language disorders

When discussing different types of genetic variation, a distinction is often made between variants at the chromosomal level and at the nucleotide (base pair) level. Chromosomal aberrations include aneuploidies (gain or loss of chromosomes; e.g. trisomy 21) and copy number variants (CNVs). CNVs are gains or losses of chromosomal material (>1,000 nucleotides), e.g. a 22q11 microdeletion, which is a deletion involving a part (band 11) of the long arm (q-arm) of chromosome 22. In addition to aneuploidies and CNVs, many other complex chromosomal rearrangements can occur: e.g. translocations, inversions, etc.

Large chromosomal aberrations (>5Mb) can be detected using karyotyping, a technique in which the chromosomes are made visible under a microscope. Using karyotyping, a study published in 1980 showed an increased prevalence of numerical chromosomal abnormalities in a cohort of children with specific speech and language delays: four out of 88 tested children had a whole chromosome aneuploidy (47XXY, 47XYY, 48XXYY and mosaic trisomy 21)[51]. This finding was confirmed in a more recent study using genome-wide SNP genotype data, in which a higher prevalence of sex chromosome aneuploidies was found in a cohort of children with speech/language delays and a cohort of children with specific language impairments, compared with a control cohort[52]. Numerical large chromosomal abnormalities, as well as chromosomal translocations have all been reported in individuals with speech or language disorders, e.g. reciprocal translocations affecting *FOXP2*[42,53,54].

Since the introduction of microarray platforms, including array CGH (comparative genomic hybridization) and SNP arrays, it is now possible to easily perform genome-wide CNV screens with high resolution[55,56]. A small number of cohort studies have been published in which CNV testing was done in children with developmental speech and language disorders, with overlapping conclusions but also conflicting evidence in terms of the enrichment of different types of CNVs in individuals with DLD. Laffin et al. identified different possibly pathogenic CNVs using array CGH in a USA-based cohort of children with CAS, including a 16p11.2 deletion[57]. Simpson et al. used SNP genotyping (in combination with CNV calling) to show an enrichment of common CNVs in children with DLD from the United Kingdom, not driven by large or *de novo* variants[58]. A *de novo* variant is a variant (SNV or CNV) that is present in an individual but not in either parent. Finally, Kalnak et al. similarly used SNP arrays to call CNVs in Swedish children with a DLD and typically developing children, but in this case reported an enrichment for rare and *de novo* CNVs in the DLD group[59].

For the broader field of neurodevelopmental disorders, the introduction of microarrays in research and diagnostic settings has led to the identification and characterization of many new microdeletion and duplication syndromes, and thus provided many affected individuals with a molecular diagnosis[60,61]. One of the best characterized CNVs is the 16p11.2 deletion, which constituted a risk factor in several different cohort studies of children with developmental speech and language disorders[57,59] and in studies of individuals with CAS[62-64]. A prospective study in 55 individuals with a 16p11.2 deletion showed that the majority of individuals had CAS, as well as receptive and expressive language impairments[65]. However, although speech and language problems are often a prominent feature of this microdeletion syndrome, the phenotype is not speech/language-specific, as a wide spectrum of additional features can be present in affected individuals, including autism spectrum disorders, mild intellectual disability and seizures[66]. Other CNVs have been associated with speech and/or language delays, e.g. 15q11.2 deletion or duplication[67,68], 7q11.23 duplication[69] and 1p21.3 deletion. Individuals with these deletions or duplications often show prominent speech and/or language impairments, but again the syndromes are not limited to the speech/language domain and often show highly variable expressivity, which is further discussed in the last paragraph of this chapter.

In summary, most chromosomal abnormalities found in children with speech and language disorders are not speech- or language-specific, but often associated with a broad range of other possible neurodevelopmental impairments. Nevertheless, it is clear that CNVs can be the cause developmental speech and language disorders, and that these pathogenic CNVs are more present in children with developmental speech and language disorders compared to typically developing children.

Next generation sequencing

Not long after the introduction of high resolution microarrays, next generation sequencing (NGS) methods became widely available[70]. NGS is a collective name for all techniques involving massive parallel sequencing. In short, these techniques have made it possible to read out the sequence of nucleotides of a very large number of DNA fragments at the same time, at costs that have become lower and lower as methods have advanced.. Due to the introduction of NGS, it is now possible to analyse entire genomes at a base pair resolution with relative ease, for example to search for disease-causing single nucleotide variants. Perhaps the most common current use of NGS techniques in human genetics is for sequencing of whole exomes (where the exome is defined as all coding genes within the genome) and genomes.

The advantages of these new sequencing techniques are numerous. With exome and genome sequencing it is now much more feasible to identify Mendelian disorders caused by single nucleotide variants, obviating the need for a prior hypothesis about the specific gene involved. This is especially important for disorders that are less recognizable and/or phenotypes that can be caused by a single variant in many different genes (genetically heterogeneous disorders). These advances have led to highly improved diagnostic possibilities for patients with various kind of disorders, as the number of new candidate genes and associated disorders has expanded very rapidly and is still growing[71]. Altogether, the widespread introduction of NGS in research and diagnostic practices has greatly contributed to our knowledge on disease genes and associated inheritance mechanisms, and also on proteins, pathways and pathogenic mechanisms involved in disease.

Exome sequencing is now implemented in diagnostic practices for many developmental disorders, such as intellectual disability and autism spectrum disorder, and has been demonstrated to outperform chromosome microarrays as a first-tier clinical diagnostic test for unexplained neurodevelopmental disorders[72]. Not only has the use of NGS confirmed the extreme genetic heterogeneity for these disorders, it has also revealed the significance of *de novo* variants. It is estimated that more than 40% of individuals with a severe neurodevelopmental disorder has a disease-causing *de novo* variant in the coding region of a gene[73].

Single nucleotide variants in developmental speech and language disorders

In contrast to several other neurodevelopmental disorders, for the field of developmental speech and language disorders, genetic testing and especially diagnostic exome/genome sequencing is not yet common practice. In a research setting, a few studies have been published in which NGS has been used to investigate developmental speech or language disorder cohorts: three studies in children with CAS, and one in a cohort of children with DLD. The first study on CAS used exome sequencing in a cohort of ten affected children[74]. Several variants of potential interest were reported, but interpretation of these variants was hampered by missing information on variant frequencies in control populations and inheritance status. Moreover, although the authors sequenced the whole exome, they prioritized most of their analysis on prior “candidate genes” from the earlier (limited) literature, leading them to propose variants in those genes as potentially pathogenic, that have subsequently turned out to be benign. Another NGS study on CAS used whole genome sequencing in nineteen unrelated individuals, a cohort partially overlapping with the previously mentioned exome sequencing study[75]. Three *de novo* likely pathogenic variants were found in three probands, affecting the genes *CHD3, SETD1A* and *WDR5*, and five additional pathogenic or likely pathogenic variants in other probands where parental information was unavailable for determining inheritance status. A very recent study that used NGS in a CAS cohort consisting of 34 children is in line with these results, as likely pathogenic SNVs or CNVs were found in 11 out of 34 probands (32%)[76]. These latter two studies show the potential of NGS for investigating developmental speech disorders, and highlight that in a significant number of cases these disorders can be caused by a high penetrance variant disrupting a single gene.

For DLD, only one NGS study has been published so far, in which exome sequencing in a research setting was performed in 43 unrelated probands with severe specific language impairment[77]. Though different potentially pathogenic variants in several genes were identified, pathogenicity is still unclear for many of those. The contribution of *de novo* variants and more general of Mendelian disorders in children with DLD remains to be explored.

Interestingly, pathogenic variants that have been reported in the speech/language cohort studies described above have often been found in known disease genes or in candidate genes for a broader group of neurodevelopmental disorders, as other individuals with variants in the same gene do not always have prominent speech or language problems. Moreover, in several different newly characterized neurodevelopmental disorders, impairments in speech and language appear frequently as part of the clinical spectrum of associated disorders.

It is hard to classify all disease genes based on the involvement of speech and language problems, as most studies do not assess and/or describe these capacities in a standardized manner. However, for some genes the link with speech and language problems is clear: e.g. when prominent speech/language disorders are recurrently reported in genotype-driven disorder characterization studies, and when rare variants in the same gene are found in NGS studies of speech and/or language disorder cohorts. A notable example is the *SETBP1* gene. Loss-of-function variants in this gene cause a neurodevelopmental disorder, commonly referred to as *SETBP1* disorder, that is characterized by severe expressive speech and/or language problems, in combination with mild intellectual disability[78,79]. Similarly, variants affecting *KAT6A* cause a neurodevelopmental disorder that can vary in severity, but speech problems and more specifically CAS, are prominent[80]. These are just two examples of a potentially large group of genes in which variants can give rise to impairments in speech and language development. A more systematic use of NGS to study developmental speech and language impairments, coupled to a more systematic characterization of phenotypes, is needed.

Genetic heterogeneity and variable expressivity

Two concepts are extremely important for understanding the genetic background of Mendelian inherited speech and language disorders: genetic heterogeneity and variable expressivity.

In Mendelian disorders, a DNA variant in a single gene can be enough to cause a phenotype. This does not mean that individuals with the phenotype always share the same underlying genetic cause. For some phenotypes there is this 1:1 relationship with the gene involved, for example the disease cystic fibrosis is always caused by variants in the CFTR gene. But for many phenotypes or disorders there is genetic heterogeneity, which means that the same phenotype in different individuals can be caused by specific variants in different genes. For example Noonan syndrome, a clinically recognizable neurodevelopmental phenotype, can be caused by pathogenic variants in >10 different genes[81]. Genetic heterogeneity can be even more extreme, depending on the phenotype of interest. For intellectual disability, over 1500 different Mendelian disease genes are already known[82]. Research so far on developmental speech and language disorders also points to significant genetic heterogeneity[75,76,83]. This means that unbiased genetic screens are preferred, as it is extremely hard to select specific disease-causing genes as candidates based only on the phenotype.

Variable expressivity is another important concept in this context. If a SNV or CNV is pathogenic, this means that the variant is disease-causing. But this does not mean that it always leads to exactly the same clinical features or the same severity of disorder. For many Mendelian neurodevelopmental disorders considerable variable expressivity has been shown, which means that the phenotype amongst individuals carrying the same variant might differ. For developmental speech and language disorders an example is the 16p11.2 deletion related disorder. While there is no doubt that this deletion is pathogenic, the resulting clinical features can vary substantially: between different families with this deletion, and even between affected individuals within the same family[66]. While one family member might have CAS, another person can have a mild intellectual disability and a third person might have prominent features of autism spectrum disorder. Some individuals who carry a pathogenic SNV or CNV may not display any phenotype at all. This phenomenon is referred to as incomplete penetrance.

All in all, when interpreting NGS data for Mendelian speech and language disorders in a clinical or research context, it is important to realize that these disorders can be caused by a very large number of different SNVs and CNVs in different individuals, and that these variants are often associated with more than only speech and language problems (Figure 2).

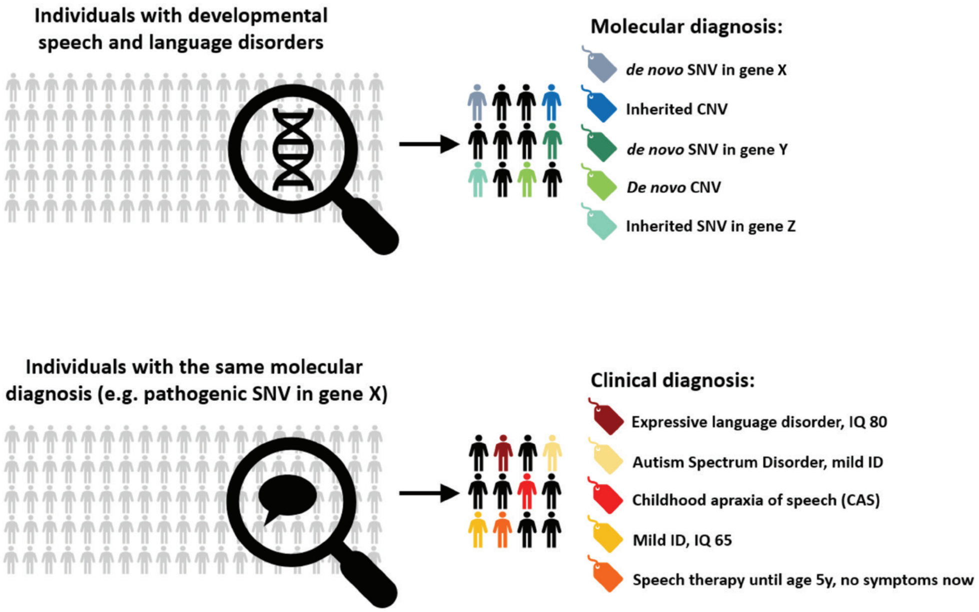

Individuals with developmental speech and language disorders
Molecular diagnosis:
de novo SNV in gene X
Inherited CNV
de novo SNV in gene Y
De novo CNV
Inherited SNV in gene Z
Individuals with the same molecular diagnosis (e.g. pathogenic SNV in gene X)
Clinical diagnosis:
Expressive language disorder, IQ 80
Autism Spectrum Disorder, mild ID
Childhood apraxia of speech (CAS)
Mild ID, IQ 65
Speech therapy until age 5y, no symptoms now

Figure 2: Genetic heterogeneity and variable expressivity

a) Genetic heterogeneity in developmental speech and language disorders: the same phenotype in different individuals can be caused by specific variants in different genes
b) Variable Expressivity of Mendelian neurodevelopmental disorders: individuals carrying the same variant might have different phenotypes

Aim, relevance and outline of this thesis

Aim and relevance of this thesis

At present, not much is known about the molecular genetic basis of developmental speech and language disorders. Although these disorders are considered to mostly reflect a multifactorial etiology, there is increasing evidence for an important additional role for Mendelian causes. NGS techniques are now frequently used to identify molecular causes in several neurodevelopmental disorders, which has led to a large increase in knowledge on underlying genetic mechanisms. In contrast, the use of NGS in developmental speech and language disorders so far is very limited, and as a consequence, the genetic architecture underlying these disorders is incompletely understood.

The aim of this thesis is to investigate Mendelian causes of developmental speech and language disorders by NGS techniques. By integrating NGS data, phenotyping data and functional data from protein-specific laboratory assays, our goal is to better understand the genetic background of these disorders and gain knowledge to improve clinical care.

A better understanding of the role of Mendelian causes in developmental speech and language disorders is highly important, as it might provide families with improved genetic counseling on recurrence risks, and on possible additional associated features, and provide guidance to specific intervention options for speech and language therapists. In general, if NGS results from studies of children with speech and language impairments indicate a larger role for Mendelian factors, this would show the need for offering genetic testing via NGS in children with these disorders. NGS results obtained through studying unique patients with abnormal developmental phenotypes have already been shown to be an inexhaustible resource of information on molecular pathways and pathogenic mechanisms involved in normal and abnormal neurodevelopmental processes.

Outline of this thesis

In **Chapter 2**, we characterize a series of rare variants in the *MED13* gene, after the initial finding of a missense variant of unknown significance in a child with a DLD. We show that pathogenic variants in *MED13* can cause a neurodevelopmental disorder with associated speech and language disorders.

In **Chapter 3**, we describe how one variant of interest in a child with a DLD has led to the identification of *POU3F3*-associated disorder. *POU3F3*, also known as *Brain-1*, is a well-known gene in the context of brain development. W show how different variants disrupt the transcription factor activity of the encoded protein, and lead to a neurodevelopmental disorder with prominent speech and language impairments and a characteristic facial phenotype.

For **Chapter 4**, a genome sequencing result from a research project in children with CAS was the starting point for the characterization of pathogenic *CHD3* variants in a larger series of patients. We found that specific missense variants in this gene disrupt ATPase activity and chromatin remodeling functions of the encoded protein. With this study, we implicate *CHD3* variants in a disorder characterized by intellectual disability, macrocephaly and impaired speech and language.

In **Chapter 5**, we characterized several missense variants in *WDR5*, after the identification of a *de novo* variant in this gene in a child with a developmental speech disorder (CAS). We show how the amino acids involved in these variants cluster in the three-dimensional structure of the WDR5 protein, and characterize the phenotypic spectrum of this neurodevelopmental disorder with variable expressivity.

The study in **Chapter 6** started with an interesting missense variant in *FOXP4*. We demonstrate that, similar to its gene family members *FOXP1* and *FOXP2*, heterozygous variants in *FOXP4* cause a neurodevelopmental phenotype with associated developmental speech and language difficulties.

In **Chapter 7** we delineate speech, language, oral motor and neuropsychological phenotypes in a cohort of individuals with *SATB2*-associated syndrome. On the one hand, we observe clinical overlap and highly recurrent features. On the other, we show that speech problems can differ between individuals, and that the severity of clinical features is highly variable.

Chapter 8 describes our study design for the GENTOS study, a prospective cohort study in which we plan to perform whole-genome sequencing using a trio approach in 50 individuals with a severe developmental speech and language disorder and their parents. The aim of this study is to determine the diagnostic yield for rare (Mendelian) genetic causes in DLD.

And finally, in **Chapter 9**, we summarize the research in this thesis, and discuss how the results from our studies have increased our understanding of rare pathogenic variants and their impact on speech and language development, as well as ideas for further research on this topic.

References

1. Mayor J, Plunkett K. A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Dev Sci.* 2011;14(4):769-785.
2. Fisher SE, Marcus GF. The eloquent ape: genes, brains and the evolution of language. *Nat Rev Genet.* 2006;7(1):9-20.
3. Van Agt H, Verhoeven L, Van Den Brink G, De Koning H. The impact on socio-emotional development and quality of life of language impairment in 8-year-old children. *Dev Med Child Neurol.* 2011;53(1):81-88.
4. Reilly S, McKean C, Morgan A, Wake M. Identifying and managing common childhood language and speech impairments. *BMJ.* 2015;350:h2318.
5. Schachinger-Lorentzon U, Kadesjo B, Gillberg C, Miniscalco C. Children screening positive for language delay at 2.5 years: language disorder and developmental profiles. *Neuropsychiatr Dis Treat.* 2018;14:3267-3277.
6. Reilly S, Tomblin B, Law J, et al. Specific language impairment: a convenient label for whom? *Int J Lang Commun Disord.* 2014;49(4):416-451.
7. Bishop DV, Snowling MJ, Thompson PA, Greenhalgh T, consortium C. CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. *PLoS One.* 2016;11(7):e0158753.
8. *Diagnostic and statistical manual of mental disorders : DSM-5.* Fifth edition. Arlington, VA : American Psychiatric Association, [2013]; 2013.
9. Bishop DVM, Snowling MJ, Thompson PA, Greenhalgh T, and the C-c. Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *J Child Psychol Psychiatry.* 2017;58(10):1068-1080.
10. Gerrits E, Beers M, Bruinsma G, Singer I. *Handboek Taalontwikkelingsstoornissen.* 2017.
11. Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res.* 1997;40(6):1245-1260.
12. NVLF (Nederlandse Vereniging voor Logopedie en Foniatrie) FFvNAC, SIAC (Samenwerkingen Instellingen voor mensen met Auditieve en/of Communicatieve beperkingen). *Ketenzorg taalontwikkelingsstoornis in beeld.* 2019.
13. Waring R, Knight R. How should children with speech sound disorders be classified? A review and critical evaluation of current classification systems. *Int J Lang Commun Disord.* 2013;48(1):25-40.
14. Leonard LB. *Children with specific language impairment, second edition.* 2014.
15. Capone Singleton N. Late Talkers: Why the Wait-and-See Approach Is Outdated. *Pediatr Clin North Am.* 2018;65(1):13-29.
16. Miniscalco C, Fernell E, Thompson L, Sandberg E, Kadesjo B, Gillberg C. Development problems were common five years after positive screening for language disorders and, or, autism at 2.5 years of age. *Acta Paediatr.* 2018;107(10):1739-1749.
17. Schlichting JEPT, Spelberg HCL. *Schlichting test voor Taalbegrip.* Bohn Stafleu van Loghum; 2010.
18. Schlichting JEPT, Spelberg HCL. *Schlichting Test voor Taalproductie-II.* Bohn Stafleu van Loghum; 2010.
19. Schlichting L. Peabody Picture Vocabulary Test Dutch-III-NL. *Amsterdam, NL: Harcourt Assessment BV.* 2005.
20. Kort W, Schittekatte M, Compaan E. CELF-4-NL: clinical evaluation of language fundamentals. *Amsterdam: Pearson Assessment and Information.* 2008.
21. Zink I, Lambrechts D. NNST. *De Nederlandstalige Nonspeech Test Aanpassing en hernormering van Nonspeech Test for receptive and expressive language (NST) van Mary Blake Huer (1988) Leuven: Acco.* 2001.
22. Olswang L, Bain B. When to Recommend Intervention. *Language Speech and Hearing Services in Schools.* 1991;22:255.
23. Stothard SE, Snowling MJ, Bishop DV, Chipchase BB, Kaplan CA. Language-impaired preschoolers: a follow-up into adolescence. *J Speech Lang Hear Res.* 1998;41(2):407-418.
24. Feeney R, Desha L, Ziviani J, Nicholson JM. Health-related quality-of-life of children with speech and language difficulties: a review of the literature. *Int J Speech Lang Pathol.* 2012;14(1):59-72.
25. Bishop D, Professor M. Neurodevelopmental Disorders: Conceptual Issues. In:2009:32-41.
26. Kjelgaard MM, Tager-Flusberg H. An Investigation of Language Impairment in Autism: Implications for Genetic Subgroups. *Lang Cogn Process.* 2001;16(2-3):287-308.
27. Baumer FM, Cardon AL, Porter BE. Language Dysfunction in Pediatric Epilepsy. *J Pediatr.* 2018;194:13-21.
28. Metz-Lutz MN, Filippini M. Neuropsychological findings in Rolandic epilepsy and Landau-Kleffner syndrome. *Epilepsia.* 2006;47 Suppl 2:71-75.
29. Stromswold K. Genetics of spoken language disorders. *Hum Biol.* 1998;70(2):297-324.

30. Bishop DVM. *Uncommon Understanding - Development and Disorders of Language Comprehension in Children.* Psychology Press Ltd; 1997.
31. Lewis BA, Thompson LA. A study of developmental speech and language disorders in twins. *J Speech Hear Res.* 1992;35(5):1086-1094.
32. Bishop DV, North T, Donlan C. Genetic basis of specific language impairment: evidence from a twin study. *Dev Med Child Neurol.* 1995;37(1):56-71.
33. Tomblin JB, Buckwalter PR. Heritability of poor language achievement among twins. *J Speech Lang Hear Res.* 1998;41(1):188-199.
34. Bishop DV, Hayiou-Thomas ME. Heritability of specific language impairment depends on diagnostic criteria. *Genes Brain Behav.* 2008;7(3):365-372.
35. Hayiou-Thomas ME. Genetic and environmental influences on early speech, language and literacy development. *J Commun Disord.* 2008;41(5):397-408.
36. Gialluisi A, Newbury DF, Wilcutt EG, et al. Genome-wide screening for DNA variants associated with reading and language traits. *Genes Brain Behav.* 2014;13(7):686-701.
37. Eicher JD, Powers NR, Miller LL, et al. Genome-wide association study of shared components of reading disability and language impairment. *Genes Brain Behav.* 2013;12(8):792-801.
38. Kornilov SA, Rakhlin N, Koposov R, et al. Genome-Wide Association and Exome Sequencing Study of Language Disorder in an Isolated Population. *Pediatrics.* 2016;137(4).
39. Nudel R, Simpson NH, Baird G, et al. Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. *Genes Brain Behav.* 2014;13(4):418-429.
40. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature.* 2001;413(6855):519-523.
41. Fisher SE, Lai CS, Monaco AP. Deciphering the genetic basis of speech and language disorders. *Annu Rev Neurosci.* 2003;26:57-80.
42. Lai CS, Fisher SE, Hurst JA, et al. The SPCH1 region on human 7q31: genomic characterization of the critical interval and localization of translocations associated with speech and language disorder. *Am J Hum Genet.* 2000;67(2):357-368.
43. Morgan A, Fisher SE, Scheffer I, Hildebrand M. FOXP2-Related Speech and Language Disorders. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews((R)).* Seattle (WA)1993.
44. Reuter MS, Riess A, Moog U, et al. FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. *J Med Genet.* 2017;54(1):64-72.
45. Carlsson P, Mahlapuu M. Forkhead transcription factors: key players in development and metabolism. *Dev Biol.* 2002;250(1):1-23.
46. Ferland RJ, Cherry TJ, Preware PO, Morrisey EE, Walsh CA. Characterization of Foxp2 and Foxp1 mRNA and protein in the developing and mature brain. *J Comp Neurol.* 2003;460(2):266-279.
47. Co M, Hickey SL, Kulkarni A, Harper M, Konopka G. Cortical Foxp2 Supports Behavioral Flexibility and Developmental Dopamine D1 Receptor Expression. *Cereb Cortex.* 2020;30(3):1855-1870.
48. Medvedeva VP, Rieger MA, Vieth B, et al. Altered social behavior in mice carrying a cortical Foxp2 deletion. *Hum Mol Genet.* 2019;28(5):701-717.
49. Norton P, Barschke P, Scharff C, Mendoza E. Differential Song Deficits after Lentivirus-Mediated Knockdown of FoxP1, FoxP2, or FoxP4 in Area X of Juvenile Zebra Finches. *J Neurosci.* 2019;39(49):9782-9796.
50. Fisher SE, Scharff C. FOXP2 as a molecular window into speech and language. *Trends Genet.* 2009;25(4):166-177.
51. Mutton DE, Lea J. Chromosome studies of children with specific speech and language delay. *Dev Med Child Neurol.* 1980;22(5):588-594.
52. Simpson NH, Addis L, Brandler WM, et al. Increased prevalence of sex chromosome aneuploidies in specific language impairment and dyslexia. *Dev Med Child Neurol.* 2014;56(4):346-353.
53. Shriberg LD, Ballard KJ, Tomblin JB, Duffy JR, Odell KH, Williams CA. Speech, prosody, and voice characteristics of a mother and daughter with a 7;13 translocation affecting FOXP2. *J Speech Lang Hear Res.* 2006;49(3):500-525.
54. Becker M, Devanna P, Fisher SE, Vernes SC. A chromosomal rearrangement in a child with severe speech and language disorder separates FOXP2 from a functional enhancer. *Mol Cytogenet.* 2015;8:69.
55. Vissers LE, de Vries BB, Osoegawa K, et al. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am J Hum Genet.* 2003;73(6):1261-1270.
56. McMullan DJ, Bonin M, Hehir-Kwa JY, et al. Molecular karyotyping of patients with unexplained mental retardation by SNP arrays: a multicenter study. *Hum Mutat.* 2009;30(7):1082-1092.

57. Laffin JJ, Raca G, Jackson CA, Strand EA, Jakielski KJ, Shriberg LD. Novel candidate genes and regions for childhood apraxia of speech identified by array comparative genomic hybridization. *Genet Med.* 2012;14(11):928-936.
58. Simpson NH, Ceroni F, Reader RH, et al. Genome-wide analysis identifies a role for common copy number variants in specific language impairment. *Eur J Hum Genet.* 2015;23(10):1370-1377.
59. Kalnak N, Stamouli S, Peyrard-Janvid M, et al. Enrichment of rare copy number variation in children with developmental language disorder. *Clin Genet.* 2018;94(3-4):313-320.
60. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of developmental delay. *Nat Genet.* 2011;43(9):838-846.
61. Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JP, Burton H. Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet Med.* 2009;11(3):139-146.
62. Fedorenko E, Morgan A, Murray E, et al. A highly penetrant form of childhood apraxia of speech due to deletion of 16p11.2. *Eur J Hum Genet.* 2016;24(2):302-306.
63. Raca G, Baas BS, Kirmani S, et al. Childhood Apraxia of Speech (CAS) in two patients with 16p11.2 microdeletion syndrome. *Eur J Hum Genet.* 2013;21(4):455-459.
64. Newbury DF, Mari F, Sadighi Akha E, et al. Dual copy number variants involving 16p11 and 6q22 in a case of childhood apraxia of speech and pervasive developmental disorder. *Eur J Hum Genet.* 2013;21(4):361-365.
65. Mei C, Fedorenko E, Amor DJ, et al. Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. *Eur J Hum Genet.* 2018;26(5):676-686.
66. Miller DT, Chung W, Nasir R, et al. 16p11.2 Recurrent Microdeletion. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews((R)).* Seattle (WA)1993.
67. Burnside RD, Pasion R, Mikhail FM, et al. Microdeletion/microduplication of proximal 15q11.2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay. *Hum Genet.* 2011;130(4):517-528.
68. Finucane BM, Lusk L, Arkilo D, et al. 15q Duplication Syndrome and Related Disorders. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews((R)).* Seattle (WA)1993.
69. Mervis CB, Morris CA, Klein-Tasman BP, Velleman SL, Osborne LR. 7q11.23 Duplication Syndrome. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews((R)).* Seattle (WA)1993.
70. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008;26(10):1135-1145.
71. Vissers LE, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet.* 2016;17(1):9-18.
72. Srivastava S, Love-Nichols JA, Dies KA, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med.* 2019;21(11):2413-2421.
73. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature.* 2017;542(7642):433-438.
74. Worthey EA, Raca G, Laffin JJ, et al. Whole-exome sequencing supports genetic heterogeneity in childhood apraxia of speech. *J Neurodev Disord.* 2013;5(1):29.
75. Eising E, Carrion-Castillo A, Vino A, et al. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol Psychiatry.* 2019;24(7):1065-1078.
76. Hildebrand MS, Jackson VE, Scerri TS, et al. Severe childhood speech disorder: Gene discovery highlights transcriptional dysregulation. *Neurology.* 2020.
77. Chen XS, Reader RH, Hoischen A, et al. Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. *Sci Rep.* 2017;7:46105.
78. Coe BP, Witherspoon K, Rosenfeld JA, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet.* 2014;46(10):1063-1071.
79. Filges I, Shimojima K, Okamoto N, et al. Reduced expression by SETBP1 haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel-Giedion syndrome. *J Med Genet.* 2011;48(2):117-122.
80. Kennedy J, Goudie D, Blair E, et al. KAT6A Syndrome: genotype-phenotype correlation in 76 patients with pathogenic KAT6A variants. *Genet Med.* 2019;21(4):850-860.
81. Allanson JE, Roberts AE. Noonan Syndrome. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews((R)).* Seattle (WA)1993.
82. Thormann A, Halachev M, McLaren W, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun.* 2019;10(1):2373.
83. Barnett CP, van Bon BW. Monogenic and chromosomal causes of isolated speech and language impairment. *J Med Genet.* 2015;52(11):719-729.

2

Chapter 2

De novo mutations in MED13, a component of the Mediator complex, are associated with a novel neurodevelopmental disorder

Lot Snijders Blok*, Susan M. Hiatt*, Kevin M. Bowling, Jeremy W. Prokop, Krysta L. Engel, J. Nicholas Cochran, E. Martina Bebin, Emilia K. Bijlsma, Claudia A. L. Ruivenkamp, Paulien Terhal, Marleen E. H. Simon, Rosemarie Smith, Jane A. Hurst, The DDD study, Heather McLaughlin, Richard Person, Amy Crunk, Michael F. Wangler, Haley Streff, Joseph D. Symonds, Sameer M. Zuberi, Katherine S. Elliott, Victoria R. Sanders, Abigail Masunga, Robert J. Hopkin, Holly A. Dubbs, Xilma R. Ortiz-Gonzalez, Rolph Pfundt, Han G. Brunner, Simon E. Fisher, Tjitske Kleefstra*, Gregory M. Cooper*

** These authors contributed equally*

Human Genetics 2018
doi:10.1007/s00439-018-1887-y

Abstract

Many genetic causes of developmental delay and/or intellectual disability (DD/ID) are extremely rare, and robust discovery of these requires both large-scale DNA sequencing and data sharing. Here we describe a GeneMatcher collaboration which led to a cohort of 13 affected individuals harboring protein-altering variants, 11 of which are de novo, in *MED13*; the only inherited variant was transmitted to an affected child from an affected mother. All patients had intellectual disability and/or developmental delays, including speech delays or disorders. Other features that were reported in two or more patients include autism spectrum disorder, attention deficit hyperactivity disorder, optic nerve abnormalities, Duane anomaly, hypotonia, mild congenital heart abnormalities, and dysmorphisms. Six affected individuals had mutations that are predicted to truncate the MED13 protein, six had missense mutations, and one had an in-frame-deletion of one amino acid. Out of the seven nontruncating mutations, six clustered in two specific locations of the MED13 protein: an N-terminal and C-terminal region. The four N-terminal clustering mutations affect two adjacent amino acids that are known to be involved in MED13 ubiquitination and degradation, p.Thr326 and p.Pro327. MED13 is a component of the CDK8-kinase module that can reversibly bind Mediator, a multi-protein complex that is required for Polymerase II transcription initiation. Mutations in several other genes encoding subunits of Mediator have been previously shown to associate with DD/ID, including *MED13L*, a paralog of *MED13*. Thus, our findings add *MED13* to the group of CDK8-kinase module-associated disease genes.

Introduction

The introduction of next-generation sequencing techniques has rapidly improved the identification of genes that associate with rare disease. Although developmental delay (DD) and intellectual disability (ID) are relatively common[1], there is extreme genetic heterogeneity among affected patients and a large fraction of patients with DD/ID remain refractory to diagnosis[2]. In unsolved cases, the understanding of gene–disease relationships has greatly benefited from collaboration between clinical genetics teams (Sobreira et al. 2015). In fact, many recently discovered DD/ID genes have come from "matchmaking"[3-5], where websites such as GeneMatcher[6] facilitate the comparison of patients with rare genotypes and phenotypes across the world.

Here we present the results of a collaboration facilitated by GeneMatcher [6] in which multiple clinical and research groups independently identified individuals with DD/ID and related phenotypes with rare protein-altering variation in *MED13*. This genotype-driven approach enabled us to characterize the phenotypes and mutational spectrum of a cohort of 13 patients, each with a likely pathogenic variant in *MED13*.

Although *MED13* has not been previously linked to a disorder, it is a paralog of *MED13L*, mutations of which have been found to cause ID, speech impairment and heart defects[7-9]. The gene products MED13 and MED13L are mutually exclusive components of the reversible CDK8-module of the Mediator complex, a multi-protein complex that is required for the expression of all protein-coding genes[10,11]. In this study, we show that variants in *MED13* are also associated with a neurodevelopmental disorder, and delineate the corresponding phenotypic features and mutational spectrum.

Materials and methods

Informed consent

Informed consent to publish de-identified data was obtained from all patients, either as part of the diagnostic workflow or as part of a research study[12]. Informed consent to publish clinical photographs was also obtained when applicable. Informed consent matched the local ethical guidelines.

Exome/genome sequencing

In patients A, B, D, E, F, G, I, K, L and M, whole exome sequencing and variant filtering were performed as previously published[13-17]. In patient C, targeted Sanger sequencing was performed to confirm the presence of the MED13 variant (L131*) that was first identified in patient B. For patient H, whole genome sequencing was performed using Illumina's HiSeq X ten platform. Sequencing reads were mapped against the hs37d5 reference using GATK. Variants were called using GATK's Haplotype Caller. Variants were filtered using frequencies from the ExAC and gnomAD databases (mean allele frequency < 0.003) and for conservation using PhastCons (> 0.5) and PhyloP (> 4). For patient J, whole genome sequencing, variant prioritization, and Sanger validation were performed as previously described[12]. In each patient, the observed *MED13* mutation was considered to be the most likely contributor to the phenotype, and no additional pathogenic or likely pathogenic variants were found.

Three-dimensional modeling

Protein modeling was performed as previously described[18]. Modeling of MED13 interacting with FBXW7 was performed using PDB 2OVQ, replacing molecule C with the MED13 amino acids 321–330. Binding energy was calculated following each patient variant insertion and energy minimization using AMBER14 force field (http://ambermd.org) in YASARA.

RNA isolation

2.5 mL of blood was collected in PAXgene RNA tubes (PreAnalytiX #762165) according to the manufacturer's instructions and stored short-term at − 20 °C. RNA was isolated using a PAX gene Blood RNA Kit (Qiagen #762164) according to the manufacturer's instructions. Isolated RNA was quantified by Qubit® (Thermo Fisher #Q32855).

cDNA synthesis

First strand synthesis of cDNA was performed from 150 ng of RNA isolated from blood using Superscript™ III (Thermo Fisher #18080044) according to manufacturer's instructions using random primers (Invitrogen #48190011) for +/− RT reactions. The products were diluted 1:10 in water before use in qPCR reactions.

qPCR

qPCR was performed according to manufacturer's protocols using Taqman gene expression master mix (ThermoFisher #4369016) and FAM-MGB Taqman probes directed against MED13 (ThermoFisher Hs01080701_m1 catalog #4331182) and GAPDH (ThermoFisher #4352934E). qPCR reactions were carried out in a QuantStudio 6 Flex Real-Time PCR system (Applied Biosystems) using 40 cycles of amplification. Raw C_T values were obtained, normalized first to the GAPDH loading control, and then to the proband. We tested an additional loading control [AGPAT-data not shown (ThermoFisher Hs00965850_g1; catalog #4331182)], but the data were like those normalized to GAPDH.

Sanger sequencing

cDNA template was amplified using primers to the region of interest: 5'-CGA GGC TCT TAT GGA ACT GAT GAA TC-3' (forward) and 5'-GAT CCA TCG TGC TTT CAG ACA CAT C-3' (reverse). No amplification was observed in the no RT condition. PCR conditions were: 500 nM primers, 3% DMSO, 1x Phusion HF (NEB #M0531L), 0.5 µL cDNA template, and cycling at (98 °C, 30 s), (98 °C, 10 s; 60 °C, 30 s; 72 °C, 45 s)x35, (72 °C, 7 m), (4 °C, ∞). The additional reverse primer 5'-AAA TGC TTC ATT GTT ACC GTC AGC T-3' and the additional forward primers 5'-TCC AAA AGA AAC GAT GTG AGT ATG CAG-3', 5'-CTC TCT TCA GCC AGT TCT TCA GGA T-3', 5'-ACA ATT TCA TAA AAT GGC TGG CCG A-3', 5'-CGA GGC TCT TAT GGA ACT GAT GAA TC-3', 5'-GTG CTT

TCT CCA TTT GCT CTT CCT T-3' were used for sequencing, along with the primers used for amplification from cDNA. Chromatograms were quantified using ab1PeakReporter (Thermo Fisher).

Western Blot

Whole blood was collected using cell processing tubes (BD #362760), isolated according to the manufacturer's instructions, and stored in liquid nitrogen in CTS™ Synth-a-Freeze® Medium (Thermo Fisher # A13713-01) until use. As a control for antibody specificity, MED13 was knocked down in neural precursor cells (clone BC1, MTI-GlobalStem #GSC-4311) by generating stable lines using puromycin selection expressing shRNA against MED13 (Sigma Aldrich # SHCLNG-NM_005121; TRCN0000234904) compared to a GFP shRNA control in the same vector (Addgene # 30323). Cell pellets were processed using the NE-PER™ (Thermo Fisher #78833) nuclear and cytoplasmic extraction kit according to the manufacturer's instructions, and nuclear extracts were used for the blot shown (whole cell extracts, even at very high concentrations, did not produce sufficient signal). 60 µg of protein was loaded for patient blood samples, and 15 µg of protein was loaded for neural precursor cell samples. Blots were blocked for 1 h at room temperature in LICOR blocking buffer (LICOR #927-40000), then blots were probed (with washes in PBS-T (0.05% Tween-20) and a secondary probe for 1 h after each primary probe) with 1:250 rabbit anti-MED13 (Bethyl #A301-277A) for 3 days at 4 °C, 1:1,000 mouse anti-HDAC2 (clone 3F3, SCBT #sc-81599) overnight at 4 °C as a loading control, and 1:1000 rabbit anti-HSP90 (abcam #ab115660) overnight at 4 °C as an additional loading control. Secondary probes were used at 1:20,000 (LICOR #926-32211 and #926-68070). Three other primary antibodies were tested for MED13, but did not show sufficient signal to detect MED13 in blood despite detecting MED13 in neural precursor nuclear lysates: Bethyl #278A, Abcam #ab49468, and Abcam #ab76923 (data not shown).

Statistical enrichment of MED13 variants in DD/ID cohorts

We compared the frequency of observed de novo MED13 variation identified in two large sequencing cohorts to the expected frequency of variation in *MED13* based on its gene specific mutation rate (Samocha et al. 2014) using an Exact Poisson Test in R [49].

Table 1: Clinical features of patients with MED13 mutations and molecular characterization

	Patient A	Patient B	Patient C	Patient D	Patient E	Patient F	Patient G	Patient H	Patient I	Patient J	Patient K	Patient L	Patient M
Molecular characterization													
cDNA variant (NM_005121.2)	c.125del	c.392T>G	c.392T>G	c.977C>T	c.975_977 delTAC	c.979C>T	c.980C>A	c.1618C>A	c.1745T>A	c.4198C>T	c.4487delC	c.6178C>A	c.6191C>T
Predicted protein effect	P42Lfs*6	L131*	L131*	T326I	T326del	P327S	P327Q	P540T	L582*	R1400*	T1496Mfs	Q2060K	A2064V
CADD v.1.3	31.0	37.0	37.0	25.0	20.5	23.4	25.2	26.3	40	41	35	24.1	25.7
GERP + + RS	5.67	5.32	5.32	5.5	5.5	5.5	5.5	6.16	6.02	4.58	5.86	6.04	6.04
Inheritance	De novo	Maternal (daughter of pt. C)	Unknown	De novo	De novo	De novo	De novo	De novo	De novo	De novo	De novo	De novo	De novo
Clinical characterization													
Gender	M	F	F	M	F	M	F	M	M	F	M	F	F
Age at last visit (years)	8	5	32	19	9	11	3	10	6	13	5	10	6
Height	+ 0.6 SD	+ 0.5 SD	average	− 0.2 SD	+ 2.2 SD	+ 0.5 SD	+ 0.3 SD	− 0.9 SD	0 SD	− 0.7 SD	− 2 SD	+ 0.5 SD	− 2 SD
Weight (for height)	+ 1.8 SD	+ 1.9 SD	average	− 0.8 SD	+ 1.2 SD	− 0.9 SD	− 1.1 SD	+ 0.6 SD	+ 0.5 SD	0 SD	− 1.6 SD	+ 0.4 SD	+ 0.7 SD
Head circumference	− 1 SD	+ 0.2 SD	NA	− 0.5 SD	+ 1.1 SD	− 0.9 SD	− 0.3 SD	− 1.5 SD	− 0.5 SD	NA	0 SD	− 2 SD	+ 1 SD
Intellectual Disability (ID) / Developmental Delay (DD)	Mild ID	Mild/ borderline ID	Borderline ID (IQ 80–85)	Mild ID (IQ 65)	Mild ID	Mild ID	DD	Borderline ID (IQ 85 with working memory score 68 on WISC-IV)	Mild ID (IQ 61)	Moderate ID	DD	DD	Mild/ borderline ID (IQ 70)
Speech delay/ disorder	+ Speech apraxia with mixed receptive and expressive language disorder, limited verbal expression and language-based learning disorder	+ Delayed speech development, mild articulation problems	+ Delayed speech development, expressive language problems in childhood. At adult age only sporadic and mild wordfinding problems	+ Delayed speech development, mild articulation problems, normal language comprehension	+ Moderate mixed receptive and expressive language disorder, decreased vocabulary and language formulation difficulties	+ Delayed expressive and receptive language	+ Mainly expressive speech problems	+/− Borderline (verbal comprehension score = 87 on WISCIV)	+ At age 6y expressive and receptive speech at age equivalent < 2y	+ Moderate expressive language disorder	+ Severe speech disorder with regression, speech apraxia, receptive language is fine	+ Severe speech delay, 5–10 words	+ Severe speech/ language disorder, expressive language most affected, signs of speech apraxia

Delayed motor development	+ (Only fine motor skills delayed)	– (walked at 14 m)	– (walked on time)	+ (walked at 20 m)	+ (walked at 25 m)	+ (walked at 22 m)	+ (walked at 26 m)	– (walked before 12 m)	+ (walked at 2 years; early delays, now mostly on target with peers)	NR	– (walked at 12 m)	–	+ (walked at 20 m)
Autism spectrum disorder (ASD) /ADHD	ADHD	NA	–	–	–	ASD	ASD	–	–	ASD, ADHD	ASD	ASD, ADHD	–
Brain MRI	Normal	NA	NA	Normal	Bulbous splenium of corpus callosum (likely normal variant)	Normal	Normal	Small area of abnormal signal in left occipital lobe	Normal	NA	NA	Normal	Mild frontal atrophy, otherwise normal
Eye/vision abnormalities	Astigmatism	probably amblyopia	–	Visual impair-ment, pale optic nerves	Congenital nystagmus, outer retinal atrophy temporal to both optic discs, optic nerves low normal in size on MRI	–	Strabismus, papil edema	–	Astigma-tism	NR	NR	Duane anomaly	Duane anomaly
Heart abnormalities	NR	–	NR	–	History of murmur, normal echo and ECG	Dilated aortic root and pulmonary artery	–	NR	NR	Subaortic stenosis	NR	NR	NR
Chronic obstipation	NR	NR	+	NR	+	NR	+	NR	NR	NR	NR	NR	+
Other features (features reported in two unrelated patients in bold)		Sloping shoulders, small and laterally deviated halluces	Small and laterally deviated halluce	Kyphosis, pes cavus	Hypotonia, mild proximal weakness, fatigues easily, clumsy gait, transient lactic acidosis with illness, congenital left hip dysplasia	Hypotonia, Conductive hearing loss, Mild scoliosis, pes cavus	Hypotonia	Epilepsy (drug-resistant with myoclonica-tonic seizures)		Chronic sleep disturbances	Chronic sleep disturbances	Conductive hearing loss, Precocious puberty	

NR not reported, *NA* not assessed

Results

Phenotypes

We collected detailed clinical information of 13 patients with rare, protein-altering *MED13* variants. Eleven variants were confirmed to be de novo, and one patient (patient B) inherited the variant from her mother who is also affected (patient C). Phenotypic data summarizing the spectrum of features of this cohort of 13 patients are shown in Table 1.

All patients had developmental delays with varying severity and course. In the patients that underwent formal intelligence testing, total IQ levels varied from 85 (lower range of normal IQ) to an IQ between 35 and 50 (moderate ID). Five patients had an Autism Spectrum Disorder (ASD), and three patients were diagnosed with Attention Deficit Hyperactivity Disorder (ADHD). All patients had speech delays and/or disorders, with delayed milestones in speech and language development. While several patients had expressive and receptive language problems, in the majority of patients, speech production was significantly more impaired than language comprehension. Three patients (patient A, K and M) showed characteristics of speech apraxia, a developmental speech disorder in which affected individuals have difficulties accurately programming the motor sequences required to produce fluent speech. Patient A had a mild ID, but showed speech apraxia with a mixed receptive and expressive language disorder, and limited verbal expression at the age of 8 years. Patient M had a non-verbal IQ of 70 with a severe speech/language disorder. Her expressive speech was severely affected, with signs of speech apraxia. At the age of 8 years she only used single words and very short sentences. Patient K developed some speech capabilities, but showed regression at the age of 13 months and has since remained non-verbal.

Seven of 13 patients showed delays in motor development, most of which affected at least the gross motor skills (6 of 7), although one patient was reported to have only fine motor delays. Three patients had hypotonia (patient E, F and G). One patient (patient H) developed severe drugresistant myoclonic-atonic epilepsy at 4 years of age with generalized clonic, myoclonic, atonic, tonic and atypical absence seizures. MRI screening of this patient showed a small abnormality in the left occipital lobe of his brain that did not correspond to the electrophysiological onset or the semiology of his seizures. In other patients, MRI scans were not performed or showed no clear abnormalities, except for mild frontal atrophy in patient M.

Eight patients (62%) presented with eye or vision abnormalities. Two patients (patients L and M) presented with Duane anomaly, a congenital type of strabismus that is characterized by non-progressive horizontal ophtalmoplegia and retraction of the globe with attempted adduction, together with narrowing of the palpebral fissure[19]. One patient (patient G) had strabismus, two patients had astigmatism (patient A and I), and one patient (patient E) had congenital nystagmus. While only one patient (patient D) had a visual impairment, three patients had optic nerve abnormalities: pale optic nerves in patient D, papilledema in patient G, and in patient E outer retinal atrophy temporal to both optic discs was reported with relatively small optic nerves on a MRI-scan.

Several other interesting phenotypes were observed in at least two patients in the cohort. Four patients presented with chronic obstipation (patients C, E, G and M). Two patients had conductive hearing loss (patients F and L). Two patients had congenital heart abnormalities: a mildly dilated aortic root and pulmonary artery (both improving over time) in patient F, and a subaortic stenosis in patient J. Two patients were reported to have chronic sleep issues (patient J and K).

Overlapping facial characteristics were reported, including widely spaced eyes with narrow palpebral fissures and peri-orbital fullness, a broad and high nasal bridge, full nasal tip, synophrys, a flat philtrum and a wide mouth with thin upper lip (Fig. 1).

Variants and predicted consequences

The *MED13* transcript (NM_005121.2) encodes a large protein consisting of 2174 amino acids (NP_005112.2). The Pfam database characterizes two domains within the MED13 protein: an N-terminal domain (aa 11–383) and a C-terminal domain (aa 1640–2163), as shown in Fig. 2a. Analysis of conservation across the length of the protein indicates several highly conserved residues that lie between these two domains (Fig. 2b).

All 12 unique variants found in our patients are absent from the gnomAD database[20] and TOPMED Bravo database (https://bravo.sph.umich.edu/freeze3a/hg19/) and are predicted to be highly deleterious by CADD v1.3[21], with scores ranging from 20.5 to 41 (Table 1). Six patients had five unique variants that are predicted to be truncating: three nonsense mutations (p.Leu131* in Patients B and C, p.Leu582* in Patient I and p.Arg1400* in Patient J) and two frameshift variants leading to a premature stop codon (p.Pro42Leufs*6 in patient A and p.Thr1496Metfs*11 in Patient K). The remaining variants include six missense variants and a single amino acid deletion. These seven variants form two apparent clusters: one in the N-terminal conserved phosphodegron domain and the other in the C-terminal domain (Fig. 2a). These seven variants were all found to lie within motifs that are highly conserved between MED13 and MED13L (Fig. 2b) and affect sites under high codon selection (Fig. 2c). These missense variants and the in-frame deletion are each located on surface-exposed sites within a three-dimensional model of the MED13 protein (Fig. 3). The four mutations that cluster in the N-terminal domain affect two adjacent amino acids (p.Thr326 and p.Pro327) that are known to be part of a conserved phosphodegron that is required for binding with SCF-Fbw7 ubiquitin ligase for degradation[22]. Using interaction data from Davis et al. and PDB structure 2OVQ, which has Fbw7 interacting with a similar motif as MED13, we modeled this interaction for MED13 followed by insertion of each variant and calculation of binding energy. All four variants (p.Thr326Ile, p.Thr326del, p.Pro327Ser, p.Pro327Gln) are predicted to alter the phosphorylation and Fbw7 interaction with drastic decreases in binding energy to Fbw7 (Figure S1). The two missense changes clustering in the C-terminal portion of the protein (p.Gln2060Lys and p.Ala2064Val; in patients L and M, respectively) were also studied in more detail. One of the changes (p.Ala2064Val) is predicted to be structurealtering through increasing hydrophobic collapse, secondary structure formation, and increasing aliphatic index of a surface exposed linear motif. This results in a decrease of the regions linear interacting peptide potential that is highly conserved and likely functional (Figure S2). The remaining missense variant (p.Pro540Thr in Patient H) lies within a highly conserved linear motif centered near amino acid 538 (Fig. 2b); it results in the formation of a high probability Casein Kinase 1 phosphorylation motif, which could lead to additional interaction with proteins containing forkhead- associated domains when analyzed through the ELM database[23] (Fig. 3).

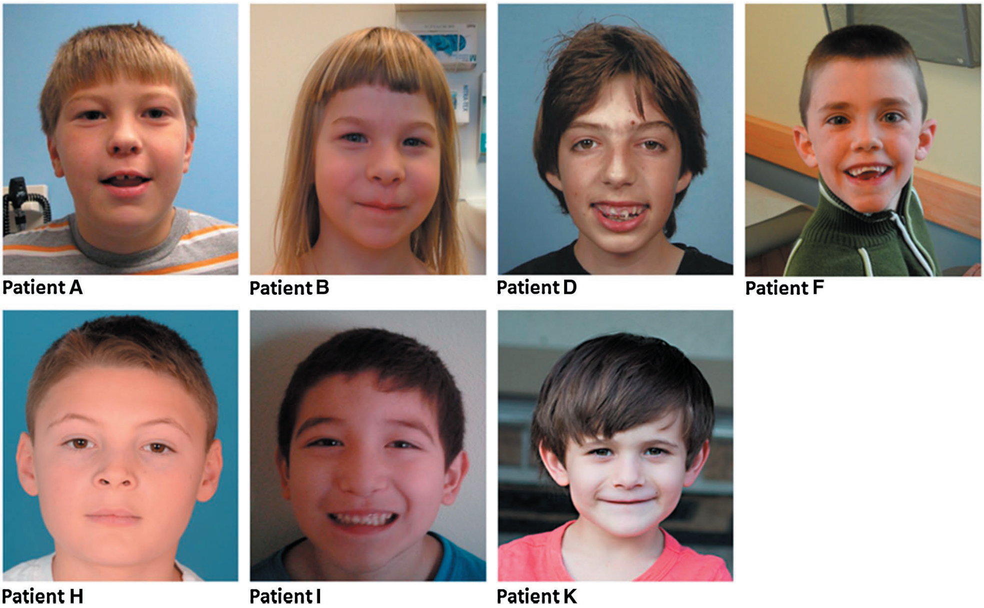


Figure 1: Facial phenotypes of seven individuals with a MED13 variant

Facial phenotypes of seven individuals with a MED13 variant. Overlapping facial characteristics include peri-orbital fullness, narrow palpebral fissures, a broad and high nasal bridge, full nasal tip, synophrys, flat philtrum, wide mouth and a thin upper lip.

Effects of truncating MED13 mutation on transcript and protein levels

As truncating mutations often lead to nonsense-mediated decay and haploinsufficiency, we aimed to examine the effects of a truncating *MED13* mutation on levels of *MED13* transcript and MED13 protein. We performed RT-PCR on cDNA transcribed from RNA of patient J, who was heterozygous for a nonsense mutation (c.4198C > T; p.Arg1400*). We compared the *MED13* transcript level of the patient to her biological parents and two healthy controls (Fig. 4a). No differences in *MED13* transcript levels were detectable between the affected patient and the unaffected parents or controls (One-way ANOVA p = 0.5913). Sanger sequencing of cDNA amplicons from the child demonstrated the presence of the aberrant transcript in the child (Fig. 4b), at ~ 70% levels relative to the normal transcript (Fig. 4c). To assess the effect of the nonsense mutation on protein levels, a western blot was performed on nuclear extracts from mononuclear blood cells of the patient and controls (Fig. 4d). While full-length MED13 protein was present in the patient (and in the controls), no truncated MED13 protein product could be detected. The MED13 protein level of the patient was not clearly different compared with the MED13 protein level of the father.

Enrichment of de novo MED13 variants in DD/ID cohorts

We quantified the extent of enrichment of de novo variants in *MED13* within DD/ID-affected probands. We used only the two largest cohorts considered within this study, each of which yielded at least two de novo *MED13* variants. Five patients described here (A, E, F, I, and K) come from a cohort of 11,149 affected individuals, and two patients, one of which is described here (patient L), were identified within the Deciphering Developmental Disorders (DDD) study of 4293 trios[24]. Both studies suggest a rate close to 1 de novo variant affecting *MED13* per ~ 2200 DD/ID-affected individuals. When comparing the number of observed de novo mutations in *MED13* to the expected number based on the gene specific mutation rate of *MED13* for missense, splicesite, nonsense and frameshift mutations [6.237×10^{-5} per chromosome[25]], we find evidence for a significant enrichment among DD/ID-affected individuals (7 variants in 30,884 alleles; $p = 0.00371$).

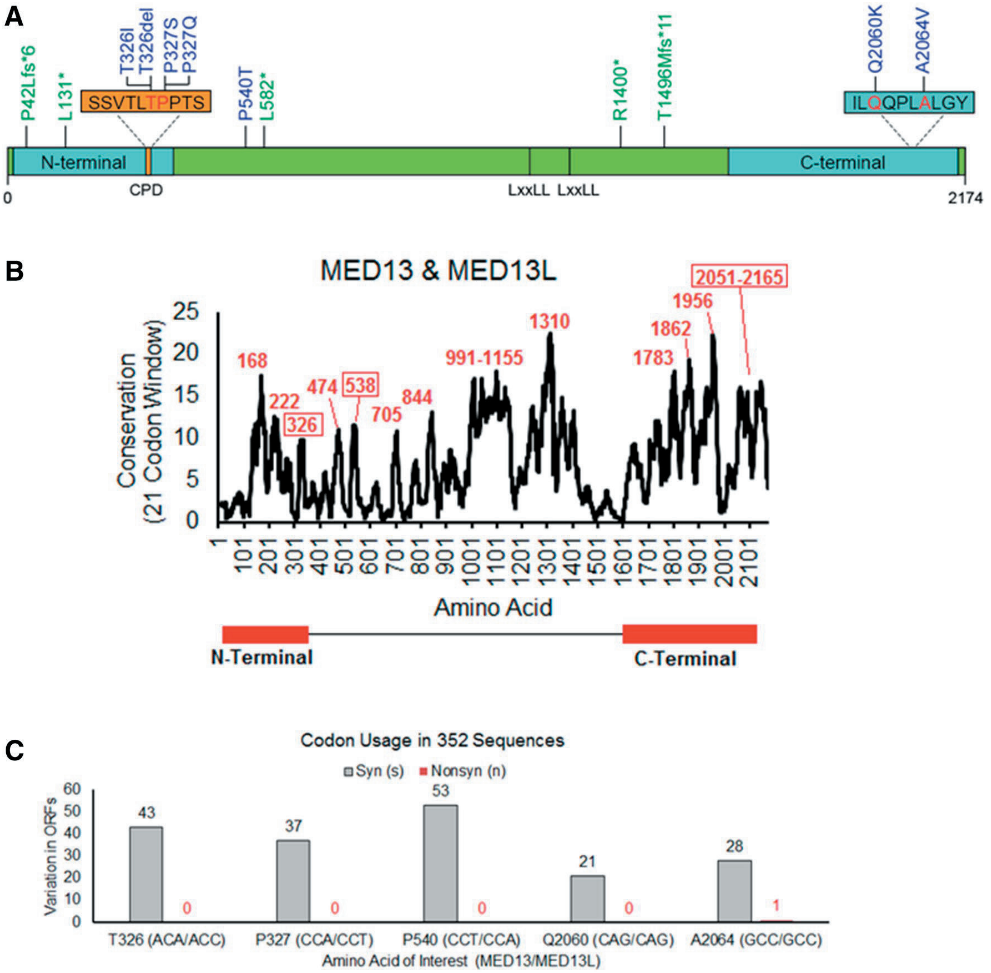


Figure 2: Analysis of mutations: location, conservation and codon usage of variant sites

(a) Identified mutations are shown within a linear representation of the MED13 protein, consisting of 2174 amino acids. Missense mutations and the in-frame deletion are shown in blue, and nonsense and frameshift mutations in green. Six of the seven non-truncating mutations in our MED13 cohort cluster in two small regions within the N-terminal and C-terminal domains of the MED13 protein. Affected amino acids p.Thr326 and p.Pro327 and are part of a conserved phosphodegron (CPD), which is shown in orange. Two LxxLL nuclear receptor-binding motifs are also noted. **(b)** Analysis of conservation throughout the protein was performed using amino acid selection scores as previously published [18], using a 21 codon sliding window for both MED13 and MED13L aligned such that the most selected motifs of a protein are identified as peaks. The center of each highly conserved linear motif is labeled and those containing variants described in this paper are boxed. **(c)** Codon usage throughout evolution for the locations of all missense mutations and the in-frame deletion. All five sites are under high selection with multiple synonymous (Syn, gray) amino acids in 352 open reading frames (ORFs) of MED13 and MED13L with only a single nonsynonymous (Nonsyn, red) change. Numbers indicate instances where ORFs in other species deviate from the conserved codon usage. Of note, for three locations (326, 327 and 540) the codon used differs between MED13 and MED13L with the amino acid conserved. In these cases, numbers indicate where ORFs in other species deviate from conserved codon usage in their respective ortholog.

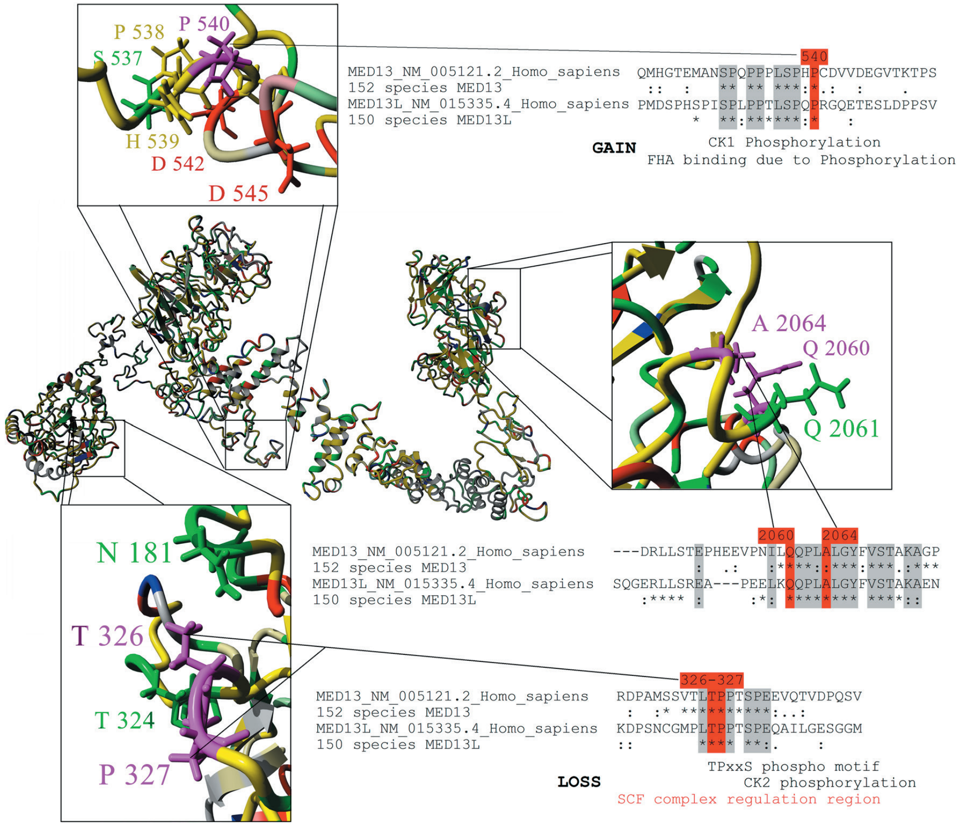

P 538
P 540
S 537
H 539
D 542
D 545
540
MED13_NM_005121.2_Homo_sapiens QMHGTEMANSPQPPPLSPHPCDVVDEGVTKTPS
152 species MED13
MED13L_NM_015335.4_Homo_sapiens PMDSPHSPISPLPPTLSPQPRGQETESLDPPSV
150 species MED13L
GAIN
CK1 Phosphorylation
FHA binding due to Phosphorylation
A 2064
Q 2060
Q 2061
N 181
T 326
T 324
P 327
2060
2064
MED13_NM_005121.2_Homo_sapiens ---DRLLSTEPHEEVPNILQQPLALGYFVSTAKAGP
152 species MED13
MED13L_NM_015335.4_Homo_sapiens SQGERLLSREA---PEELKQQPLALGYFVSTAKAEN
150 species MED13L
326-327
MED13_NM_005121.2_Homo_sapiens RDPAMSSVTLTPPTSPEEVQTVDPQSV
152 species MED13
MED13L_NM_015335.4_Homo_sapiens KDPSNCGMPLTPPTSPEQAILGESGGM
150 species MED13L
LOSS
TPxxS phospho motif
CK2 phosphorylation
SCF complex regulation region

Figure 3: Location of missense mutations and in-frame deletion in threedimensional structure of MED13 and conservation of affected amino acids

A full model of MED13 protein created with I-TASSER modeling was combined with 152 species sequences for MED13 using ConSurf mapping. Amino acid coloring is as followed: gray = not conserved, yellow = conserved hydrophobic, green = conserved hydrophilic, red = conserved polar acidic, blue = conserved polar basic, magenta = conserved human variants of interest. A zoomed in view of the three different affected regions are shown, along with amino acid alignments from MED13 and MED13L. An asterisk (*) indicates 100% conservation in all sequences and a colon (:) indicates functional conservation. Linear motifs mapped with the Eukaryotic Linear Motif tool are shown below sites for 326–327 and 540.

Discussion

By molecular and clinical characterization of a cohort of 13 patients with variants in *MED13*, we here provide evidence for a new neurodevelopmental disorder. This *MED13*-associated syndrome is characterized by DD/ ID with speech delay and/or speech disorders. Additionally a broad spectrum of other common features is seen, including ASD, ADHD, various eye abnormalities and mild facial dysmorphisms. Based on the phenotypes of patients presented here, we do not yet see a clear genotype- phenotype correlation between type and location of the mutations and severity of clinical features. However, it is notable that the two patients with Duane anomaly have a missense mutation in a similar location in the C-terminal domain of the MED13 protein, and that the optic nerve abnormalities are reported in patients with mutations affecting residues p.Thr326 or p.Pro327 only.

MED13 is a component of the CDK8-kinase module, which can reversibly bind the Mediator complex. Mediator is a multi-protein complex that is required for assembly and stabilization of the pre-initiation complex, which is essential for transcription initiation[26,27]. The core function of Mediator is to transmit signals from various transcription factors to RNA polymerase II (Pol II)[28]. Binding of the CDK8-module to Mediator has been reported to prevent the association of Mediator with the Pol II preinitiation complex, thus preventing transcription initiation and/or re-initiation. In this way, the CDK8-module is considered a key molecular switch in Pol II mediated transcription[29]. MED13, as well as the other subunits of the CDK8-module, are known to be critical regulators of developmental gene expression programs in Drosophila, zebrafish and *C. elegans*[30,31]. MED13, or its paralog MED13L, forms a direct connection of the CDK8 module with the core Mediator complex[32], and protein turnover of MED13 (or MED13L) may be critical in modulating the pools of Mediator-CDK8 kinase complex in cells[22,29,33].

Three missense mutations (p.Thr326Ile, p.Pro327Ser and p.Pro327Gln) and one in-frame-deletion (p.Thr326del) in our cohort are likely to affect MED13 protein turnover due to their location within a conserved phosphodegron. This phosphodegron is recognized by the SCF-Fbw7 ubiquitin ligase, which targets the MED13 protein for ubiquitination and degradation[22]. In fact, it has already been shown that a specific amino acid substitution at position 326 in MED13 (p.Thr326Ala) leads to impaired binding of Fbw7 to the phosphodegron of MED13/ MED13L, thus preventing MED13/MED13L ubiquitination and degradation[22]. Therefore, a variant at this position may lead to increased levels of MED13 protein in the cell. As Fbw7 is proposed to target only MED13 or MED13L proteins that are bound to the core Mediator complex[22], these mutations may have an effect on the CDK8 module-Mediator association and subsequently on transcription regulation. The potential effects of the p.Pro540Thr missense variant are also intriguing. Protein modeling suggests that this variant could introduce an additional Casein Kinase 1 phosphorylation site, thus potentially increasing interactions with forkhead-associated domains involved in protein–protein interactions.

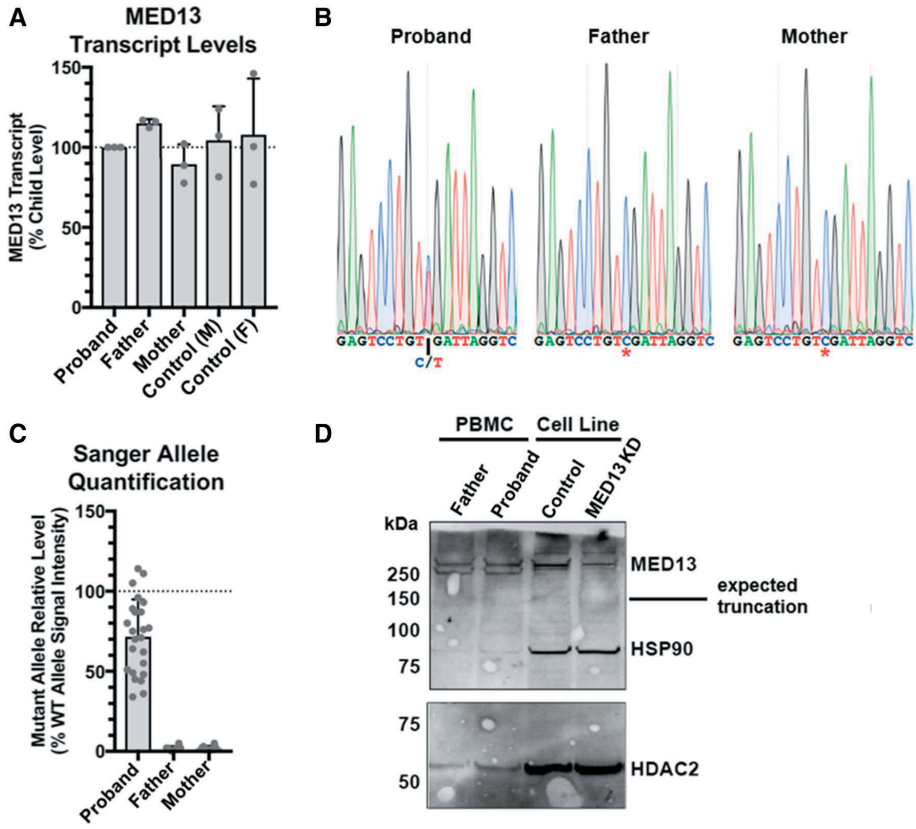


Figure 4: Analysis of transcript and protein levels in patient with nonsense mutation

(a) Level of MED13 transcript was measured by qPCR and normalized to GAPDH and proband (patient J). No differences were detectable between groups (One-way ANOVA $p = 0.5913$). An additional loading control (AGPAT) produced very similar results (data not shown). **(b)** Representative Sanger traces from cDNA amplicons demonstrating the presence of the variant in the proband, and absence in the father and mother. **(c)** Quantification of the chromatograms of all Sanger sequences reveals less signal from the base on the mutant allele ($p < 0.0001$ by paired t-test compared to the wildtype base signal by trace). The father and mother do not have any signal at the mutant base above the level of noise. **(d)** Western blot for MED13 (and HSP90 and HDAC2 as loading controls) from nuclear extracts of patient peripheral blood mononuclear cells or a neural precursor cell line (present to demonstrate antibody specificity with a knockdown (KD) control). If the nonsense mutation resulted in a stable protein, a product at approximately 150 kDa would be expected, which is not present. No protein was recoverable from the blood sample from the mother.

We also observed five unique mutations predicted to truncate MED13. In assessments of RNA and protein levels in Patient J and her unaffected parents, the variant transcript was detected in the proband but no truncated protein could be observed. While these results are inconclusive with regards to the molecular mechanism of pathogenicity in this particular proband, loss-of-function mechanisms remain an attractive possibility. Patterns of variation in *MED13* in human population databases indicate that *MED13* is relatively intolerant to loss-of-function variation; MED13 has a Rare Variant Intolerance Score (RVIS) that ranks among the top 1.66% of all genes[34] and an ExAC pLI score of 1.00[20].

We show an enrichment of de novo *MED13* mutations compared to what is expected under a null model (*p* = 0.00371) in two large ID/DD patient cohorts. We acknowledge that this *p* value does not exceed a genomewide evidence threshold and by itself proves association. However, the enrichment *p* value does not account for five de novo variants described here from smaller cohorts that were discovered independent of, and prior to, assessment of the statistical evidence from the larger cohorts. We also observed clustering of missense mutations in our cohort, which by itself is an argument for pathogenicity[35]. Additionally, independent genetic studies also support the disease relevance of variation in *MED13*. There is one report of an 800-kb microdeletion including *MED13* and five other genes in a patient with moderate ID, short stature, mild dysmorphisms, and hearing loss[36]; the authors proposed *MED13* as the most likely causal candidate gene. Additionally, a de novo frameshift (p.Pro286Leufs*86) and a de novo variant that likely affects splicing (D + 3; c.814+3A>G) were observed in a cohort of 2508 probands with ASD[37], and three rare protein-altering variants in *MED13* (p.Ala418Thr, p.Arg512*, p.Tyr1649*) were also found in a separate ASD cohort[38].

Other Mediator subunits, including other CDK8-kinase module-associated disease genes, have been associated with various neurodevelopmental disorders. Variants in *MED12* have been associated with ID syndromes with congenital abnormalities, including Opitz-Kaveggia syndrome (MIM 305,450)[39], Lujan-Fryns syndrome (MIM 309,520)[40] and X-linked Ohdo syndrome (MIM 300,896)[41]. Mutations in *MED12* have also been associated with intellectual disability. In addition to ID and speech delays both MED12 patients and several MED13 probands described here present with eye abnormalities (eye movement disorders, and abnormalities of the retina and optic nerves)[42,43] and chronic obstipation[43,44]. In addition to the MED12 subunit, a disruption of *CDK19* was reported in a patient with ID, microcephaly and congenital retinal folds[45].

It is of particular relevance to this study that variation in the *MED13*-paralog *MED13L* has been shown to cause a neurodevelopmental disorder as well[46]. Given the similar molecular roles for MED13 and MED13L, we aimed to compare and contrast phenotypes presented by both groups of individuals using information provided in the literature. The main phenotypic characteristics of MED13L-associated syndrome are (borderline) ID with delayed speech and language development, and a variable spectrum of other features including autism, hypotonia, characteristic facial features and heart defects[7-9,47,48]. Many of these features clearly overlap with the phenotypes in our *MED13* cohort. However, similar to the heterogeneity observed here in patients with *MED13* variation, the spectrum of phenotypes observed among *MED13L* mutation carriers is quite broad. The identification and detailed phenotyping of additional patients with *MED13* and *MED13L* mutations is needed to elucidate the complete spectrum of associated features, and to reveal the similarities and differences between the two syndromes.

We believe that the data presented in this study coupled to the additional evidence available from other studies strongly support the conclusion that rare protein-altering variation in *MED13* underlie a new neurodevelopmental disorder. Key results from this study include: a significant enrichment of de novo mutations in *MED13* within ID/DD cohorts (p = 0.00371); the clustering and conservation levels of the positions affected by the observed missense variation (Fig. 2a, b); the computationally predicted deleteriousness of the observed mutations (Table 1; Fig. 3, Fig. S1); and the overlap of phenotypic features among the 13 patients presented here, including speech difficulties (13/13), intellectual disability (at least 9/13), and eye or vision problems (8/13). Supporting evidence from other studies include: the existence of mutations affecting *MED13* in at least six independent families affected by pediatric neurodevelopmental disorders; the intolerance of *MED13* to mutations in the general human population (pLI = 1.00, RVIS score of 1.66%); and the previously established disease-associations of several other Mediator subunits, including *MED13L*, a functionally related paralog of *MED13*. While the precise pathogenic mechanisms have yet to be elucidated—some of the mutations observed here are predicted to stabilize MED13 protein while others are predicted to lead to loss-offunction— we find it highly likely that mutational disruption of normal MED13 function leads to disease, adding *MED13* to the list of Mediator-associated, in particular CDK8-kinase module-associated, neurodevelopmental disorders.

Acknowledgements

We thank all patients and families for their contributions. This work was supported by the Max Planck Society (S.E.F.), a grant from the US National Human Genome Research Institute (NHGRI; UM1HG007301) (S.M.H., K.M.B., J.N.C., K.L.E., E.M.B., G.M.C.), and the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 24.001.006 to the Language in Interaction Consortium (L.S.B., H.G.B., S.E.F.). Patient L was part of the DDD study cohort (DECIPHER ID: DDD263479). The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003], a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute [grant number WT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. This study makes use of DECIPHER (http://decip her.sange r.ac.uk), which is funded by the Wellcome Trust.

References

1. Boat TF, Wu JT (eds) (2015) mental disorders and disabilities among low-income children. Washington (DC).
2. Vissers LE, Gilissen C, Veltman JA (2016) Genetic studies in intellectual disability and related disorders. Nat Rev Genet 17:9–18.
3. Au PYB et al (2015) GeneMatcher aids in the identification of a new malformation syndrome with intellectual disability, unique facial dysmorphisms, and skeletal and connective tissue abnormalities caused by de novo variants in. HNRNPK Hum Mut 36:1009–1014
4. Harms FL et al (2017) Mutations in EBF3 disturb transcriptional profiles and cause intellectual disability, ataxia, and facial dysmorphism Am J Hum Genet 100:117–127.
5. Kernohan KD et al. (2017) Matchmaking facilitates the diagnosis of an autosomal-recessive mitochondrial disease caused by biallelic mutation of the tRNA isopentenyltransferase (TRIT1) gene. Hum Mut 38:511–516.
6. Sobreira N, Schiettecatte F, Valle D, Hamosh A (2015) GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Human Mutation 36:928–930.
7. Adegbola A et al (2015) Redefining the MED13L syndrome Eur. J Hum Genet 23:1308–1317.
8. Muncke N et al (2003) Missense mutations and gene interruption in PROSIT240, a novel TRAP240-like gene, in patients with congenital heart defect (transposition of the great arteries. Circulation 108:2843–2850.
9. van Haelst MM, Monroe GR, Duran K, van Binsbergen E, Breur JM, Giltay JC, van Haaften G (2015) Further confirmation of the MED13L haploinsufficiency syndrome. Eur J Hum Genet 23:135–138.
10. Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway JW (2005) The mammalian Mediator complex and its role in transcriptional regulation. Trends Biochem Sci 30:250–255.
11. Malik S, Roeder RG (2005) Dynamic regulation of pol II transcription by the mammalian Mediator complex. Trends Biochem Sci 30:256–263.
12. Bowling KM et al (2017) Genomic diagnosis for children with intellectual disability and/or developmental delay. Genome Med 9:43.
13. de Ligt J et al (2012) Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367:1921–1929.
14. Deciphering Developmental Disorders S (2015) Large-scale discovery of novel genetic causes of developmental disorders. Nature 519:223–228.
15. Neveling K et al (2013) A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Hum Mut 34:1721–1726.
16. Sollis E et al (2017) Equivalent missense variant in the FOXP2 and FOXP1 transcription factors causes distinct neurodevelopmental disorders. Human mutation 38:1542–1554.
17. Tanaka AJ et al (2015) Mutations in SPATA5 are associated with microcephaly, intellectual disability, seizures, and hearing loss. Am J Human Genetics 97:457–464.
18. Prokop JW, Lazar J, Crapitto G, Smith DC, Worthey EA, Jacob HJ (2017) Molecular modeling in the age of clinical genomics, the enterprise of the next generation. J Mol Model 23:75.
19. Andrews CV, Hunter DG, Engle EC (1993) Duane Syndrome. In: Pagon RA et al (eds) GeneReviews(R). Seattle (WA)
20. Lek M et al (2016) Analysis of protein-coding genetic variation in 60,706. Hum Nat 536:285–291.
21. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315.
22. Davis MA, Larimore EA, Fissel BM, Swanger J, Taatjes DJ, Clurman BE (2013) The SCF-Fbw7 ubiquitin ligase degrades MED13 and MED13L and regulates CDK8 module association with. Mediator Genes Dev 27:151–156.
23. Dinkel H et al (2016) ELM 2016—data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res 44:D294-300.
24. Deciphering Developmental Disorders S (2017) Prevalence and architecture of de novo mutations in developmental disorders. Nature 542:433–438.
25. Samocha KE et al (2014) A framework for the interpretation of de novo mutation in human disease. Nat Genet 46:944–950.
26. Chen XF et al (2012) Mediator and SAGA have distinct roles in Pol II preinitiation complex assembly and function. Cell Rep 2:1061–1067.
27. Hantsche M, Cramer P (2017) Conserved RNA polymerase II initiation complex structure Curr. Opin Struct Biol 47:17–22.

28. Allen BL, Taatjes DJ (2015) The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Biol 16:155–166.
29. Knuesel MT, Meyer KD, Bernecky C, Taatjes DJ (2009) The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function. Genes Dev 23:439–451.
30. Carrera I, Janody F, Leeds N, Duveau F, Treisman JE (2008) Pygopus activates Wingless target gene transcription through the mediator complex subunits Med12 and Med13. Proc Natl Acad Sci USA 105:6644–6649.
31. Poss ZC, Ebmeier CC, Taatjes DJ (2013) The Mediator complex and transcription regulation. Crit Rev Biochem Mol Biol 48:575–608.
32. Daniels DLF, Schwinn M, Benink MK, Galbraith H, Amunugama MD, Jones R, Allen R, Okazaki D, Yamakawa N, Futaba H, Nagase M, Espinosa T, Urh JM, M. (2013) Mutual Exclusivity of MED12/ MED12L, MED13/13L, and CDK8/19 paralogs revealed within the CDK-mediator kinase module. J Proteom Bioinf S2
33. Tsai KL, Sato S, Tomomori-Sato C, Conaway RC, Conaway JW, Asturias FJ (2013) A conserved mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction. Nat Struct Mol Biol 20:611–619.
34. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 9:e1003709.
35. Lelieveld SH et al (2017) Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorderassociated genes. Am J Hum Genet 101:478–484.
36. Boutry-Kryza N et al (2012) An 800 kb deletion at 17q23.2 including the MED13 (THRAP1) gene, revealed by aCGH in a patient with a SMC 17. Am J Med Genet A 158A:400–405.
37. Iossifov I et al (2014) The contribution of de novo coding mutations to autism spectrum disorder. Nature 515:216–221.
38. Yuen RK et al (2017) Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci 20:602–611.
39. Risheg H et al (2007) A recurrent mutation in MED12 leading to R961W causes Opitz–Kaveggia syndrome. Nat Genet 39:451–453.
40. Schwartz CE et al (2007) The original Lujan syndrome family has a novel missense mutation (p.N1007S) in the MED12 gene. J Med Genet 44:472–477.
41. Vulto-van Silfhout AT et al (2013) Mutations in MED12 cause X-linked Ohdo syndrome. Am J Human Genetics 92:401–406.
42. Clark RD et al (2009) FG syndrome, an X-linked multiple congenital anomaly syndrome: the clinical phenotype and an algorithm for diagnostic testing. Genet Med 11:769–775.
43. Donnio LM et al (2017) MED12-related XLID disorders are dosedependent of immediate early genes (IEGs) expression. Hum Mol Genet 26:2062–2075.
44. Lyons MJ (1993) MED12-Related Disorders. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Stephens K, Amemiya A (eds) GeneReviews((R)). Seattle (WA).
45. Mukhopadhyay A et al (2010) CDK19 is disrupted in a female patient with bilateral congenital retinal folds, microcephaly and mild mental retardation. Hum Genet 128:281–291.
46. Asadollahi R et al (2013) Dosage changes of MED13L further delineate its role in congenital heart defects and intellectual disability. Eur J Hum Genet 21:1100–1104.
47. Caro-Llopis A, Rosello M, Orellana C, Oltra S, Monfort S, Mayo S, Martinez F (2016) De novo mutations in genes of mediator complex causing syndromic intellectual disability: mediatorpathy or transcriptomopathy? Pediatr Res 80:809–815.
48. Martinez F, Caro-Llopis A, Rosello M, Oltra S, Mayo S, Monfort S, Orellana C (2017) High diagnostic yield of syndromic intellectual disability by targeted next-generation sequencing. J Med Genet 54:87–92.
49. R Core Team (2013) R: A language and environment for statistical computing http://www.r-proje ct.org. Vienna A, R Foundation for Statistical Computing.

Supplemental Figures

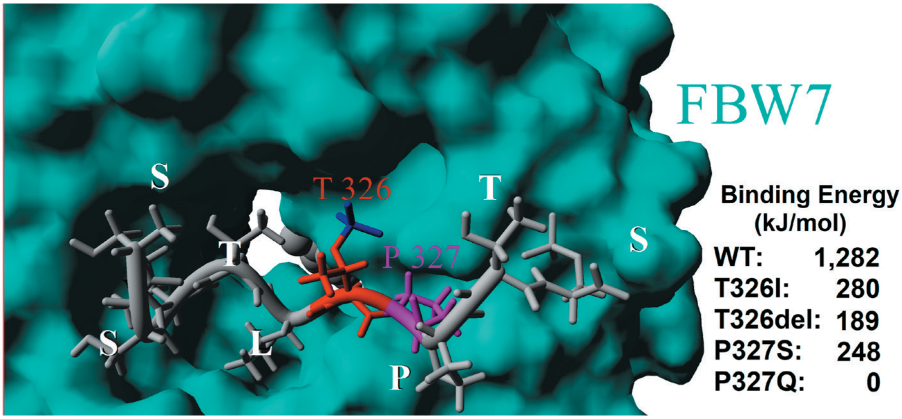


Figure S1: Analysis of variants at position 326-327 in relation to Fbw7-interaction

The interaction of MED13 with Fbw7 was modeled, by using PDB structure 2OVQ and amino acids 321-330 of the MED13 protein. All four different variants in our cohort that affect this binding region (T326I, T326del, P327S, P327Q) were subsequently inserted in the model, and binding energy was calculated using AMBER14 force field (http://ambermd.org/) in YASARA. All four variants are predicted to alter the phosphorylation and Fbw7 interaction with a severe decrease in binding energy to Fbw7 (PNG 2012 KB).

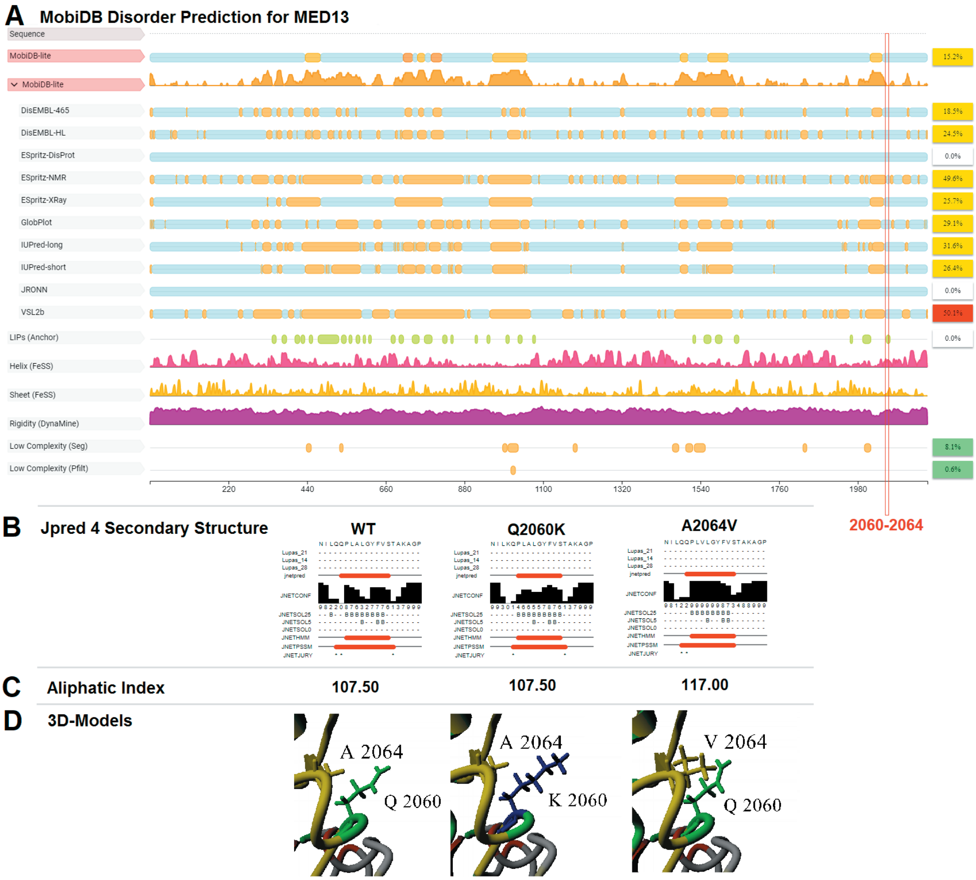


Figure S2: Structural packing of MED13 and variants at amino acid position 2060-2064

(a) MobiDB breakdown (http://mobidb.bio.unipd.it/Q9UHV7/predictions) for MED13 showing structural disorder (orange) from various databases, linear interacting peptides (LIPs, green) helical prediction (pink), beta sheet prediction (light orange), and rigidity (magenta). The 2060-2064 region is boxed in red with low prediction of disorder and a predicted LIP from 2060-2070, suggesting this highly conserved surface exposed region has a high potential to form secondary structure when bound to some unknown protein binding partner. **(b)** Jpred4 secondary structure predictions(Drozdetskiy et al. 2015) showing predicted changes in secondary structures, with a score of 9 being most likely to form secondary structure at each residue. p.Ala2064Val has highest probability to form stable secondary structure (average residue score of 7.05 compared to the wild type WT 5.6 and p.Gln2060Lys of 5.7). A variant increasing secondary structure would decrease formation rates with the unknown binding partner, thus likely resulting in loss of binding. **(c)** Aliphatic index score(Ikai 1980) showing p.Ala2064Val to increase thermostability of the linear motif. An increase in intrinsic thermostability likely decreases formation rates with the unknown binding partner similar to secondary structure predictions. **(d)** Location and effect of the two missense mutations p.Gln2060Lys and p.Ala2064Val shown on our predicted models for the region (PNG 351 KB).

3

Chapter 3

De *novo* variants disturbing the transactivation capacity of POU3F3 cause a characteristic neurodevelopmental disorder

Lot Snijders Blok, Tjitske Kleefstra, Hanka Venselaar, Saskia Maas, Hester Y. Kroes, Augusta M.A. Lachmeijer, Koen L.I. van Gassen, Helen V. Firth, Susan Tomkins, Simon Bodek, The DDD Study, Katrin Õunap, Monica H. Wojcik, Christopher Cunniff, Katherine Bergstrom, Zoë Powis, Sha Tang, Deepali N. Shinde, Catherine Au, Alejandro D. Iglesias, Kosuke Izumi, Jacqueline Leonard, Ahmad Abou Tayoun, Samuel W. Baker, Marco Tartaglia, Marcello Niceta, Maria Lisa Dentici, Nobuhiko Okamoto, Noriko Miyake, Naomichi Matsumoto, Antonio Vitobello, Laurence Faivre, Christophe Philippe, Christian Gilissen, Laurens van de Wiel, Rolph Pfundt, Pelagia Deriziotis, Han G. Brunner and Simon E. Fisher

American Journal of Human Genetics 2019
doi: 10.1016/j.ajhg.2019.06.007

Abstract

POU3F3, also referred to as Brain-1, is a well-known transcription factor involved in the development of the central nervous system, but it has not previously been associated with a neurodevelopmental disorder. Here, we report the identification of 19 individuals with heterozygous *POU3F3* disruptions, most of which are *de novo* variants. All individuals had developmental delays and/or intellectual disability and impairments in speech and language skills. Thirteen individuals had characteristic low-set, prominent, and/or cupped ears. Brain abnormalities were observed in seven of eleven MRI reports. *POU3F3* is an intronless gene, insensitive to nonsense-mediated decay, and 13 individuals carried protein-truncating variants. All truncating variants that we tested in cellular models led to aberrant subcellular localization of the encoded protein. Luciferase assays demonstrated negative effects of these alleles on transcriptional activation of a reporter with a *FOXP2*-derived binding motif. In addition to the loss-of-function variants, five individuals had missense variants that clustered at specific positions within the functional domains, and one small in-frame deletion was identified. Twomissense variants showed reduced transactivation capacity in our assays, whereas one variant displayed gain-of-function effects, suggesting a distinct pathophysiological mechanism. In bioluminescence resonance energy transfer (BRET) interaction assays, all the truncated POU3F3 versions that we tested had significantly impaired dimerization capacities, whereas all missense variants showed unaffected dimerization with wild-type POU3F3. Taken together, our identification and functional cell-based analyses of pathogenic variants in *POU3F3*, coupled with a clinical characterization, implicate disruptions of this gene in a characteristic neurodevelopmental disorder.

POU3F3 (MIM: 602480) encodes a member of class III of the POU family of transcription factors. These proteins all carry a characteristic POU domain that binds with high affinity to a specific octamer (5'-ATGCAAAT-3') or closely related DNA sequences in the enhancers and promoters of various different target genes[1]. The importance of POU3F3 for the developing brain is reflected in its original name Brain-1 (Brn1)[2]. Best known as a marker of upper-layer projection neurons in the cortex[3], POU3F3 is implicated in the regulation of many key processes in the development of the central nervous system; these processes include cortical neuronal migration[4], upper-layer specification and production, and neurogenesis[5–7]. However, the phenotypic consequences of pathogenic germline variants in human *POU3F3* are currently unknown.

We identified a *de novo* missense variant disrupting *POU3F3* in a female with a severe developmental speech and language disorder, autism spectrum disorder, and mild intellectual disability. This variant was absent in control databases and affected a highly conserved amino acid, and *in silico* analyses consistently predicted deleterious effects on the function of the encoded protein. In addition, we noted a case report describing a *de novo* chromosome 2q12.1 deletion in a male with intellectual disability; in the report, *POU3F3* haploinsufficiency was discussed as a possible pathogenic mechanism[8]. Chromosome 6q16.1 deletions that span the closely related, co-expressed ortholog, *POU3F2* (also known as Brn2; MIM: 600494), cause a neurodevelopmental disorder with obesity[9]. Moreover, it has been shown that *FOXP2* (MIM: 605317), disruptions of which cause a developmental speech disorder (speech language disorder-1; MIM: 602081), contains an intronic binding site for POU3F2[10,11].

We used matchmaking initiatives such as Gene- Matcher[12] and the Decipher Database[13] to identify additional individuals with rare germline variants in *POU3F3*. Here, we delineate the characteristic phenotypic features and mutational spectrum of a cohort of 19 individuals with pathogenic variants in *POU3F3*. Nearly all (17 out of 19) individuals had a *de novo* variant in the gene, and all variants were identified via exome sequencing with a trio approach. One person (individual 18) had inherited the variant from an affected mother (individual 19); in this family, exome sequencing had been performed in the proband and the mother only (duo approach).

The phenotypes of all 19 individuals with *POU3F3* variants were systematically assessed and analyzed. A summary of the most common phenotypic characteristics is shown in Table 1, and more details per individual can be found in Table S1. All individuals with a *POU3F3* variant in our cohort (19/19; 100%) had developmental delays (DD) and/or an intellectual disability (ID). The level of functioning was broad, ranging from severe ID in two individuals to borderline intellectual functioning (WPPSI-IQ 77) in one individual. For individuals in which the severity of ID and/or DD was known, the majority (8/10; 80%) had a borderline to moderate level of ID and/or DD.

Table 1: Summary of Phenotypes in Individuals with POU3F3 Variants

	Non-Truncating Variants	**Truncating Variants**	**All Variants Combined**	
Feature	**Amount**	**Amount**	**Amount**	**Percentage**
Development				
Developmental delay (DD) and/or intellectual disability (ID)	6/6	13/13	19/19	100%
Borderline or mild ID	2/6	3/13	5/19	26%
Moderate ID	1/6	2/13	3/19	16%
Severe ID	2/6	0/13	2/19	11%
DD or ID, severity unknown	1/6	8/13	9/19	47%
Speech delay or disorder	6/6	13/13	19/19	100%
Autism Spectrum Disorder (ASD) diagnosis	1/6	6/13	7/19	37%
Neurology				
Abnormalities reported on brain MRI	3/4	4/7	7/11[a]	64%
Hypotonia	3/5	7/13	10/18[a]	56%
Epilepsy	2/6	0/13	2/19	11%
Drooling	2/5	7/9	9/14[a]	64%
Other features				
Cupped and/or low-set ears	3/6	13/13	16/19	84%
Vision problems	2/6	8/13	10/19	53%
Sleeping problems (often waking up at night)	1/5	4/9	5/14[a]	36%
Cryptorchidism	0/1	3/10	3/11[b]	27% of males

A more detailed overview of phenotypic features per individual can be found in Table S1.

[a] Feature not assessed or not known for all 19 individuals in the cohort.

[b] Feature only applicable to the 11 males in the cohort.

Given that the first proband in our study (individual 3 in Table S1) had a severe developmental speech and language disorder, we paid special attention to the speech and language capacities across the entire cohort. All individuals with *POU3F3* variants had delayed speech and language development, often with a delayed onset of producing first words. In two children who spoke their first words at an appropriate age, a halt in development or regression of speech in the first years of life was reported. Although both receptive and expressive language problems were reported, in many children expressive speech capacities were more impaired than language comprehension. Almost all individuals received or had received speech therapy, and commonly reported speech-related problems were oral motor problems, word finding problems, and social communication issues. In addition to this, drooling was reported in 9/14 individuals (64%), and open mouth behavior was seen in four individuals.

Many individuals had autism-like features, and a formal diagnosis of autism spectrum disorder (ASD) was made in 7/19 individuals (37%). Although 3/19 (16%) individuals reached their motor milestones in time, most had delays in both fine and gross motor development. Hypotonia was reported in 10/18 individuals (56%). The two individuals with severe ID each had a form of epilepsy: Lennox- Gastaut syndrome with tonic-clonic seizures in one individual and a severe seizure disorder with myoclonic seizures, drop attacks, absences, and tonic-clonic seizures in the other. In both individuals, capillary hemangiomas were reported; this feature is not present in the rest of the cohort. Magnetic resonance imaging (MRI) revealed cerebral atrophy in both individuals; additionally, white matter cysts were present in one of them. In total, brain anomalies were reported in 7 of the 11 individuals (64%) in which a brain MRI was performed. Anomalies observed in at least two individuals were delayed myelination, cerebral atrophy, and corpus callosum abnormalities.

No significant additional congenital abnormalities were noted in our collected cohort of 19 individuals. Vision problems, which mainly included (mild) refraction errors and strabismus, were reported in 10/19 individuals (53%). Hearing loss was present in one individual, and narrow auditory canals were reported in two other individuals. Although growth parameters were generally normal, small hands with short and broad digits, especially broad thumbs, as well as flat feet and high-arched feet were reported in some of the individuals. Five children (5/14; 36%) had sleeping problems at a young age, waking up several times per night. Three of the eleven males (27%) had cryptorchidism.

A comparison of facial features revealed a striking overlap in dysmorphisms in individuals with *POU3F3* variants. The cupped, prominent, and often low-set ears, present in 16 of the 19 individuals (84%), were most remarkable. Other common facial features included full lips, an openmouth appearance, a broad and bulbous nasal tip, hypertelorism, epicanthal folds, and peri-orbital fullness (Figure 1).

The types of *POU3F3* variants in our cohort were diverse and included nonsense variants, frameshift variants, missense variants, and an in-frame deletion of five amino acids. All variants in this study are annotated with respect to the GenBank: NM_006236.2 transcript. None of the identified variants were present in the gnomAD database. The pLI-score of *POU3F3* in gnomAD is 0.89, and no high-confidence truncating variants in this gene are present in this dataset, indicating that *POU3F3* is especially intolerant for loss-of-function variance. In 13 of the 19 individuals, we found 12 different nonsense or frameshift variants, predicted to truncate POU3F3. For most genes, when such variants arise, the post-transcriptional surveillance mechanism of nonsense-mediated mRNA decay (NMD) helps to prevent translation of aberrant truncated versions of proteins[14]. In mammalian cells this surveillance mechanism is tightly linked to pre-mRNA splicing[15]. Because *POU3F3* is an intronless gene, aberrant transcripts with truncating variants are insensitive toNMD and so will still be expressed. These truncating variants were distributed widely across *POU3F3* (Figure 2A) and are thus predicted to yield truncated proteins of a range of different sizes.

Five individuals in our cohort had a missense variant. All were located in one of the two known functional domains of POU3F3: the POU-specific (POU-S) domain and the POU-homeobox (POU-H) domain (Figure 2A). Even with this relatively small number of missense variants, a clear clustering was seen: two unrelated individuals had an identical *de novo* missense variant (c.1085G>T, [p.Arg362Leu]), and another two individuals had missense variants that affected the same amino acid (c.1219C>G, [p.Arg407Gly] and c.1220G>T, [p.Arg407Leu]). One individual had a c.1367A>G, (p.Asn456Ser) substitution. In addition to the missense variants, one individual had an in-frame deletion (c.992_1006del, [p.Gln331_ Lys335del]), which removes five amino acids from within the POU-specific domain of the encoded protein. For all the missense variants and the in-frame deletion, highly conserved residues are affected (Figure 2B). All the missense variants of our cohort are predicted to be pathogenic by both PolyPhen- 2 and SIFT and have high CADD scores (range 26.9–32.0; Table S1). We also visualized the non-truncating variants in a tolerance landscape of POU3F3 by using the MetaDome web server. The tolerance landscape was computed as a missense over synonymous ratio, on the basis of single nucleotide variants in gnomAD in the proteincoding part of *POU3F3*. All the non-truncating variants were located in regions of the protein that are extremely intolerant to missense variation (Figure S1).

The two known functional domains of POU3F3, connected via a flexible linker, are both required for site-specific DNA binding with high affinity[16]. The POU-S domain forms four alpha-helices; several direct and sequence-specific hydrogen bonds are made between residues in the third helix and the DNA[17]. In the POU-H domain, the third helix is also responsible for sequence-specific DNA-binding. Two of the missense variants, c.1085G>T, (p.Arg362Leu) and c.1367A>G, (p.Asn456Ser), affect residues that are located in the third helix of the POU-S domain and the POU-H domain, respectively (Figure 2B).

We used three-dimensional protein modeling to further investigate the potential functional impact of the identified non-truncating variants. The PDB file PDB: 2XSD[18] (POUdomain of POU3F1 bound to DNA) provided a template for modeling the DNA-binding region of POU3F3, which spans amino acids 316–466 (Figure 2C). Thismodel revealed that two of the amino acids (Arg362 and Asn456) that are substituted in our cohort are directly involved in binding the major groove of target DNA, consistent with the prior literature on other POU proteins[16,17]. In our model, Arg362 forms hydrogen bonds with a guanine base in the DNA of the transcription-factor binding site (Figure S2). Substitution of a leucine residue at this point of the protein is predicted to abolish DNA-binding. Similarly, Asn456 is directly involved in DNA-binding by forming hydrogen bonds with adenine, an interaction that will be disrupted by a substitution of serine at this position (Figure S2).

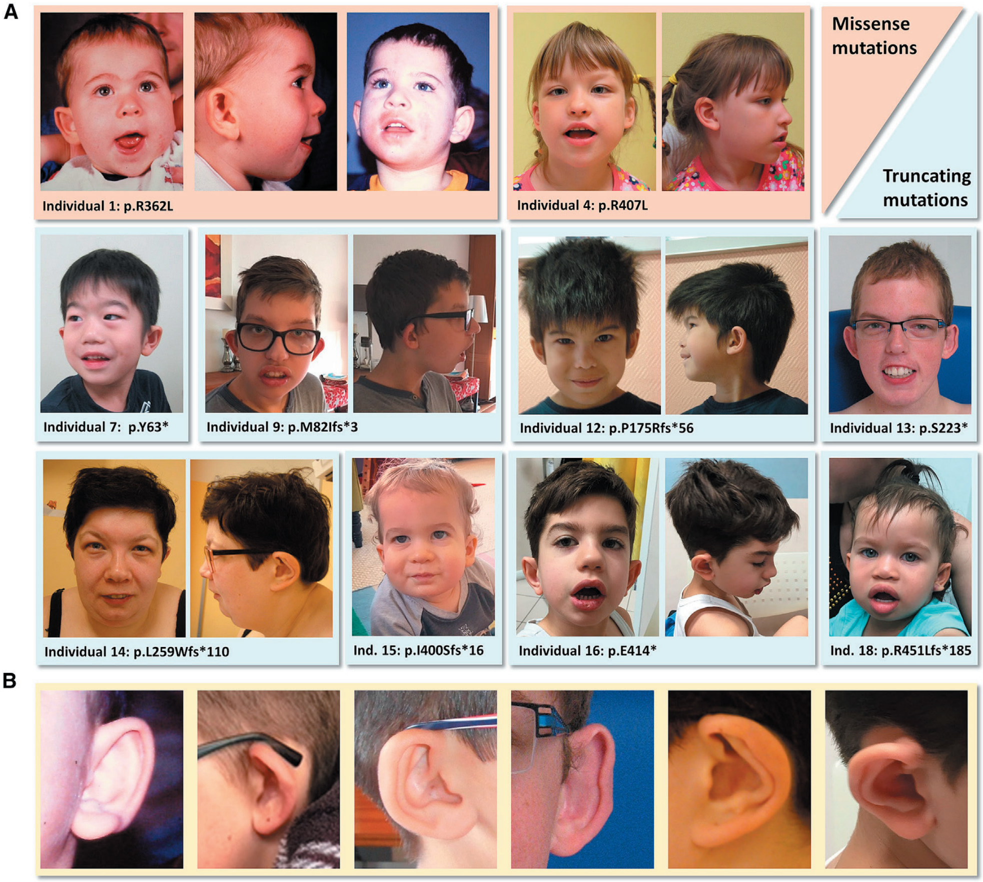


Figure 1: Facial Features of Ten Individuals with a Pathogenic POU3F3 Variant

(a) All individuals in this picture have cupped and/or prominent and often low-set ears, except for individual 4. Other overlapping features are full lips, an open-mouth appearance, thick ear helices, a broad and bulbous nasal tip, hypertelorism, epicanthal folds, and periorbital fullness. (b) Magnification of the ear abnormalities in individuals 1, 9, 12, 13, 14, and 16, respectively.

The remaining two missense variants that we identified are located at position 407 on the edge of the homeodomain and at the flexible linker in our linear representation of POU3F3 (Figure 2B). In the three-dimensional model, Arg407 lies in the flexible linker between both POU domains, but it is unclear what impact a substitution at this point would have on protein function (Figure 2C). Lastly, the in-frame deletion in our cohort is located in the POU-S domain. Although the amino acids that are deleted do not directly bind to DNA themselves, it is likely that their loss will alter domain structure and indirectly disturb DNA-binding capacities.

To assess the potential functional effects of the *POU3F3* variants, we performed a variety of complementary cellbased assays. We expressed a representative set of nine POU3F3 variants, as well as wild-type (WT) POU3F3, as fusions to YFP tags in HEK293 cells. The set of

POU3F3 variants included all four missense variants, the in-frame deletion, and four of the truncating variants. Immunoblot analysis showed that all the expressed YFP-fusion proteins had the expected molecular weights (Figure S3).

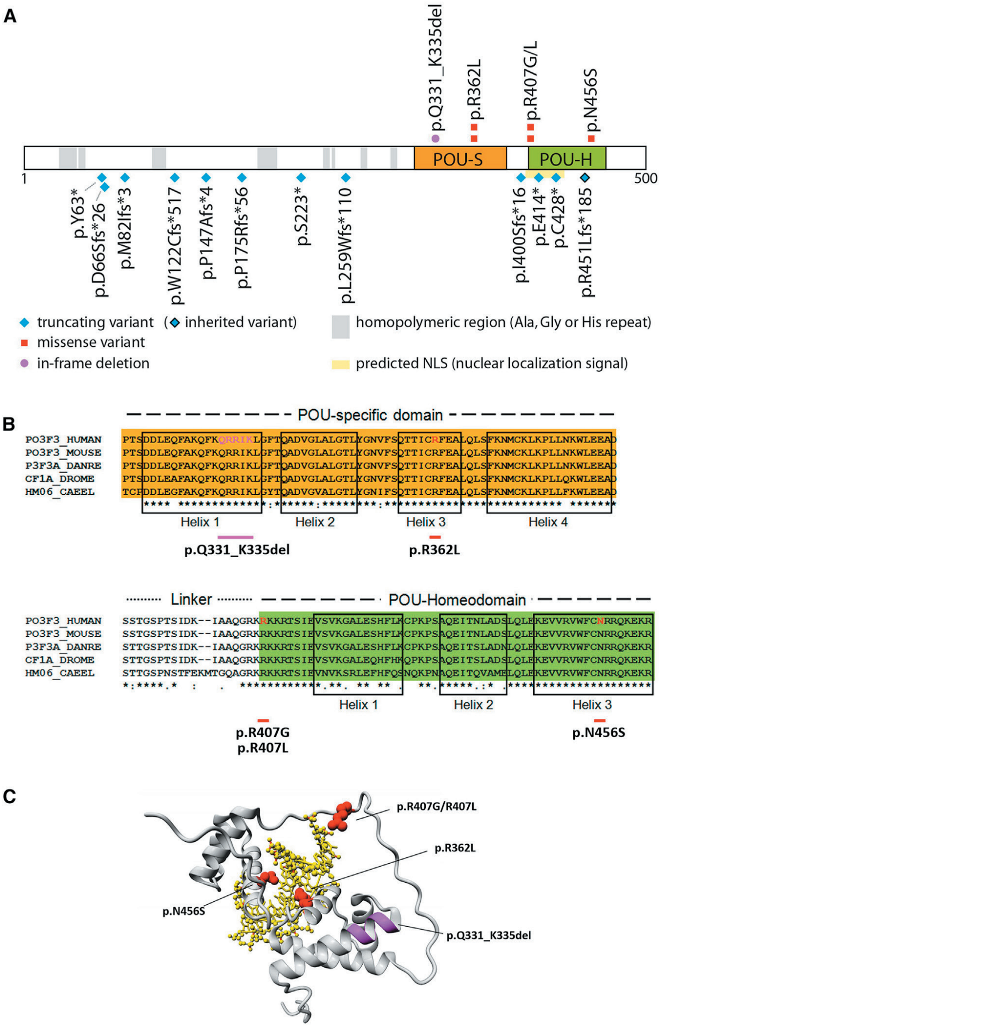

A
p.Q331_K335del
p.R362L
p.R407G/L
p.N456S
POU-S
POU-H
1
500
p.Y63*
p.D66Sfs*26
p.M82Ifs*3
p.W122Cfs*517
p.P147Afs*4
p.P175Rfs*56
p.S223*
p.L259Wfs*110
p.I400Sfs*16
p.E414*
p.C428*
p.R451Lfs*185
◆ truncating variant (◆ inherited variant)
■ missense variant
● in-frame deletion
homopolymeric region (Ala, Gly or His repeat)
predicted NLS (nuclear localization signal)
B
POU-specific domain
PO3F3_HUMAN PTSDDLEQFAKQFKQRRIKLGFTQADVGLALGTLYGNVFSQTTICRFEALQLSFKNMCKLKPLLNKWLEEAD
PO3F3_MOUSE PTSDDLEQFAKQFKQRRIKLGFTQADVGLALGTLYGNVFSQTTICRFEALQLSFKNMCKLKPLLNKWLEEAD
P3F3A_DANRE PTSDDLEQFAKQFKQRRIKLGFTQADVGLALGTLYGNVFSQTTICRFEALQLSFKNMCKLKPLLNKWLEEAD
CF1A_DROME PTSDDLEAFAKQFKQRRIKLGFTQADVGLALGTLYGNVFSQTTICRFEALQLSFKNMCKLKPLLQKWLEEAD
HM06_CAEEL TCPDDLEGFAKQFKQRRIKLGYTQADVGVALGTLYGNIFSQTTICRFEALQLSFKNMCKLKPLLFKWLEEAD
**** *************:******:********:******************************** *******
Helix 1
Helix 2
Helix 3
Helix 4
p.Q331_K335del
p.R362L
Linker
POU-Homeodomain
PO3F3_HUMAN SSTGSPTSIDK--IAAQGRKRKKRTSIEVSVKGALESHFLKCPKPSAQEITNLADSLQLEKEVVRVWFCNRRQKEKR
PO3F3_MOUSE SSTGSPTSIDK--IAAQGRKRKKRTSIEVSVKGALESHFLKCPKPSAQEITNLADSLQLEKEVVRVWFCNRRQKEKR
P3F3A_DANRE STTGSPTSIDK--IAAQGRKRKKRTSIEVSVKGALESHFLKCPKPSAQEITSLADNLQLEKEVVRVWFCNRRQKEKR
CF1A_DROME STTGSPTSIDK--IAAQGRKRKKRTSIEVSVKGALEQHFHKQPKPSAQEITSLADSLQLEKEVVRVWFCNRRQKEKR
HM06_CAEEL STTGSPNSTFEKMTGQAGRKRKKRTSIEVNVKSRLEFHFQSNQKPNAQEITQVAMELQLEKEVVRVWFCNRRQKEKR
*:****.* : . ************.**. ** ** . **.*****.:* .*********************
Helix 1
Helix 2
Helix 3
p.R407G
p.R407L
p.N456S
C
p.R407G/R407L
p.R362L
p.N456S
p.Q331_K335del

Figure 2: Facial Features of Ten Individuals with a Pathogenic POU3F3 Variant

(a) Linear representation of POU3F3 (Uni- Prot: P20264) showing the location of variants from unrelated families in this cohort. There are twelve truncating variants (blue), five missense variants (red), and one in-frame deletion (magenta). POU-S (orange) is the POU-specific domain, and POU-H (green) is the POUhomeodomain. The shown NLS (nuclear localization signal) prediction is derived from cNLS Mapper[27]. An overview with mutation details per subject is provided in Table S1. (b) Alignment of part of the POU3F3 amino acid sequence (using ClustalW) with orthologous sequences from the following species: *Mus musculus, Danio rerio, Drosophila melanogaster, and Caenorhabditis elegans.* Helix boundaries are defined as previously described[16]. (c) Three-dimensional modeling of the functional domains of POU3F3 binding to a target DNA sequence (yellow). Amino acids that are affected by the missense variants are shown in red (wild-type side chains are depicted), and the location of the in-frame deletion is shown in magenta. A more detailed picture for two missense variants, p.Arg362Leu and p.Asn456Ser, can be found in Figure S2.

We assessed the subcellular localization of the mutant proteins by using direct fluorescence imaging (Figure 3). Although two missense variants (p.Arg407Leu and p.Arg407Gly) map within a computationally predicted nuclear localization signal (NLS) motif (Figure 2A), none of the missense variants disturbed subcellular localization in this assay. All four tested proteins with a missense variant were located in the nucleus in a similar manner to WT. However, all the other tested constructs showed abnormalities in subcellular localization patterns compared to WT. For three truncating constructs (c.196_197delinsT, [p.Asp66Serfs*26], c.668C>A, [p.Ser223*], and c.1197delG, [p.Ile400Serfs*16]), aberrant cytoplasmic expression was noted, in addition to the normal nuclear expression of the protein. For two of these constructs (p.Ser223* and p.Ile400Serfs*16) we observed protein aggregates just around the nuclear membrane in a subset of cells, possibly indicating degradation of mutant protein (Figures 3 and S4). Aberrant localization patterns within the nucleus were observed for the c.1284C>A, (p.Cys428*) and c.992_1006del, (p.Gln331_Lys335del) proteins, and the former showed small nuclear aggregates in a minority of cells (Figures 3 and S4).

We next investigated whether the variants affect the transcription factor activity of the encoded protein. POU3F3 belongs to the POU family of transcription factors and is known to share important roles in neurodevelopment with its close paralog POU3F2[5,6]. *In vitro* experiments suggest that POU3F2 is able to activate an intronic binding site in *FOXP2*[10,11], a gene that has been implicated in a rare neurodevelopmental disorder mainly characterized by severe speech problems (MIM: 602081)[19]. We hypothesized that POU3F3 might also be able to activate transcription via this binding site within *FOXP2*. To test this hypothesis, we performed luciferase assays in which a YFP-fusion protein with POU3F3 or POU3F2 was expressed together with a Firefly luciferase construct containing the conserved FOXP2 binding site (Figure S5), as well as a Renilla luciferase construct, providing a normalization control (Figure 4A). POU3F3 was able to increase luciferase expression as strongly as POU3F2; there was a six-fold increase in expression compared to the negative control (Figure 4B). This finding indicates that the known intronic binding site for POU3F2 can also serve as a functional binding site for POU3F3.

We used the same luciferase assay to compare our POU3F3 variant constructs with the POU3F3 WT construct. All four POU3F3 constructs with truncating variants showed a severe impairment in transcriptional activation function (Figure 4C). The relative luciferase expression for these variants was similar to that for the negative control (a YFP-expression vector without POU3F3), consistent with the complete or partial loss of the DNA-binding POU domains of POU3F3. Three of the non-truncating variants (p.Arg362Leu, p.Asn456Ser, and p.Gln331_Lys335del) showed partial transactivation capacity that was significantly lower than that of the WT construct. The p.Arg407Leu variant led to a significant increase in relative luciferase expression compared to the WT construct. No significant difference compared to WT was seen for the other missense variant at this position (p.Arg407Gly). In summary, all variants except for the p.Arg407Gly substitution led to significantly disturbed transactivation capabilities in our assays.

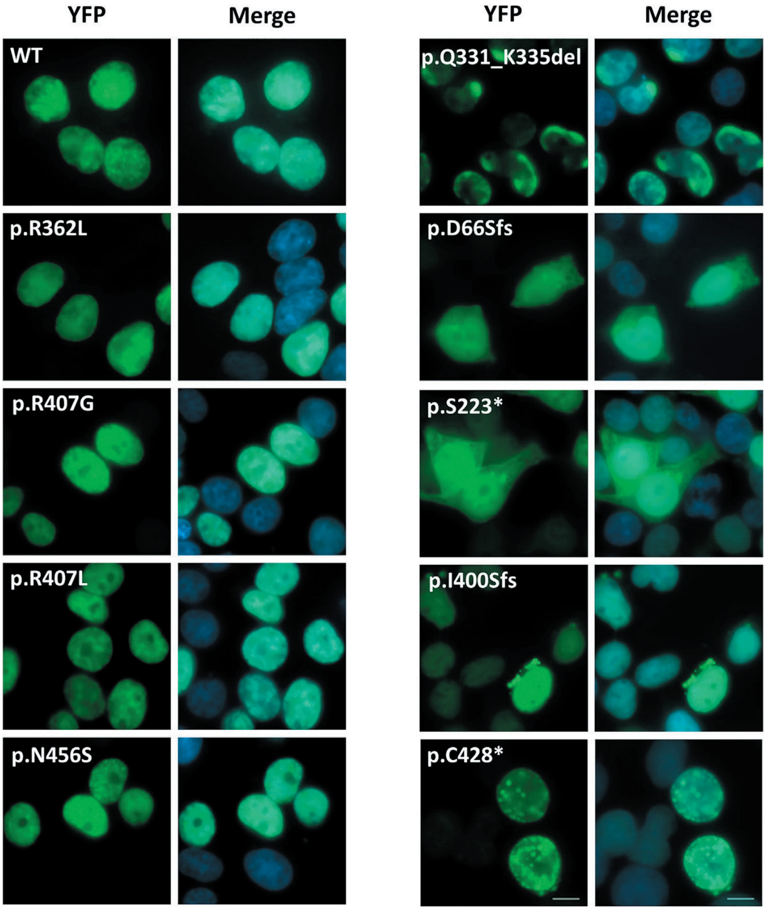


Figure 3: Subcellular Localization

(Direct fluorescence imaging of HEK293 cells expressing YFP-POU3F3 fusion proteins carrying different variants found in our cohort (green). The nuclei are stained with DAPI (blue). The scale bar ¼ 10mm. Pictures showing the aberrant subcellular localization patterns in a larger amount of cells for the variants p.Gln331_Lys335del, p.S223*, p.Ile400Serfs*16, and p.Cys428* can be found in Figure S4.

POU proteins are well known to have highly conserved dimerization properties[10]. They bind to target genes as monomers or dimers and can form either homo-dimers or hetero-dimers involving other family members[20]. To investigate whether the variants in our cohort affected the dimerization capacities of POU3F3, we used bioluminescence resonance energy transfer (BRET), a sensitive live-cell assay, to test putative protein-protein interactions.[21] In our assays, the bioluminescent donor construct encodes a Renilla luciferase (RLuc) fusion protein, and the fluorescent acceptor construct encodes a protein fused to YFP. If the proteins of interest are in close proximity, energy transfer can take place from donor to acceptor. We tested the ability of each mutant protein to form dimers with WT POU3F3 (Figure 5A) and with itself (Figure 5B). In these experiments the missense variants showed generally intact dimerization capacity, although the interactions for the p.Asn456Ser variant were slightly decreased. Two variants that are predicted to cause an early truncation of POU3F3 (p.Asp66Serfs*26 and p.Ser223*) showed a complete loss of dimerization capacity in both conditions. The two other truncating constructs (p.Ile400Serfs*16 and p.Cys428*) showed a less severe decrease, although the dimerization capacity was still significantly different from that of the WT construct. The p.Gln331_Lys335 protein showed impaired dimerization with WT POU3F3 but normal capacities for forming homo-dimers.

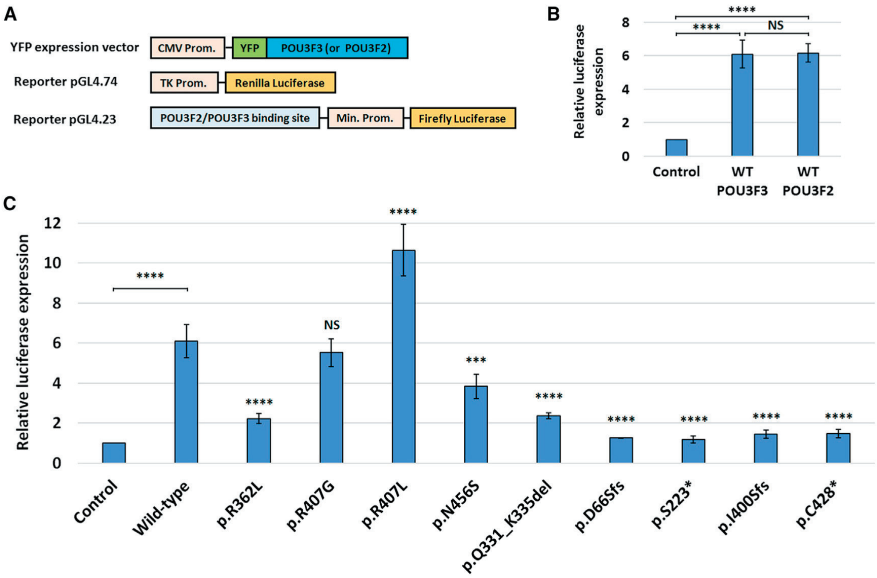


Figure 4: Luciferase Assays

(a) Expression constructs used in the luciferase assays: a YFP-fused POU3F3 or POU3F2 construct with a CMV promoter; a Firefly luciferase reporter construct with a minimal promoter and a preceding intronic FOXP2-derived binding site; and a control construct with Renilla luciferase under control of a TK promoter. (b) Results of luciferase assays with WT POU3F3 and WT POU3F2 and the reporter construct with the FOXP2-derived binding site. Values are expressed relative to the control and represent the mean5SD of three independent experiments, each performed in triplicate (**** ¼ $p < 0.0001$ and NS ¼ not significant, using one-way ANOVA and a post-hoc Tukey's test). (c) Results of luciferase assay withWT POU3F3 and nine constructs with POU3F3 variants. Values are expressed relative to the control and represent the mean5SD of three independent experiments, each performed in triplicate (*** ¼ $p < 0.001$; **** ¼ $p < 0.0001$; and NS ¼ not significant when compared to WT POU3F3 using one-way ANOVA and a post-hoc Dunnett's test).

All in all, the results of our clinical and molecular characterization show that diverse variants at different locations within *POU3F3* lead to a neurodevelopmental disorder with overlapping symptoms. When comparing genotypes and phenotypes within the cohort, several findings are of interest. First, two individuals (individuals 1 and 2) have a distinct and more severe phenotype compared to the rest of the cohort; this phenotype includes severe ID, epilepsy, and capillary hemangiomas. These individuals are unrelated but have an identical p.Arg362Leu variant. In luciferase assays, this variant showed impaired transcription- activation capacity, but it was not lower than that observed for other mutant constructs. Although the reason for the more severe and distinct phenotype associated with the p.Arg362Leu variant remains unclear, a dominantnegative effect is one possibility, given that the mutant protein showed normal subcellular localization and dimerization capacities in our assays.

Second, several pathogenic POU3F3 variants appear to be associated with characteristic facial features, especially the prominent, often cupped, and low-set ears. These ear abnormalities were reported independently in all individuals with a truncating variant and in the individuals with the p.Arg362Leu and p.Gln331_Lys335del variant (Figure 1 and Table S1). Prominent ears were also reported in the previously published individual with a microdeletion that included *POU3F3*.[8] The cupped or prominent ears are not present in the three individuals with the missense variants p.Arg407Leu, p.Arg407Gly, and p.Asn456Ser. The two missense variants affecting amino acid Arg407 did not show loss-of-function effects in our luciferase assay; in fact, p.Arg407Leu showed evidence of a possible gain-of-function. The p.Asn456Ser variant had a mild loss-of-function effect on transcriptional activity. These results suggest that both loss-of-function and gain-of-function mechanisms of different severity might lead to neurodevelopmental disorders with differences on a phenotypic level.

Although the missense variants at positions Arg362 and Asn456 mediate DNA-binding in the major groove, this is not the case for Arg407. This residue is located in the flexible linker between the POU-domains. POU3F2 and POU3F3 are known to be flexible in terms of spacing preference[16], meaning that they can bind to short binding motifs that are separated by 0, 2, or 3 bp, in contrast to other POU proteins that have more fixed preferences. Findings from a study of POU3F2 suggest that the highly conserved arginine residue at a position analogous to POU3F3 residue 407 is one of the residues that form a critical region in regulating the spacing preference of the protein[16]. Binding activity experiments showed that mutation of this critical region on the edge of the flexible linker and the homeodomain leads to less flexibility in spacing preference for the POU protein[16].

The missense variants p.Arg407Gly and p.Arg407Leu show different effects in our luciferase assay: although p.Arg407Gly did not show any difference compared with theWT POU3F3 construct, the p.Arg407Leu variant showed a gain-of-functioneffect. It isunclearwhythese twodifferent variants affecting the same Arg407 residue show different effects on transactivation capacity and how this relates to pathogenic mechanisms. Possibly, the missense variants at Arg407 alter the spacing properties of the encoded POU3F3 protein, and the alteration might affect transcriptional activation depending on the characteristics of the binding site. The architecture of regulatory DNA sites has been shown to influence the structure and organization of POU dimerization, interaction with other proteins, and

DNA-binding properties and can therefore be critical in determining the functionality of a transcription factor[18,22].

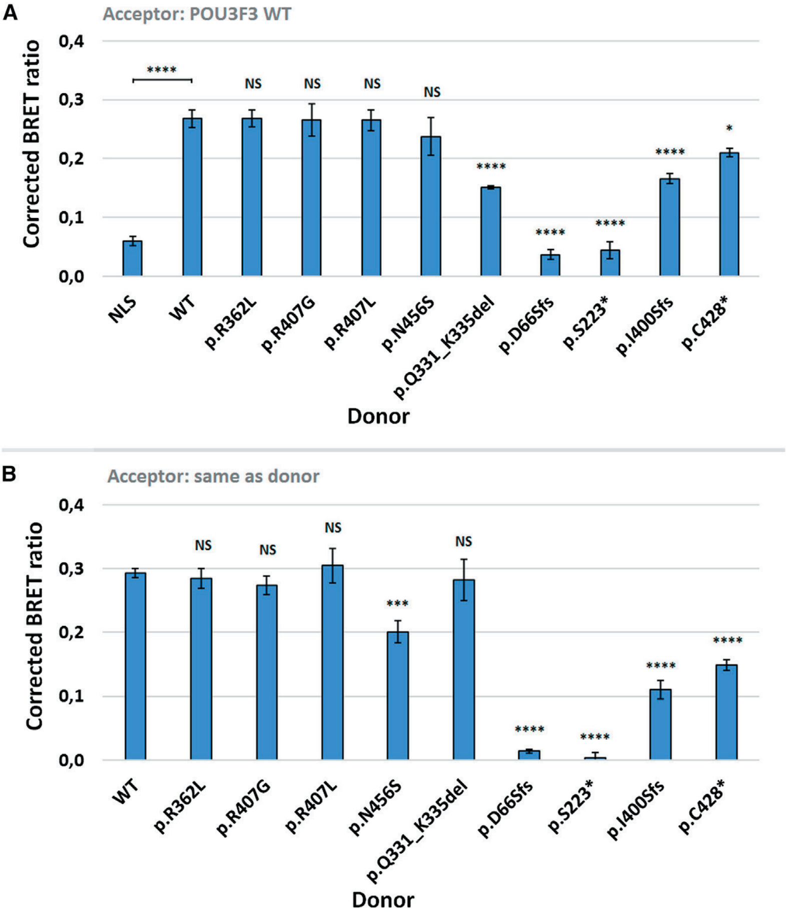


Figure 5: Bioluminescence Resonance Energy Transfer Assays

(a) Bioluminescence resonance energy transfer (BRET) assays to measure interactions between WT POU3F3 and mutant POU3F3 constructs. Bars represent the corrected mean BRET ratio 5 SD of three independent experiments performed in triplicate (**** ¼ p < 0.0001; * ¼ p < 0.05; and NS ¼ not significant when compared to WT using one-way ANOVA and a post-hoc Tukey's test). The NLSdonor construct is a Renilla luciferase construct with a nuclear localization signal. (b) BRET assays to measure homodimerization capacity of each mutant POU3F3 construct. Bars represent the corrected mean BRET ratio 5 SD of three independent experiments performed in triplicate (**** ¼ p < 0.0001; *** ¼ p < 0.001; and NS ¼ not significant when compared to WT using one-way ANOVA and a posthoc Tukey's test)

POU3F3 is highly similar to POU3F2; it has nearly identical (98.7%) amino acid sequences for the POU domains and the flexible linker. The main differences are found within the N-terminal region, which contains homopolymeric repeats that can function as transcriptional activation domains[1,23]. POU3F3 and POU3F2 share some roles and have partially redundant functions in cortical development[4,5]. Nonetheless, our study and the previously published study on POU3F2 haploinsufficiency[9] underscore the fact that two functional copies of both POU3F3 and POU3F2 are required for normal neurodevelopment. Microdeletions that span *POU3F2* have been shown previously to cause a neurodevelopmental disorder with obesity[9]. In contrast to this *POU3F2*-related disorder, pathogenic variants in *POU3F3* do not seem to be associated with obesity, because this feature is only reported in one of the 19 individuals in our cohort. In addition to the microdeletions encompassing *POU3F2*, a single *de novo* missense variant in *POU3F2* has recently been reported, but the specific location of this variant does not correspond to any variant reported here for *POU3F3*[24].

Our functional data indicate that a known POU3F2 regulatory site mapping within the *FOXP2* locus11 can also be bound by POU3F3. By using this binding site, we could develop luciferase-based assays to index the transactivation capacities of POU3F3 proteins carrying different etiological variants. Nevertheless, it remains undetermined whether pathogenic *POU3F2* and/or *POU3F3* variants actually have a significant impact on *FOXP2* expression in the proper genomic context *in vivo*. Future studies (for example by directly testing for *FOXP2* misregulation in individuals with pathogenic *POU3F3* variants) might shed light on whether putative functional links between the different genes have physiological relevance for the speech and language impairments observed in the associated neurodevelopmental disorders[25].

We emphasize that exome sequencing coverage is variable for *POU3F3*; the 5' half of the gene has poor coverage and the 3' part has good coverage[26]. So if the characteristic facial phenotype as shown in Figure 1 is recognized in an individual with an overlapping neurodevelopmental phenotype, it might be prudent to re-assess any existing next-generation-sequencing data and/or perform targeted sequencing of *POU3F3*. The specific variants identified in this study were all covered by a sufficient number of allele counts in gnomAD, and none of these alleles were found in this large dataset[26].

In conclusion, we have shown that pathogenic *POU3F3* variants cause a neurodevelopmental disorder with a broad phenotypic spectrum that includes ID and/or DD, speech and language problems, hypotonia, and autism spectrum disorder. Most individuals have mild to moderate delays in neurodevelopment, but a distinct phenotype of severe ID and epilepsy is also reported in two individuals with an identical missense variant. Although most variants result in loss-of-function effects on the transactivation capacities of POU3F3, other possible pathogenic mechanisms cannot be excluded. By showing the effects of POU3F3 dysfunction in humans, our data highlight the essential functions of POU3F3 for normal brain development.

Acknowledgements

We thank all individuals and families for their contribution.

This work was supported by the Netherlands Organization for Scientific Research (NWO) Gravitation Grant 24.001.006 to the Language in Interaction Consortium (L.S.B., S.E.F., and H.G.B.), the Netherlands Organization for Health Research and Development (ZonMw grant 91718310 to T.K.), the Max Planck Society (P.D. and S.E.F.), Fondazione Bambino Gesù (Vite Coraggiose) (M.T.), the Italian Ministry of Health (Ricerca Corrente 2019) (M.L.D.), the Japan Agency for Medical Research and Development (AMED) under grant numbers JP18ek0109280, JP18dm0107090, JP18ek0109301, JP18ek0109348, and JP18kk0205001 (N.Ma.), Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (JSPS KAKENHI) under grant numbers JP17H01539 (N.Ma.) and JP16H05357 (N.Mi.), grants by the Ministry of Health, Labour, and Welfare in Japan (N.Ma.), grants by the Takeda Science Foundation (N.Ma. and N.Mi.), and an Estonian Research Council grant PRG471 (K.Õ.). M.H.W. is supported by grant T32GM007748 (National Institutes of Health). The Broad Center for Mendelian Genomics (UM1 HG008900) is funded by the National Human Genome Research Institute with supplemental funding provided by the National Heart, Lung, and Blood Institute under the Trans-Omics for Precision Medicine (TOPMed) program and the National Eye Institute. Individuals 1 and 2 were part of the DDD study cohort. Acknowledgments of the DDD Study are included in the Supplemental Data.

Web Resources

CADD, https://cadd.gs.washington.edu
ClustalW, https://www.genome.jp/tools-bin/clustalw
Decipher, https://decipher.sanger.ac.uk
gnomAD, https://gnomad.broadinstitute.org
MetaDome, https://stuart.radboudumc.nl/metadome
cNLS Mapper, http://nls-mapper.iab.keio.ac.jp
OMIM, https://www.omim.org
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2
SIFT, https://sift.bii.a-star.edu.sg
UniProt, https://www.uniprot.org

References

1. Sumiyama, K.,Washio-Watanabe, K., Saitou, N., Hayakawa, T., and Ueda, S. (1996). Class III POU genes: Generation of homopolymeric amino acid repeats under GC pressure in mammals. *J. Mol. Evol.* **43**, 170–178.
2. He, X., Treacy, M.N., Simmons, D.M., Ingraham, H.A., Swanson, L.W., and Rosenfeld, M.G. (1989). Expression of a large family of POU-domain regulatory genes in mammalian brain development. *Nature* **340**, 35–41.
3. Hagino-Yamagishi, K., Saijoh, Y., Ikeda, M., Ichikawa, M.,Minamikawa-Tachino, R., and Hamada, H. (1997). Predominant expression of Brn-2 in the postmitotic neurons of the developing mouse neocortex. *Brain Res.* **752**, 261–268.
4. 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.
5. 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 Dev.* **16**, 1760–1765.
6. 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. *Cereb. Cortex* **23**, 2632–2643.
7. Castro, D.S., Skowronska-Krawczyk, D., Armant, O., Donaldson, I.J., Parras, C., Hunt, C., Critchley, J.A., Nguyen, L., Gossler, A., Go¨ttgens, B., et al. (2006). Proneural bHLH and Brn proteins coregulate a neurogenic program through cooperative binding to a conserved DNA motif. *Dev. Cell* **11**, 831–844.
8. Dheedene, A., Maes, M., Vergult, S., and Menten, B. (2014). A de novo POU3F3 deletion in a boy with intellectual disability and dysmorphic features. *Mol. Syndromol.* **5**, 32–35.
9. Kasher, P.R., Schertz, K.E., Thomas, M., Jackson, A., Annunziata, S., Ballesta-Martinez, M.J., Campeau, P.M., Clayton, P.E., Eaton, J.L., Granata, T., et al. (2016). Small 6q16.1 deletions encompassing POU3F2 cause susceptibility to obesity and variable developmental delay with intellectual disability. *Am. J. Hum. Genet.* **98**, 363–372.
10. Rhee, J.M., Gruber, C.A., Brodie, T.B., Trieu, M., and Turner, E.E. (1998). Highly cooperative homodimerization is a conserved property of neural POU proteins. *J. Biol. Chem.* **273**, 34196–34205.
11. Maricic, T., Gu¨nther, 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. *Mol. Biol. Evol.* **30**, 844–852.
12. 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.
13. 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.
14. Holbrook, J.A., Neu-Yilik, G., Hentze, M.W., and Kulozik, A.E. (2004). Nonsense-mediated decay approaches the clinic. Nat. Genet. 36, 801–808.
15. Chang, Y.F., Imam, J.S., and Wilkinson, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.* **76**, 51–74.
16. Li, P., He, X., Gerrero, M.R., Mok, M., Aggarwal, A., and Rosenfeld, M.G. (1993). Spacing and orientation of bipartite DNA-binding motifs as potential functional determinants for POU domain factors. *Genes Dev.* **7** (12B), 2483–2496.
17. Phillips, K., and Luisi, B. (2000). The virtuoso of versatility: POU proteins that flex to fit. *J. Mol. Biol.* **302**, 1023–1039.
18. Jauch, R., Choo, S.H., Ng, C.K., and Kolatkar, P.R. (2011). Crystal structure of the dimeric Oct6 (POU3f1) POU domain bound to palindromic MORE DNA. *Proteins* **79**, 674–677.
19. 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.
20. Smit, D.J., Smith, A.G., Parsons, P.G., Muscat, G.E., and Sturm, R.A. (2000). Domains of Brn-2 that mediate homodimerization and interaction with general and melanocytic transcription factors. *Eur. J. Biochem.* **267**, 6413–6422.
21. 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.* **87**, e51438.
22. Reme´nyi, A., Tomilin, A., Scho¨ler, H.R., and Wilmanns, M. (2002). Differential activity by DNA-induced quarternary structures of POU transcription factors. *Biochem. Pharmacol.* **64**, 979–984.

23. Mitchell, P.J., and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378.
24. Westphal, D.S., Riedhammer, K.M., Kovacs-Nagy, R., Meitinger, T., Hoefele, J., and Wagner, M. (2018). A de novo missense variant in POU3F2 identified in a child with global developmental delay. *Neuropediatrics* **49**, 401–404.
25. Deriziotis, P., and Fisher, S.E. (2017). Speech and language: Translating the genome. *Trends Genet.* **33**, 642–656.
26. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291.
27. Kosugi, S., Hasebe, M., Tomita, M., and Yanagawa, H. (2009). Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc. Natl. Acad. Sci. USA* **106**, 10171–10176.

Supplemental Table

Table S1 is available online via doi:10.1016/j.ajhg.2019.06.007

Supplemental Figures

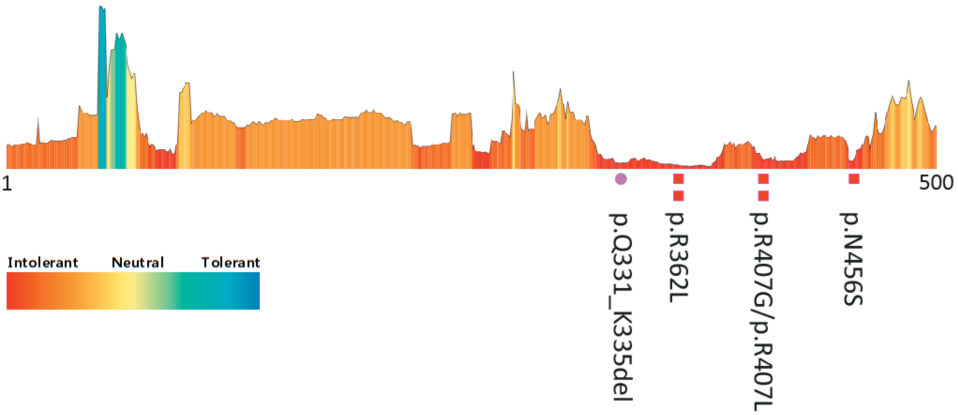
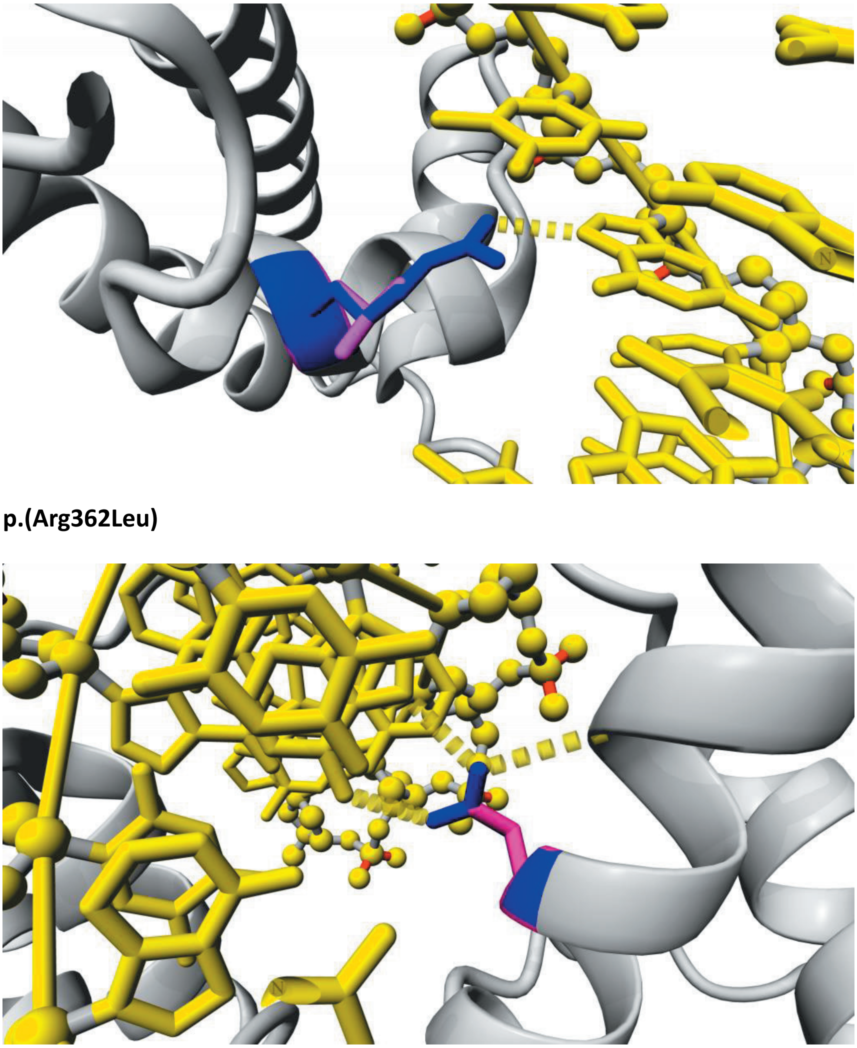


Figure S1: Non-truncating variants visualised in tolerance landscape of POU3F3

Tolerance landscape of the POU3F3 protein based on transcript NM_006236.2 (ENST00000361360.2) visualized via the MetaDome web server1. The tolerance landscape is computed based on single nucleotide variants present in the gnomAD database. It is calculated as a missense over synonymous ratio in a sliding window of 21 residues over the entire POU3F3 protein. The green and blue peaks correspond to regions more tolerant to missense variation, and the red valleys indicate intolerant regions. The locations of the non-truncating variants in our cohort are displayed within the tolerance landscape of POU3F3. All these variants are located in regions that are highly intolerant to missense variation.



p.(Arg362Leu)

p.(Asn456Ser)

Figure S2: Three-dimensional visualisation for the p.(Arg362Leu) and p.(Asn456Ser) variants

A detailed visualization of the three-dimensional modeling analysis for the two missense variants affecting amino acids that directly bind to the major groove of DNA . Wild-type residues are shown in blue, while substitutions (caused by variants) are shown in magenta. DNA is depicted in yellow.

a) Arg362 (blue) is able to form hydrogen bonds with guanine in the DNA binding site. This binding is disturbed by substitution into leucine (magenta).

b) Asn456 (blue) is able to form hydrogen bonds with adenine in the DNA binding site. This binding is disturbed by substitution into serine (magenta).

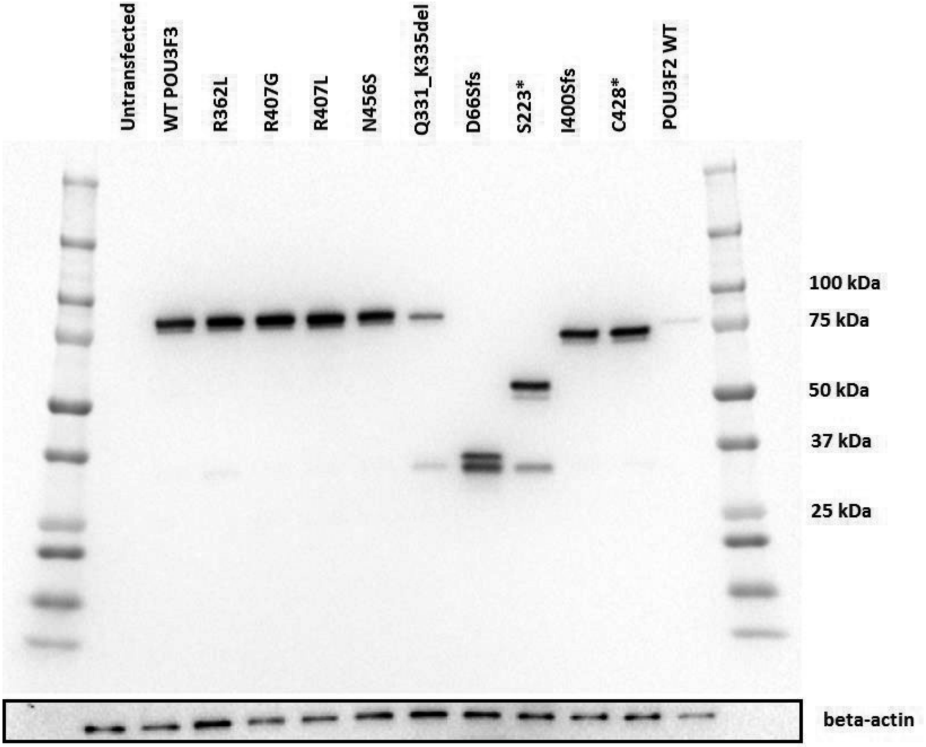


Figure S3: Immunoblot analysis

Western blot analysis of whole-cell lysates expressing eleven different YFP-tagged constructs, probed with an anti-EGFP antibody: wild-type POU3F3, nine different POU3F3 variants and wild-type POU3F2. The immunoblot was stripped and then re-probed with beta-actin as a loading control. All different expressed YFP-fusion proteins are visible at the expected molecular weights.

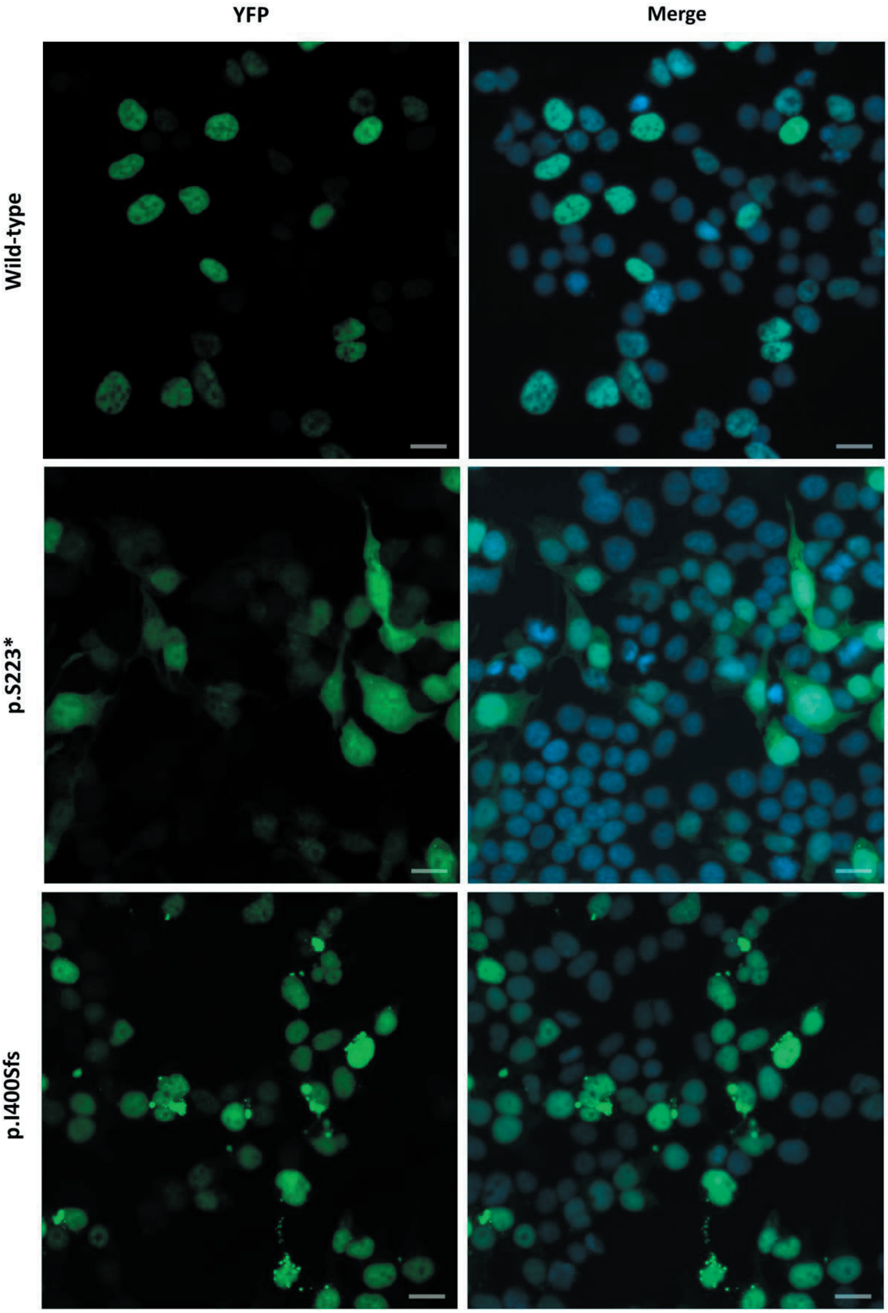
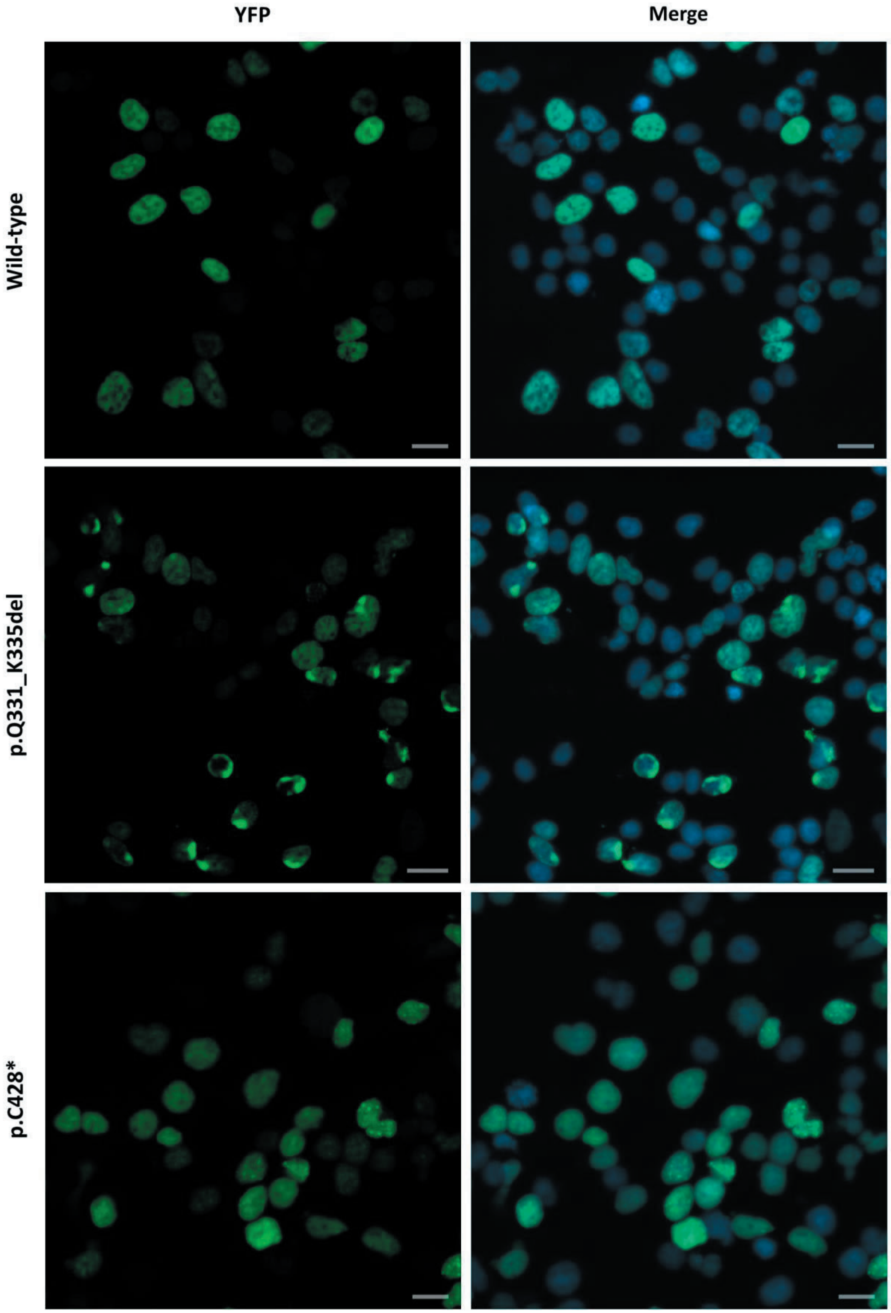
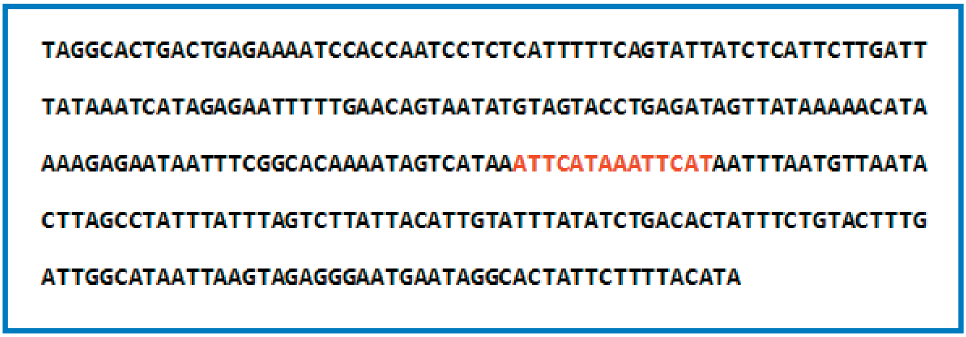


Figure S4: Aberrant subcellular localization patterns in a subset of cells for four variants

a) Direct fluorescence imaging of cells expressing YFP-tagged variants of the POU3F3 protein: wild-type POU3F3, and POU3F3 with the p.(Ser223*) variant and the p.(Ile400Serfs*16) variant. In addition to the cytoplasmic localization of the p.(Ser223*) variant, both variant conditions show perinuclear aggregates in a subset of cells. Nuclei are stained with DAPI (blue). Scale bar = 10µm.



b) Direct fluorescence imaging of cells expressing YFP-tagged variants of the POU3F3 protein: wild-type POU3F3, and POU3F3 with the p.(Gln331_Lys335del) variant and the p.(Cys428*) variant. Both variant conditions show an aberrant localization pattern within the nucleus in a subset of cells. Nuclei are stained with DAPI (blue). Scale bar = 10µm.



```
TAGGCACTGACTGAGAAAATCCACCAATCCTCTCATTTTTCAGTATTATCTCATTCTTGATT

TATAAATCATAGAGAATTTTTGAACAGTAATATGTAGTACCTGAGATAGTTATAAAAACATA

AAAGAGAATAATTTCGGCACAAAATAGTCATAAATTCATAAATTCATAATTTAATGTTAATA

CTTAGCCTATTTATTTAGTCTTATTACATTGTATTTATATCTGACACTATTTCTGTACTTTG

ATTGGCATAATTAAGTAGAGGGAATGAATAGGCACTATTCTTTTACATA
```

Figure S5: POU3F3 binding site in intronic region of *FOXP2*

This figure shows the ~300bp region of intron 8 of FOXP2 (chr7:114,289,482-114,289,778 (hg19/GRCh37)), that was cloned into a luciferase reporter vector to investigate transcriptional activation. The previously described POU3F2 consensus binding site2 is shown in red.

Supplemental Material and Methods

Research subjects

Informed consent was obtained from all participating families. For all pictures of probands in this study, specific consent to publish clinical photographs was obtained. All procedures in this study matched the local ethical guidelines of the participating centres, and are in accordance with the Declaration of Helsinki. Probands with possible pathogenic POU3F3 variants were found using the GeneMatcher website[3], the Decipher Database[4] and matchbox[5], and table S1 contains details on which specific matchmaking platform was used to identify each individual.

Exome sequencing, variant filtering and annotation

Exome sequencing and variant filtering were performed as previously described[6-1]7. In individuals 1-17, whole exome sequencing and variant filtering was performed using a trio approach, in which sequencing was performed in the proband and both parents. In individual 18 and 19, whole exome sequencing was performed with a duo approach, in which the proband (individual 18) and the mother (individual 19) were sequenced. In all individuals, the POU3F3 variant was considered to be the most likely variant contributing to the phenotype, and there were no additional pathogenic or likely pathogenic variants reported. Additional variants (SNVs and CNVs) considered to be possibly pathogenic and/or to possibly contribute to the phenotype, are listed in Table S1. All variants in this study are annotated with respect to the NM_006236.2 transcript.

Cell culture and transfection

Human embryonic kidney cells 293 (HEK293) cells were grown in Dulbecco's Modified Eagle's Medium (Gibco) that was supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin mix (both Gibco). The cells were cultured at 37°C in an incubator with 5%

CO₂. Transient transfection was performed with GeneJuice Transfection Reagent (Merck MilliPore) according to the manufacturer’s instructions.

Cloning of DNA constructs and site-directed mutagenesis

A synthetic clone of wildtype POU3F3 cDNA with flanking restriction sites EcoR1 and Xba1 in a pUC57-vector was synthesized by GenScript. The POU3F3 insert was subcloned using the EcoR1/Xba1 restriction sites into a modified pLuc and pYFP vector as previously described [18].

Variants in POU3F3 constructs were created using site-directed mutagenesis (SDM) with the QuikChange II Site-Directed Mutagenesis Kit (Agilent) according to the manufacterer’s instructions. The following mutated constructs were created (corresponding SDM primers in parentheses; F = Forward primer, R = Reverse primer): p.Arg362Leu (F: 5’-ACCACCATCTGCCTCTTCGAGGCCCTG-3’; R: ’-CAGGGCCTCGAAGAGGCAGATGGTGGT-3’), p.Arg407Gly (F: 5’-CAGGGCCGCAAGGGCAAGAAGCGGA- 3’; R: 5’-TCCGCTTCTTGCCCTTGCGGCCCTG-3’), p.Arg407Leu (F: 5’- CAGGGCCGCAAGCTCAAGAAGCGGACC-3’; R: 5’-GGTCCGCTTCTTGAGCTTGCGGCCCTG-3’), p.Asn456Ser (F: 5’-GCGGGTCTGGTTCTGCAGTCGGCGCCA-3’; R: 5’-TGGCGCCGACTGCAGAACCAGACCCGC-3’), p.Gln331_Lys335del (F: 5’-CCTGCGTGAAGCCCAGCTTGAACTGCTTGG-3’; R: 5’- CCAAGCAGTTCAAGCTGGGCTTCACGCAGG-3’), p.Asp66Serfs (F: 5’-GCCTACCGGGGGTCCCGTCCTCTGT- 3’; R: 5’-ACAGAGGACGGACCCCCGGTAGGC-3’), p.Ser223* (F: 5’- CCGGGCTGCTAGTAGAGCAGACTCTGC-3’; R: 5’-GCAGAGTCTGCTCTACTAGCAGCCCCG-3’), p. Ile400Serfs (F: 5’-GCGCCGCGATTTGTCGATGCTTGTGGGG-3’; R: 5’- CCCCACAAGCATCGACAAATCGCGGCGC-3’) and p.Cys428* (F: 5’- CACTTCCTCAAGTGACCCAAGCCCTCCGC-3’; R: 5’-GCGGAGGGCTTGGGTCACTTGAGGAAGTG-3’).

After SDM, variants were validated using Sanger sequencing, and POU3F3 inserts were subcloned into new pLuc and pYFP vectors. An ‘empty YFP-vector’ (modified pYFP expression vector without POU3F3 insert) was used as a control construct for the luciferase assays.

POU3F2 cDNA was cloned into TOPO vector using the following primers: 5’-GAGGATCCTGGCGACCGCAGCGTCTAACCAC-3’ (Forward primer) and
5’- GATCTAGATTACTGGACGGGCGTCTGCACCCCG-3’ (Reverse primer).
The POU3F2 insert was then subcloned into modified pLuc and pYFP vectors (as previously described; 18) using BamHI and XbaI restriction sites.

To create the firefly reporter construct for luciferase assays, the previously described POU3F2 binding site in FOXP2 (Figure S5) was cloned into TOPO from gDNA using the following primers: 5'- CTCGAGTAGGCACTGACTGAGAAAATC-3' (Forward primer) and 5'-AGATCTATATGTAAAAGAATAGTGCCT-3' (Reverse primer). The binding site was then subcloned into a p.GL4.23 vector (Promega) using the restriction sites BglII and XhoI. Control constructs used in the BRET assays (pYFP-vector with NLS-insert, and pRLuc-vector with NLSinsert) were made as previously described18. All constructs were validated by Sanger sequencing.

Subcellular localization

HEK293 cells were seeded on poly-D-lysine (Sigma) coated coverslips, and transfected after 24 hours. At 36 hours post-transfection, the cells were fixed by incubation in 4% Paraformaldehyde (Electron Microscopy Sciences) for ten minutes at room temperature. Coverslips were mounted onto slides using Vectashield mounting medium for fluorescence with DAPI (Vector). The proteins of interest were expressed as fusion proteins to YFP. Fluorescence images were obtained with an Axiovert A-1 fluorescence microscope and ZEN imaging software (Zeiss).

BRET assay

HEK293 cells were transfected 24 hours after plating in 96 well plates, with pairs of Renilla luciferase and YFP fusion proteins, as previously described[18]. Luciferase substrate (EnduRen; Promega) was added at 60µM 36 hours post-transfection. After four hours of incubation, the emission was measured using a TECAN F200 PRO or M200 PRO microplate reader using the Blue1 and Green1 filter sets. To determine the YFP-fusion protein expression level, fluorescence measurements were taken using a filter and a dichroic mirror suitable for green fluorescent protein (GFP) fluorescence (excitation 480nm, emission 535nm). The corrected BRET ratio was obtained using the following formula [$\text{Green1}_{\text{(experimental condition)}}/\text{Blue1}_{\text{(experimental condition)}}$] - [$\text{Green1}_{\text{(control condition)}}/\text{Blue1}_{\text{(control condition)}}$]. The BRET assay set-up that was used for this study is discussed in more detail in Deriziotis et al.[18]

Luciferase assay

HEK293 cells were transfected 24 hours after seeding in 96 well plates with the firefly luciferase reporter construct (2µl of 36nM), a pGL4.74 (*hRluc*/TK) *Renilla* Reniformis luciferase construct (2µl of 36nM; Promega) and a YFP-expression construct or empty YFP-expression vector (6µl of 36nM). Firefly luciferase and *Renilla* luciferase activities were measured 36 hours post-transfection using the Dual-Luciferase Reporter Assay System (Promega) and a TECAN F200 PRO microplate reader.

Western blotting

Whole cell lysates were collected 40 hours post-transfection by treatment with RIPA buffer (Cell Signaling Technology) supplemented with 0.1mM PMSF (Sigma), Protease Inhibitor Cocktail (Roche) and 1mM DTT (Sigma). Cells were lysed for 30 minutes at 4°C followed by centrifugation for 10 minutes at 13,000rpm at 4°C. Laemmli buffer (Bio-Rad) was added to the supernatants, and the proteins were loaded on 4-15% Mini Protean-TGX Precast Gels (Bio-Rad) and transferred onto polyvinylidene fluoride membranes (Bio-Rad). Membranes were blocked in 5% milk in PBS-T (Phospate Buffered Saline supplemented with 0.1% Tween) for 1.5h at room temperature and then probed with 1:8000 mouse anti-EGFP (Clontech) in 1% milk in PBS-T at 4°C overnight. This was followed by incubation with 1:3000 HRP-conjugated goat anti-mouse (Bio-Rad) in 1% milk at room temperature for 1h. Proteins were visualized with Novex ECL Chemluminescent Substrate Reagent Kit (Invitrogen), using the ChemiDoc XRS+ System (Bio-Rad). To check for equal loading of proteins, the blot was subsequently stripped for 25 minutes in Re-blot Plus Strong stripping solution (Millipore) at room temperature, and blocked in 5% milk in PBS-T for 30 minutes at room temperature, followed by incubation with 1:1000 mouse anti-beta-actin (Sigma) in 1% milk overnight at 4°C, and incubation with 1:3000 HRP-conjugated goat anti-mouse (Bio-Rad) in 1% milk at room temperature.

Three-dimensional modeling

The exact three-dimensional structure of human POU3F3 is not known. Therefore, we created a homology model based on the crystal structure of the mouse POU3F1 structure (PDB file 2XSD)[19]. The human POU3F3 and mouse POU3F1 sequences show 94% identity over 147 residues in the Cterminal domain, containing both the POU-specific and the POU-homeo domain. We used the YASARA & WHAT IF Twinset modeling algorithm with standard parameters for modeling and subsequent analysis[20, 21].

Supplemental Acknowledgements

Individuals 1 and 2 were part of the DDD study cohort. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003], a parallel funding partnership between Wellcome and the Department of Health, and the Wellcome Sanger Institute [grant number WT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of Wellcome or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. This study makes use of DECIPHER (http://decipher.sanger.ac.uk), which is funded by the Wellcome.

Supplemental References

1. 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*
2. 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. *Mol Biol Evol* **30**, 844-852.
3. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Human mutation* **36**, 928-930.
4. 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.
5. Arachchi, H., Wojcik, M.H., Weisburd, B., Jacobsen, J.O.B., Valkanas, E., Baxter, S., Byrne, A.B., O'Donnell-Luria, A.H., Haendel, M., Smedley, D., et al. (2018). matchbox: An open-source tool for patient matching via the Matchmaker Exchange. *Hum Mutat* **39**, 1827-1834.
6. Deciphering Developmental Disorders, S. (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-228.
7. Farwell, K.D., Shahmirzadi, L., El-Khechen, D., Powis, Z., Chao, E.C., Tippin Davis, B., Baxter, R.M., Zeng, W., Mroske, C., Parra, M.C., et al. (2015). Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. *Genet Med* **17**, 578-586.
8. Hempel, M., Cremer, K., Ockeloen, C.W., Lichtenbelt, K.D., Herkert, J.C., Denecke, J., Haack, T.B., Zink, A.M., Becker, J., Wohlleber, E., et al. (2015). De Novo Mutations in CHAMP1 Cause Intellectual Disability with Severe Speech Impairment. *Am J Hum Genet* **97**, 493-500.
9. Flex, E., Niceta, M., Cecchetti, S., Thiffault, I., Au, M.G., Capuano, A., Piermarini, E., Ivanova, A.A., Francis, J.W., Chillemi, G., et al. (2016). Biallelic Mutations in TBCD, Encoding the Tubulin Folding Cofactor D, Perturb Microtubule Dynamics and Cause Early-Onset Encephalopathy. *Am J Hum Genet* **99**, 962-973.
10. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291.
11. Lelieveld, S.H., Reijnders, M.R., Pfundt, R., Yntema, H.G., Kamsteeg, E.J., de Vries, P., 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. *Nature Neuroscience* **19**, 1194-1196.
12. Thevenon, J., Duffourd, Y., Masurel-Paulet, A., Lefebvre, M., Feillet, F., El Chehadeh-Djebbar, S., St-Onge, J., Steinmetz, A., Huet, F., Chouchane, M., et al. (2016). Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. *Clin Genet* **89**, 700-707.
13. Smith, E.D., Radtke, K., Rossi, M., Shinde, D.N., Darabi, S., El-Khechen, D., Powis, Z., Helbig, K., Waller, K., Grange, D.K., et al. (2017). Classification of Genes: Standardized Clinical Validity Assessment of Gene-Disease Associations Aids Diagnostic Exome Analysis and Reclassifications. *Hum Mutat* **38**, 600-608.
14. Bauer, C.K., Calligari, P., Radio, F.C., Caputo, V., Dentici, M.L., Falah, N., High, F., Pantaleoni, F., Barresi, S., Ciolfi, A., et al. (2018). Mutations in KCNK4 that Affect Gating Cause a Recognizable Neurodevelopmental Syndrome. *Am J Hum Genet* **103**, 621-630.
15. Bouman, A., Waisfisz, Q., Admiraal, J., van de Loo, M., van Rijn, R.R., Micha, D., Oostra, R.J., and Mathijssen, I.B. (2018). Homozygous DMRT2 variant associates with severe rib malformations in a newborn. *Am J Med Genet A* **176**, 1216-1221.
16. Gibson, K.M., Nesbitt, A., Cao, K., Yu, Z., Denenberg, E., DeChene, E., Guan, Q., Bhoj, E., Zhou, X., Zhang, B., et al. (2018). Novel findings with reassessment of exome data: implications for validation testing and interpretation of genomic data. *Genet Med* **20**, 329-336.
17. Suzuki, T., Behnam, M., Ronasian, F., Salehi, M., Shiina, M., Koshimizu, E., Fujita, A., Sekiguchi, F., Miyatake, S., Mizuguchi, T., et al. (2018). A homozygous NOP14 variant is likely to cause recurrent pregnancy loss. *J Hum Genet* **63**, 425-430.
18. 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.* **87**:e51438.
19. Jauch, R., Choo, S.H., Ng, C.K., and Kolatkar, P.R. (2011). Crystal structure of the dimeric Oct6 (POU3f1) POU domain bound to palindromic MORE DNA. *Proteins* **79**, 674-677.
20. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J Mol Graph* **8**, 52-56, 29.

21. 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.

4

Chapter 4

CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language

Lot Snijders Blok[1], Justine Rousseau[1], Joanna Twist[1], Sophie Ehresmann, Motoki Takaku, Hanka Venselaar, Lance H. Rodan, Catherine B. Nowak, Jessica Douglas, Kathryn J. Swoboda, Marcie A. Steeves, Inderneel Sahai, Connie T.R.M. Stumpel, Alexander P.A. Stegmann, Patricia Wheeler, Marcia Willing, Elise Fiala, Aaina Kochhar, William T. Gibson, Ana S.A. Cohen, Ruky Agbahovbe, A. Micheil Innes, P.Y. Billie Au, Julia Rankin, Ilse J. Anderson, Steven A. Skinner, Raymond J. Louie, Hannah E. Warren, Alexandra Afenjar, Boris Keren, Caroline Nava, Julien Buratti, Arnaud Isapof, Diana Rodriguez, Raymond Lewandowski, Jennifer Propst, Ton van Essen, Murim Choi, Sangmoon Lee, Jong H. Chae, Susan Price, Rhonda E. Schnur, Ganka Douglas, Ingrid M. Wentzensen, Christiane Zweier, André Reis, Martin G. Bialer, Christine Moore, Marije Koopmans, Eva H. Brilstra, Glen R. Monroe, Koen L.I. van Gassen, Ellen van Binsbergen, Ruth Newbury-Ecob, Lucy Bownass, Ingrid Bader, Johannes A. Mayr, Saskia B. Wortmann, Kathy J. Jakielski, Edythe A. Strand, Katja Kloth, Tatjana Bierhals, The DDD study, John D. Roberts, Robert M. Petrovich, Shinichi Machida, Hitoshi Kurumizaka, Stefan Lelieveld, Rolph Pfundt, Sandra Jansen, Pelagia Deriziotis, Laurence Faivre, Julien Thevenon, Mirna Assoum, Lawrence Shriberg, Tjitske Kleefstra, Han G. Brunner, Paul A. Wade, Simon E. Fisher[2] & Philippe M. Campeau[2]

[1] These authors contributed equally
[2] These authors jointly supervised this work

Nature Communications 2018
doi:10.1038/s41467-018-06014-6

Abstract

Chromatin remodeling is of crucial importance during brain development. Pathogenic alterations of several chromatin remodeling ATPases have been implicated in neurodevelopmental disorders. We describe an index case with a de novo missense mutation in *CHD3*, identified during whole genome sequencing of a cohort of children with rare speech disorders. To gain a comprehensive view of features associated with disruption of this gene, we use a genotype-driven approach, collecting and characterizing 35 individuals with de novo *CHD3* mutations and overlapping phenotypes. Most mutations cluster within the ATPase/helicase domain of the encoded protein. Modeling their impact on the three-dimensional structure demonstrates disturbance of critical binding and interaction motifs. Experimental assays with six of the identified mutations show that a subset directly affects ATPase activity, and all but one yield alterations in chromatin remodeling. We implicate de novo *CHD3* mutations in a syndrome characterized by intellectual disability, macrocephaly, and impaired speech and language.

The Chromodomain Helicase DNA-binding (CHD) protein family is a key class of ATP-dependent chromatin remodeling proteins, which utilize energy derived from ATP hydrolysis to regulate chromatin structure, thereby modulating gene expression[1,2]. CHD proteins are crucial for developmental processes[1,3], with various members implicated in major neurodevelopmental disorders including CHD2 in epileptic encephalopathy[4], CHD7 in CHARGE syndrome[5], CHD8 in autism[6,7], and more recently CHD4 and CHD1 in neurodevelopmental syndromes[8,9]. Three CHD proteins (CHD3, CHD4, and CHD5) can exert their chromatin remodeling activity by forming the core ATPase subunit of the NuRD complex[1,10–12]. The NuRD complex is associated with various fundamental cellular mechanisms, including genomic integrity and cell cycle progression[13], and plays important roles in embryonic stem cell differentiation[14]. A recent study reports that the different CHD factors within the NuRD complex (CHD3, CHD4, and CHD5) are developmentally regulated in the mouse brain, each having distinct and mostly non-redundant functions during cortical development[15]. In particular, the CHD3 protein has been implicated in late neural radial migration and cortical layer specification.

In contrast to most other members of the CHD protein family, a specific syndrome associated with mutations in *CHD3* (MIM 602120) has not yet been characterized. In this study, based on an index case from whole genome sequencing of children with rare speech disorders, we assemble a set of 35 probands carrying de novo mutations that disrupt *CHD3*. We characterize the overlapping phenotypic features of probands with *CHD3* mutations, including intellectual disability (with a wide range of severity), developmental delays, macrocephaly, impaired speech and language skills, and characteristic facial features. We identify mainly missense mutations that cluster in and around the ATPase/helicase domain of the CHD3 protein, and are predicted to disturb function, based on three-dimensional modeling. We use functional assays to describe the effects of multiple different CHD3 mutations on ATPase activity and chromatin remodeling capacities. Taken together, our data demonstrate that de novo missense mutations in *CHD3* disturb chromatin remodeling activities of the encoded protein, thereby causing a neurodevelopmental disorder.

Results

De novo CHD3 mutations cause a neurodevelopmental phenotype.

During whole genome sequencing of a cohort of 19 unrelated children with a primary diagnosis of Childhood Apraxia of Speech (CAS)[16], we discovered a de novo missense mutation in *CHD3*, predicted to disrupt the helicase domain of the encoded protein. CAS is a rare neurodevelopmental disorder characterized by impairments in learning to produce the coordinated sequences of mouth and face movements underlying fluent speech. Remarkably, the CHD3 protein is one of the few documented interaction partners of FOXP2 (see Supplementary Table S1 in ref. [17]), a transcription factor that has been implicated in monogenic forms of CAS, accompanied by wide-ranging language problems, in multiple families and unrelated cases[18–20].

Discovery of the *CHD3* mutation (NM_001005273.2, p.Arg1169Trp) in our index case motivated a search for other de novo mutations in this gene. Studies of large numbers of simplex families with an autistic proband have documented just two single non-synonymous de novo variants in *CHD3* in probands[21,22], while eight additional non-synonymous variants were recently recorded in a study of thousands of children with unexplained developmental disorders from the UK[23], with limited information on phenotypic profiles of carriers of CHD3 variants. Via GeneMatcher[24] we assemble a cohort of 35 independently diagnosed probands with de novo mutations disrupting *CHD3*, to systematically assess the phenotypic consequences of damage to this gene.

The 35 probands with de novo mutations in *CHD3* show overlapping phenotypes, summarized in Table 1 and in more detail in Supplementary Data 1. All individuals have global developmental delays and/or intellectual disability, with a total IQ varying from 70–85 (borderline intellectual functioning) to below 35 (severe intellectual disability). Nine individuals (29%) show autism or autism-like features, including stereotypic and handflapping behavior. Interestingly, the majority of individuals (19 individuals; 58%) have macrocephaly, and in cases where neuroimaging has been performed, widening of cerebrospinal fluid spaces is noted in 10 out of 30 MRI reports (33%). One individual (individual 5) has microcephaly. Hypotonia is reported in 21 individuals (75%). The facial phenotype consists of widely spaced eyes, a broad and bossing forehead, periorbital fullness and narrow palpebral fissures, laterally sparse eyebrows, low-set and often simple ears with thick helices, and a pointed chin (Fig. 1). Joint dislocations and/or hyperlaxity are reported in 12 cases, and five individuals have inguinal or umbilical hernias. Five of the 21 male individuals have undescended testes. Vision problems are quite common and include hypermetropia (11 individuals), strabismus (10 individuals), and cerebral visual impairment (three individuals). One individual (individual 34) developed epilepsy, two additional individuals had neonatal convulsions. In many individuals an abnormal and often unsteady gait is reported, and one individual (individual 13) developed symptoms of Parkinsonism at a later age.

Given that our index case was ascertained on the basis of a formal diagnosis of CAS, we pay special attention to the association of *CHD3* mutations with speech and language deficits. The index case was diagnosed with severe speech apraxia at the age of 3 years, and then used sign language to communicate effectively. He has severe problems with expressive speech, against relatively normal scores on language comprehension tests and a composite IQ (KBIT) of 72. In all 33 subjects that were at least 2 years old at the last evaluation, *CHD3* disruptions are associated with delayed milestones in the speech and language domain. The average age for first spoken words in this cohort is 2 years and 10 months (range: 1.5–5.5 years, after excluding six individuals that were non-verbal at the last evaluation). Our data suggest that expressive language is more affected than receptive language, and intelligibility is often impaired. Speech-related problems identified in our cohort include dysarthria, speech apraxia, oromotor problems, and stuttering.

Table 1: Summary of phenotypes found in this cohort of probands with CHD3 mutations

	Amount	Percentage
Development		
ID/DD	35/35	100%
Degree of ID/DD		
Borderline ID	3/35	9%
Mild or mild–moderate ID	9/35	26%
Moderate or moderate–severe ID	8/35	23%
Severe ID	7/35	20%
DD/level unknown	8/35	23%
Speech delay/disorder	33/33	100%
Autism or autism-like features	9/31	29%
Neurology		
Hypotonia	21/28	75%
Macrocephaly	19/33	58%
Widened CSF spaces (MRI)	10/30	33%
Neonatal feeding problems	10/32	31%
Dysmorphisms		
High, broad, and/or prominent forehead	28/33	85%
Widely spaced eyes	24/31	77%
Other		
Joint laxity (generalized and/or local)	12/30	40%
Vision problems		
Hypermetropia	11/29	38%
Strabism	10/33	30%
Cerebral visual impairment	3/33	9%
Genital abnormalities in males	6/17	35%
Hernia (inguinal, umbilical, hiatal)	5/28	18%

More extensive clinical information per individual is provided in Supplementary Data 1. As information on the different features was not always applicable or known for each patient, the denominator in the "Amount" column is different for different clinical characteristics.

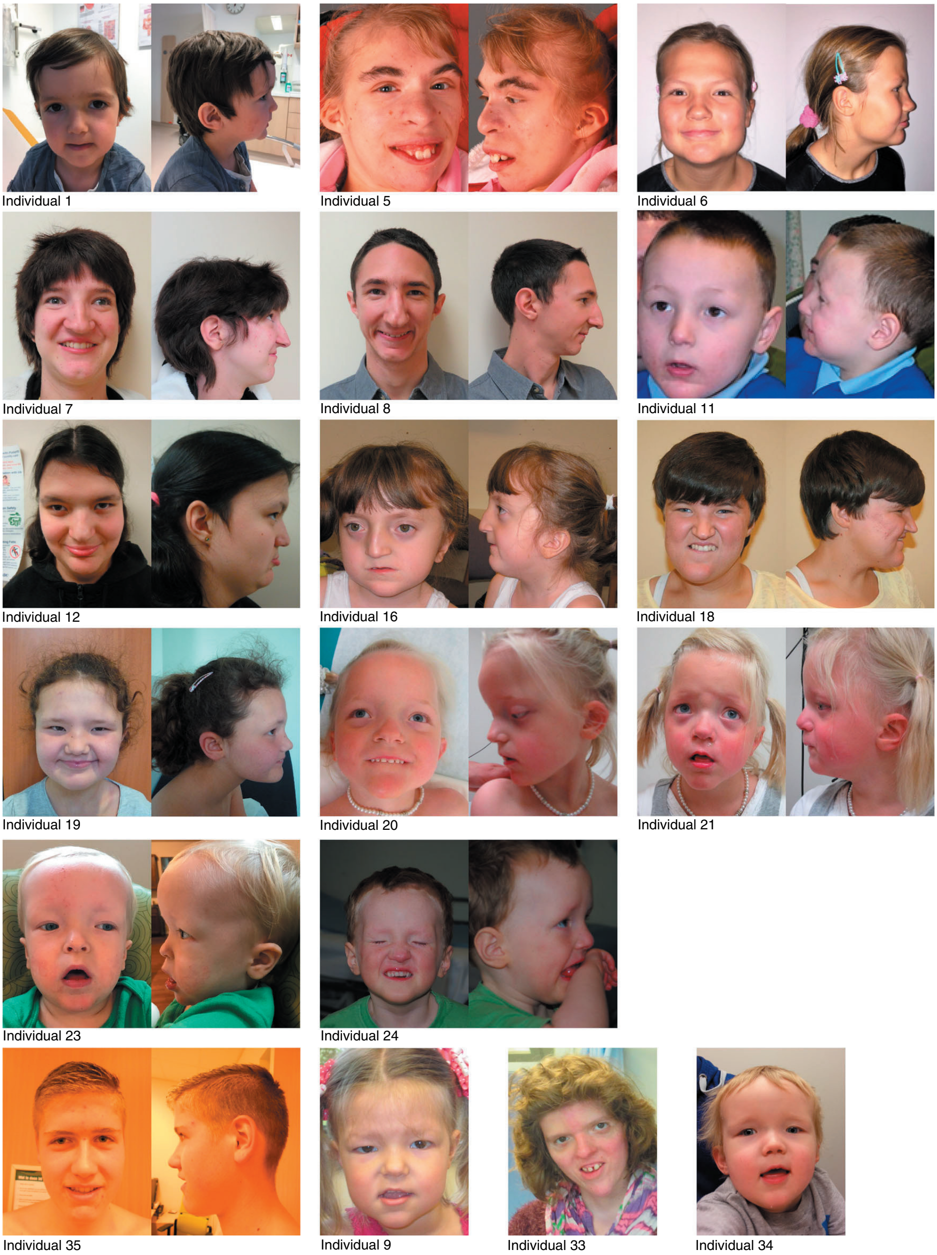


Figure 1: Photographs of affected individuals

Facial photographs showing dysmorphisms in 18 individuals with de novo CHD3 mutations. The majority of individuals have macrocephaly with a prominent or bossing forehead, individual 5 has microcephaly. Hypertelorism or telecanthus is common, often accompanied by narrow palpebral fissures, deep-set eyes, peri-orbital fullness, and/or epicanthal folds. The combination of macrocephaly and deep-set eyes leads to a more prominent supra-orbital ridge. Some individuals show midface hypoplasia. Many individuals have low-set ears that can be posteriorly rotated, and sometimes simple with thick helices. A broad nasal base, prominent nose, a bifid nasal tip, and characteristic pointy chin is also frequently seen, as well as laterally sparse eyebrows.

De novo CHD3 mutations cluster in the helicase domain

The 35 unrelated probands have 23 different de novo mutations in *CHD3* (Fig. 2a, b). None of these mutations are present in the GnomAD database (http://gnomad.broadinstitute.com). Except for four individuals, all individuals have missense mutations. Interestingly, within our cohort there are multiple cases of recurrent identical de novo mutations, revealing mutational hotspots. The most striking is p.Arg985Trp, found in six children from five different families, while two additional individuals have a different substitution affecting the same residue (p.Arg985Gln).

The *CHD3* protein is characterized by a SNF2-like ATPase/ helicase domain, together with two plant homeodomain (PHD) fingers and two chromodomains (Fig. 2b, c)[1,11], which mediate chromatin interactions and nucleosome remodeling[1]. The overwhelming majority of missense mutations (17/19) cluster within and around the ATPase/helicase motif, a functional domain that consists of two subdomains: a Helicase ATP-binding lobe and a Helicase-C-terminal lobe. This domain provides energy for nucleosome remodeling through its ATPase activity. All missense mutations affect amino acids that are highly conserved, both in different species and also in the other CHD proteins that can be part of the NuRD complex (Fig. S1), and clearly cluster in and around highly conserved SF2-family helicase motifs (Fig. S2). All are predicted to be pathogenic by Polyphen-2 and/or SIFT, and have CADD scores above 24 (Supplementary Data 1).

The identified de novo mutations also include one in-frame deletion of one amino acid (p.Gly1109del) and two truncating mutations (p.Glu457* and p.Phe1935Glufs*108), although the latter causes a frameshift at the very end of the protein, leading to a stop codon after 108 amino acids. RNA sequencing of transcripts with and without cycloheximide showed that this mutation escapes nonsense-mediated decay (Fig. S2). Finally, one case has a splice-site mutation (c.4073-2A>G) which is predicted to yield skipping of exon 27, while preserving the reading frame (Fig. 2a). Data from the ExAC database (http://exac.broadinstitute.com) indicate that CHD3 is extremely intolerant for loss-of-function mutations (loss-of-function intolerance score of 1.0) and highly intolerant for missense mutations (*Z*-score of +7.15)[25], supporting the pathogenicity of the mutations that we found.

All *CHD3* mutations were determined to be the most likely causal variant contributing to the disorder of the proband. In proband 15 who has a de novo *CHD3* p.Asp1120His mutation, a de novo truncating mutation in *CIC* was also identified (NM_015125.3:c.1444G>T; p.Glu482*). Since truncating mutations in *CIC* were recently suggested as a potential cause of intellectual disability (ID)[26], both mutations might be involved in the phenotype of this proband.

A subset of CHD3 mutations directly affects ATP hydrolysis

The striking clustering of almost all missense mutations in the ATPase/helicase domain of the CHD3 protein led us to hypothesize that disturbance of ATPase and/or chromatin remodeling activities of CHD3 could be potential pathogenic mechanisms. Three-dimensional modeling and mutation analysis of all missense mutations, including analysis of the conserved SF2-characteristic helicase motifs, demonstrates clear clustering of mutations and disturbance of important binding and interaction domains (Fig. 2d and Supplementary Note 1). Direct fluorescence imaging of mCherry-tagged CHD3 mutations in cellular models revealed no differences in subcellular localization for the mutated proteins as compared to wild-type CHD3 (Fig. S3).

We experimentally assessed ATPase activity of six representative mutations, selected to include one mutation in the Helicase ATP-binding lobe and several mutations in the Helicase Cterminal lobe. FLAG-tagged full-length wild-type CHD3 protein and each of the six mutant proteins were transiently expressed in mammalian HEK293 cells and purified (Fig. S4). Radiometric ATPase assays were performed to assess the activity of these mutant proteins relative to wild-type, in the presence of dsDNA (Fig. 3), recombinant nucleosomes (Fig. 3), or in the absence of DNA substrates as a control (Fig. S5). ATPase activities of p.Arg1121Pro and p.Arg1172Gln were significantly lower than wild-type for both substrate conditions. These findings are consistent with the modeling data, since p. Arg1121Pro is predicted to disrupt a helix integral to motif V, while p.Arg1172Gln is located in helicase motif VI, and both motifs are known to be critical in ATP hydrolysis. The activity of p.Asn1159Lys was significantly lower only in the presence of dsDNA, although the reason for the different activity depending on the substrate is currently unknown. The protein with the p. Leu915Phe mutation, located in conserved SNF2-motif III, is significantly hyperactive under both conditions. The p.Arg1187-Pro and p.Trp1158Arg mutations do not show statistically significant differences from the wild-type protein in these ATPase assays. According to the three-dimensional structure, the location of p.Arg1187Pro is not close to the ATP-binding or interaction surface. To assess whether mutant protein could impact activity of wild-type enzyme, we mixed wild-type protein with equimolar amounts of several mutant proteins, finding no biochemical evidence in this assay for interference (Fig. S6).

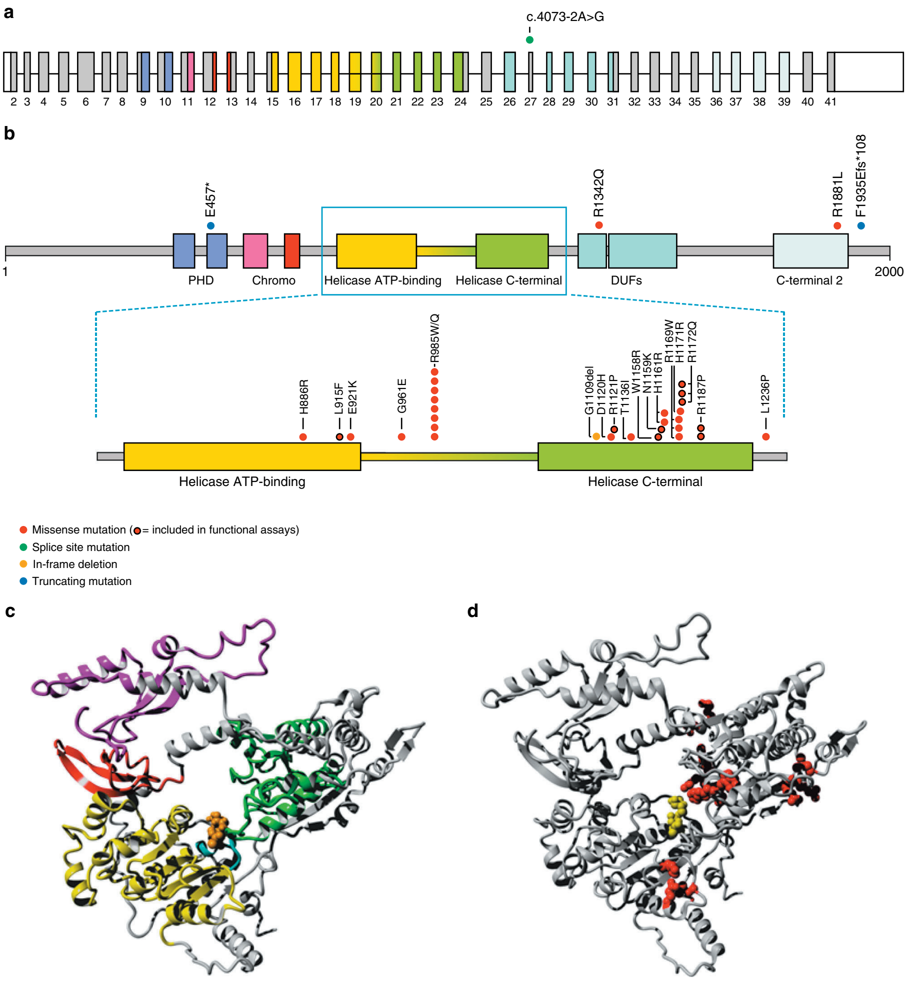


Figure 2: Schematic view of CHD3 transcript and protein with de novo mutations

a Schematic view of CHD3 exons (transcript 1, NM_001005273.2) with the splice site mutation c.4073-2A>G shown that most likely leads to skipping of exon 27 (22 amino acids), while preserving the reading frame. Exon 27 is part of the beginning of the second DUF domain (DUF 1086). Colors of the domains in **a** match with colors of domains in **b** and **c**. Five different types of domains are specified: plant homeodomains (PHD), chromodomains (Chromo), a Helicase domain consisting of two parts (Helicase ATP-binding and Helicase C-terminal), domains of unknown function (DUF), and a C-terminal 2 domain. **b** Schematic view of linear CHD3 protein (transcript 1, NM_001005273.2) with all mutations, except for the splice site mutation that is shown in **a**, found in our cohort. Almost all missense mutations cluster in or around the Helicase domain of the CHD3 protein. **c** Overview of one of the two CHD3-models used in this study, based on the 3MWY protein structure. This figure shows the different domains of the protein in their three-dimensional conformation: chromo domain 1 494–595 (magenta), chromo domain 2 631–673 (red), helicase ATP binding domain (yellow), helicase C-terminal domain (green), ATP binding residues 761–768 (cyan). ATP is orange, and gray residues do not belong to an indicated domain. Colors of the domains in **c** match with colors of domains in **a** and **b**. **d** The same structure as **c**, but in this figure the positions of the mutated residues are indicated in red, the sidechains of these residues are shown as red balls. The ATP molecule is shown in yellow. This figure illustrates the clustering of mutations on specific sites within the Helicase ATP-binding domain and Helicase C-terminal domain. A more detailed analysis of the different missense mutations in our cohort can be found in Supplementary Note 1.

CHD3 mutations disturb chromatin remodeling capacities

We measured the effects of six mutations on the chromatin remodeling activity of CHD3, by assessing restriction enzyme accessibility to nucleosomal DNA[27]. Consistent with its reduced activity in the ATPase assays, the p.Arg1172Gln mutant was partially, but not fully, active at chromatin remodeling (Fig. 4). p. Arg1121Pro, which showed severely reduced ATPase activity, was highly compromised in the chromatin remodeling assay. Moreover, p.Leu915Phe demonstrated hyperactivity in this assay, mirroring its elevated ATPase activity. Crucially, chromatin remodeling assays can also detect functional defects beyond ATP hydrolysis[27]. Two of the mutant proteins, p.Trp1158Arg and p. Asn1159Lys, exhibited severely compromised ability to remodel chromatin (Fig. 4) against a background of some preserved ATPase activity (c.f. Fig. 3). In sum, with the sole exception of p. Arg1187Pro, all the mutant versions of CHD3 that we tested differ from wild-type protein in their ability to remodel chromatin, with some mutants exhibiting decreased activity while one shows increased activity.

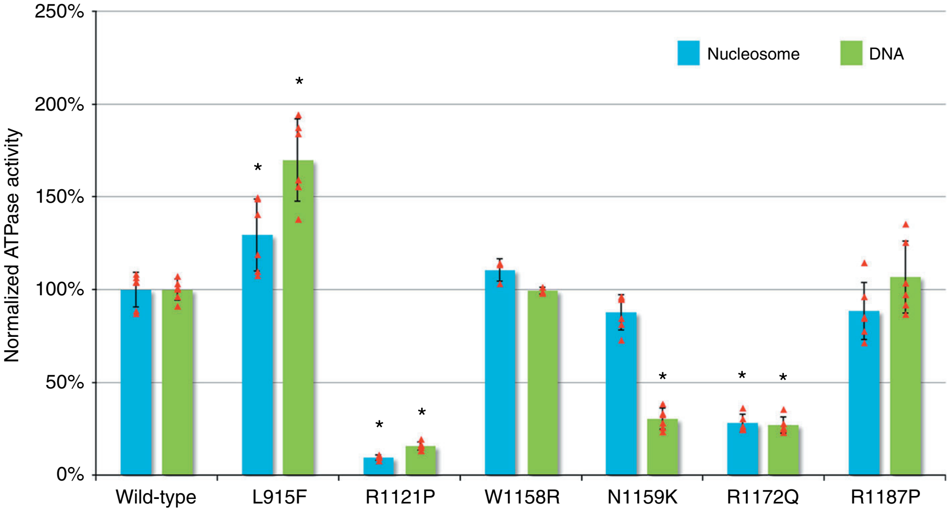


Figure 3: ATPase assays

Radiometric ATPase assays were performed to assess the activity of the mutant proteins relative to wild-type, in the presence of recombinant nucleosomes (blue), dsDNA (green), or in the absence of DNA substrates as a control (Fig. S4). Released phosphate was separated from unhydrolyzed ATP by thin layer chromatography, and detected by exposure to a phosphorimager. The experimental values (percentage hydrolyzed ATP) for the different mutant conditions were normalized to values for the wild-type condition within the experiment, to derive a normalized ATPase activity. The experimental data are presented as means ± standard deviation, individual data points are shown as red triangles. Three independent experiments from two individual purifications (wild-type, p.Leu915Phe, p.Arg1121Pro, p.Asn1159Lys, p.Arg1172Gln, and p.Arg1187Pro) (*N*= 6) or one purification (p.Trp1158Arg) (*N* = 3) were performed. Raw values from the individual experiments can be found in Supplementary Data 2. Asterisk (*) indicates significant difference for mutant values compared to wild-type values (unpaired *t*-test, $P < 0.05$) within the same substrate condition.

Discussion

In this study, we show that de novo *CHD3* mutations cause a neurodevelopmental disorder. We demonstrate defining clinical features of this syndrome. The characteristic phenotype of individuals with *CHD3* mutations overlaps with that reported for de novo mutations in *CHD4*, in which intellectual disability, macrocephaly, ventriculomegaly, undescended testes, and similar facial features have been reported. However, comparisons to the CHD4-related syndrome are currently limited because so far only five individuals with *CHD4* mutations have been clinically characterized. Also interesting in this context is the fact that four of the six recently described patients with missense mutations in *CHD1* have a diagnosis of speech apraxia[9], a relatively rare condition. Although CHD1 does not function in the same protein complex as CHD3 and has different expression patterns[9], there might be shared pathogenic mechanisms leading to speech problems in patients with mutations in these chromatin remodelers.

Based on the molecular and phenotypic data of individuals in our cohort, there is no obvious correlation between the precise type or location of the mutation, and the severity of the variable features of the resulting syndrome. However, the only individual in our cohort with epilepsy is also the only case with a missense mutation in the C-terminal domain of the protein. Future identification of more individuals with missense mutations in this region of the protein will help resolve whether this reflects a phenotype–genotype correlation.

In addition to defining the phenotype associated with *CHD3* mutations, we aimed to characterize the effects of *CHD3* mutations at a molecular and functional level. ATPase assays with six different mutant CHD3 constructs showed a clearly decreased ATPase activity for two mutations (p.Arg1121Pro and p.Arg1172Gln) and increased ATPase activity for one mutation (p.Leu915Phe). The disturbed ATPase activities are associated with corresponding effects on chromatin remodeling capacities for these three mutants, as shown by the restriction enzyme accessibility assays. It is currently unclear how deactivating and activating mutations can both yield similarly disruptive effects on neurodevelopmental outcomes. However, a recent study of cancer-specific mutations in the chromatin remodeling ATPase SMARCA4 concluded that mutations in the ATPase core of this enzyme had dominant-negative impacts on the global chromatin landscape regardless of whether they displayed increased or decreased dynamic recovery in fluorescence after photobleaching[28]. By analogy, it seems plausible that perturbed chromatin remodeling activity of CHD3, whether by gain or loss of activity relative to wild-type or by affecting its interactions, might likewise alter chromatin landscapes, to contribute to a neurodevelopmental phenotype.

Two mutations (p.Trp1158Arg and p.Asn1159Lys) show severely decreased chromatin remodeling capacities, despite unaffected ATPase activity in the presence of recombinant nucleosomes. In line with these findings, the highly conserved tryptophan residue at a position analogous to CHD3 residue 1158 has recently been shown to be critical for chromatin remodeling, but not for ATP hydrolysis, in the context of yeast SNF2[27]. Interestingly, the mutation in our cohort affecting this amino acid (p.Trp1158Arg) directly matches the position of a previously published mutation in CHD4[8] (Fig. S1), while the other previously published de novo missense mutations in CHD4-related syndrome are also mainly affecting the orthologous Helicase domain of CHD4 (Fig. S1)[8,29].

To systematically assess whether the distribution of the missense mutations in CHD3 reflects mutational hotspots in the gene, we performed a formal clustering analysis based on mutual distances, as previously described[30]. This analysis revealed significant clustering within the transcript (P = 0.0017), a finding that argues against simple haploinsufficiency as an underlying molecular mechanism. The paucity of patients with truncating mutations compared to the 31 patients with missense mutations in our cohort also supports this view, although the precise mechanistic effects of CHD3 mutations during neurodevelopment are a topic for future study.

Taken together, with our research we identify a recognizable neurodevelopmental disorder. We define the phenotypic spectrum associated with mutations in *CHD3*, and show the effects of several different mutations on ATPase activity and chromatin remodeling capacities. Our findings highlight the importance of chromatin remodeling factors, and specifically the CHD3 protein, in human brain development.

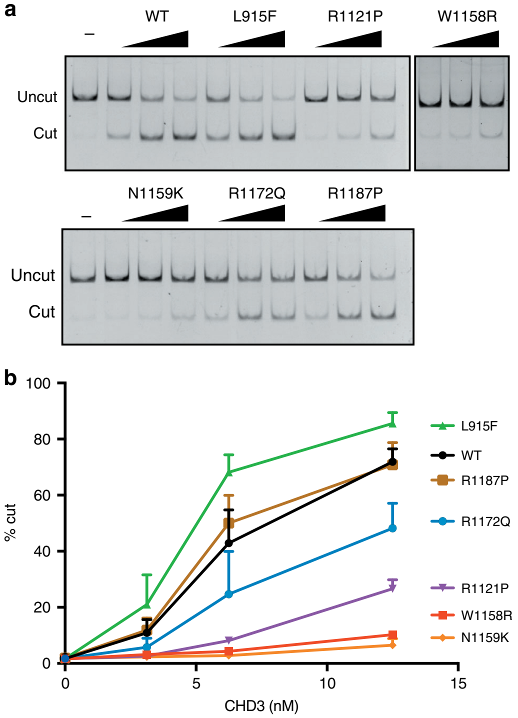


Figure 4: Restriction enzyme accessibility assay

a Restriction enzyme accessibility analysis of CHD3 wild-type and mutant proteins. 3.125, 6.25, or 12.5 nM of CHD3 proteins were incubated with 347 bp mononucleosomes. Digested fragments were analyzed by native polyacrylamide gel. **b** Quantitative analysis of restriction enzyme accessibility. Three individual experiments from two individual purifications (wild-type, p. Leu915Phe, p.Arg1121Pro, p.Asn1159Lys, p.Arg1172Gln, and p.Arg1187Pro) (N = 6) or one purification (p.Trp1158Arg) (N =3) were conducted. The experimental data are presented as means with standard deviations.

Methods

Individuals and consents

The authors affirm that (the legal representatives of) all human research participants provided informed consent for publication of the images in Fig. 1. Informed consent was also derived for the use of biological materials from all individuals or their legal representative. Genetic testing and research were performed in accordance with protocols approved by the local Institutional Review Boards where the patients were followed. Specifically, research exomes were performed after informed consent on protocols approved by the Institutional Review Boards of the following institutions: University of British Columbia, Augustana College, CHU Dijon, Mass General Hospital for Children, University of Erlangen-Nuremberg, Hamburg Chamber of Physicians, Cambridge South—UK Research Ethics Committee, University of Wisconsin-Madison Social & Behavioral Sciences.

Annotation of mutations

All mutations in this report are annotated in GRCh37 (hg19) and CHD3 transcript variant 1 (NM_001005273.2).

Next-generation sequencing

For the index case (individual 22), whole genome sequencing was performed using Illumina's HiSeq X Ten technology, the Burrows–Wheeler Aligner (BWA) software version 0.7.8-r455[31] and GATK v.3.4[32]. In other individuals, exome or genome sequencing and data analysis were performed as previously described[33–44].

Expression and purification of FLAG-CHD3

CHD3 proteins were prepared as previously described[45], with the following modifications. FLAG-CHD3 constructs were cloned into expression vectors (kindly provided by Guang Hu) using Gateway Cloning technology. Primer sequences are provided in Fig. S1. HEK293-f (ThermoFisher, FreeStyle™ 293-F Cells) were grown in suspension culture using FreeStyle™ 293 Expression Medium (ThermoFisher) in optimum growth flasks (Thomson) using a shaking incubator set at 8% CO_2, 80% humidity, and 150 rpm shaker rate. The cell count was 10^6 cells/ml on the day of transfection. Cells were transfected with 1 mg of expression vector using PEI max (Polysciences). Cells were harvested 48 h after transfection by centrifugation at 400 × g for 6 min. Cells were washed once with phosphate buffer saline solution prior to storage at −80 °C or protein purification.

The cell pellet was resuspended in lysis buffer (20mM HEPES, 1.5mM $MgCl_2$, 10mM KCl, 1mM DTT, 1mM PMSF, and 1× cOmplete® protease-inhibitor EDTA-free (Roche), pH 7.6). Cells were incubated on ice for 30 min, vortexed briefly, and nuclei were collected by centrifugation (5 min, 3300 × g, 4 °C). The supernatant was discarded and the nuclear pellet was resuspended in nuclear extraction buffer (20mM HEPES, 0.5M KCl, 1.5mM $MgCl_2$, 0.2mM EDTA, 20% glycerol, 0.2% NP-40, 1mM DTT, 1 mM PMSF, and 1× cOmplete®

proteaseinhibitor EDTA-free (Roche), pH 7.6). The nuclear pellet was homogenized using a Dounce homogenizer, incubated on ice for 30 min, and insoluble material was removed by centrifugation (20 min, 110,000 × g, 4 °C). The supernatant (nuclear extract) was incubated with α-FLAG M2 affinity gel (Sigma-Aldrich) and rotated overnight at 4 °C. The α-FLAG beads were then washed twice with nuclear extraction buffer, followed by 2 additional washes with wash buffer (20mM HEPES, 0.1M KCl, 0.2% NP-40, 20% glycerol, and 1mM DTT, pH 7.6). The FLAG-CHD3 protein was eluted with 0.3 mg/ml 3XFLAG peptide (in 20mM HEPES, 0.1M KCl, 0.05% NP-40, 20% glycerol, and 1 mM DTT, pH 7.6). Wildtype and mutant protein samples were analyzed by SDS-PAGE and stained with Coomassie Brilliant Blue (Fig. S4). The concentration of the CHD3 proteins was estimated from BSA standards in SDS-PAGE gels stained with Coomassie Brilliant Blue.

Radiometric ATPase assay

Each ATPase reaction (10 µL) contained 20mM Tris–HCl, pH 7.5, 1 mM MgCl2, 0.1 mg/ml BSA, 1mM DTT, 100 µM ATP, 1 µCi of [γ-^{32}P]ATP as a tracer. 25 nM of each CHD3 purified protein was incubated with 70 nM of recombinant nucleosomes or naked dsDNA. Nucleosome was reconstituted by the salt gradient dialysis method using recombinant histone octamer and 201 bp 601 DNA fragment[46]. The reactions were initiated by the addition of nucleosome or DNA substrate and incubated at 37 °C for 40 min. The reaction was quenched by the addition of EDTA to a final concentration of 100 mM. Aliquots (2.5 µL) were removed and spotted on PEI-cellulose thin-layer chromatography plates and developed in 1M formic acid and 0.5M LiCl. ATP hydrolysis was quantified using a Phosphorimager with Image Quant Software. For the mixing experiment, all reaction components except for CHD3 protein were incubated for 10 min at 37 °C, and the CHD3 protein mixture was added last to start the reaction. This experiment was performed 3 times per condition (N = 3) for all conditions, except for the conditions “no CHD3”, “WT 12.5 nM” and “WT 25 nM” (N = 2).

For the quantification analysis, we performed 3 individual experiments for each of the two biological replicates (total N = 6), except for the p.Trp1158Arg mutant (one biological replicate, total N = 3). An unpaired t-test was used to determine whether the activity of the mutant proteins differed significantly from wild-type protein activity.

Restriction enzyme accessibility assay

Remodeling activities were measured with a restriction enzyme accessibility assay as previously described[27]. 12.5 nM nucleosomes (347 bp) were incubated with the indicated amounts of CHD3 proteins at 37 °C for 60 min in the remodeling buffer (20mM Tris–HCl pH 7.5, 1mM DTT, 1mM $MgCl_2$, 1mM ATP, 0.1 mg/ml BSA, and 5 U HhaI). The reactions were stopped by adding 2 µL of proteinase K buffer containing 6.7 mg/ml proteinase K, 167mM EDTA, and 1.7% SDS. After incubation at 50 °C for 10 min, the DNAs were analyzed by 6% native polyacrylamide gel electrophoresis. The separated DNA fragments were visualized with UV light on the ChemiDox XRS system (BIO-RAD). The band intensities were quantified by ImageJ.

Cloning constructs for immunofluorescence

Wild-type CHD3 (NM_001005273.2) was amplified by PCR and cloned into pCR2.1-TOPO (Invitrogen) as described[47]. CHD3 mutation constructs were generated using the QuikChange II Site-Directed Mutagenesis Kit (Agilent), primer sequences are provided in Fig. S1. CHD3 cDNAs were subcloned using BamHI/ NheI restriction sites into a modified pmCherry-C1 vector (Clontech). All constructs were verified by Sanger sequencing.

Immunofluorescence

HEK293 cells were obtained from ECACC (Catalogue number 85120602) and grown in Dulbecco's modified Eagle's medium (Invitrogen), supplemented with 10% fetal bovine serum (Invitrogen). Transfection was performed using GeneJuice (Merck-Millipore). The cells were seeded onto coverslips coated with poly-L-lysine (Sigma). At 36 h post-transfection, cells were fixed using 4% paraformaldehyde solution (Electron Microscopy Sciences) for 10 min at room temperature. The mCherry fusion proteins were visualized by direct fluorescence, nuclei were visualized with Hoechst 33342 (Invitrogen). Fluorescence images were obtained using an Axiovert A-1 fluorescent microscope with ZEN Image Software (Zeiss).

Three-dimensional modeling

As no experimentally solved 3D-structure of CHD3 exists, we performed homology modeling using the modeling option with standard parameters in the YASARA[48] & WHAT IF[49] twinset. Several models of the ATPase/helicase domain were created. The best scoring model was based on template PDB-file 5JXR (sequence identity 41% over the aligned residues). We also studied the model based on PDB-file 3MWY (sequence identity 45%), which shows a more open conformation and contains an ATP substitute. These two models provided information about the relative position of the mutated residues in the different conformation of the protein complex.

Clustering analysis of missense mutations

The locations of observed de novo missense mutations were permutated 1,000,000 times over the cDNA of the *CHD3* gene (RefSeq transcript: NM_001005273.2). The distances between missense mutations were adjusted to take into account the total size of the coding region of *CHD3* (6003 bp). Then, the geometric mean (the *n*th root of the product of *n* of all distances separating the mutations) was calculated, giving an index of clustering, as previously described[30]. To circumvent a mean distance of 0 as the result of recurrent mutations, pseudocount (adding 1 to all distances and 1 to the gene size) was used. To avoid artificial deflation of the clustering *P*-value, only one of the recurrent mutations present in the sibling-pair (individuals 7 and 8) and twin-pair (individuals 20 and 21) were included for the analysis.

Acknowledgements

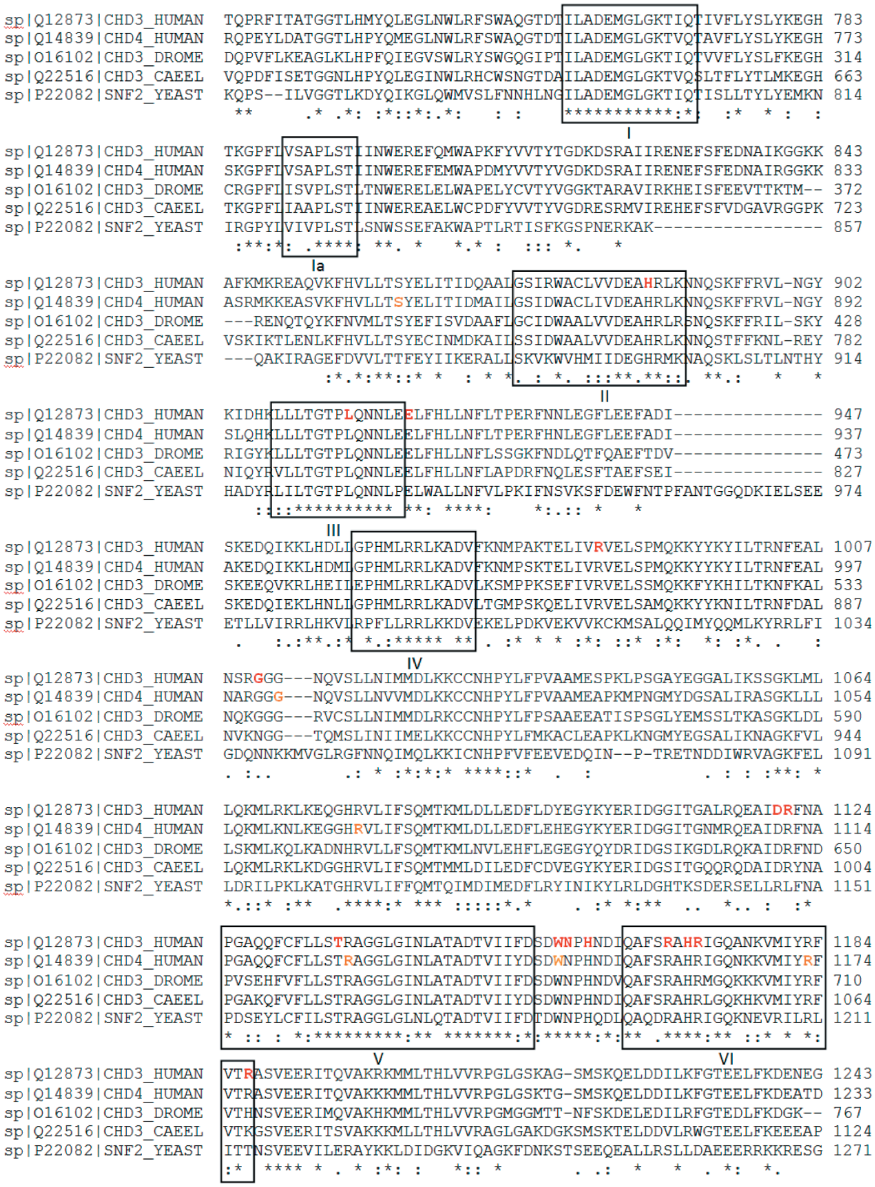
We thank all individuals and families for their contribution. We thank Amaia Carrión Castillo and Else Eising for assistance with the WGS analysis of the index individual, and Sarah Graham and Elliot Sollis for cloning the wild-type CHD3 construct for immunofluorescence. This work was supported by the Netherlands Organization for Scientific Research (NWO) Gravitation Grant 24.001.006 to the Language in Interaction Consortium (L.S.B., S.E.F., and H.G.B.), the Max Planck Society (S.E.F.), the National Institute on Deafness and Other Communication Disorders Grant DC000496 (L.Sh.) and a core grant to the Waisman Center from the National Institute of Child Health and Human Development (Grant U54 HD090256) to L.Sh., the Canadian Institutes of Health Research Grants MOP-119595 and PJT-148830 to W.T.G. Individuals 11, 16, 24, and 28 were part of The DDD Study cohort. The DDD Study presents independent research commissioned by the Health Innovation Challenge Fund [Grant number HICF-1009- 003], a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute [Grant number WT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The DDD study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/ 284/12, granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network.

References

1. Marfella, C. G. & Imbalzano, A. N. The Chd family of chromatin remodelers. *Mutat. Res.* **618**, 30–40 (2007).
2. Hargreaves, D. C. & Crabtree, G. R. ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell Res.* **21**, 396–420 (2011).
3. Ho, L. & Crabtree, G. R. Chromatin remodelling during development. *Nature* **463**, 474–484 (2010).
4. Carvill, G. L. et al. Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat. Genet.* **45**, 825–830 (2013).
5. Vissers, L. E. et al. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat. Genet.* **36**, 955–957 (2004).
6. O'Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
7. Bernier, R. et al. Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* **158**, 263–276 (2014).
8. Weiss, K. et al. De novo mutations in CHD4, an ATP-dependent chromatin remodeler gene, cause an intellectual disability syndrome with distinctive dysmorphisms. *Am. J. Hum. Genet.* **99**, 934–941 (2016).
9. Pilarowski, G. O. et al. Missense variants in the chromatin remodeler CHD1 are associated with neurodevelopmental disability. *J. Med. Genet.* **55**, 561–566 (2017).
10. Zhang, Y., LeRoy, G., Seelig, H. P., Lane, W. S. & Reinberg, D. The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. *Cell* **95**, 279–289 (1998).
11. Woodage, T., Basrai, M. A., Baxevanis, A. D., Hieter, P. & Collins, F. S. Characterization of the CHD family of proteins. *Proc. Natl Acad. Sci. USA* **94**, 11472–11477 (1997).
12. Kolla, V. et al. The tumour suppressor CHD5 forms a NuRD-type chromatin remodelling complex. *Biochem. J.* **468**, 345–352 (2015).
13. Lai, A. Y. & Wade, P. A. Cancer biology and NuRD: a multifaceted chromatin remodelling complex. *Nat. Rev. Cancer* **11**, 588–596 (2011).
14. Basta, J. & Rauchman, M. The nucleosome remodeling and deacetylase complex in development and disease. *Transl. Res.* **165**, 36–47 (2015).
15. Nitarska, J. et al. A functional switch of NuRD chromatin remodeling complex subunits regulates mouse cortical development. *Cell Rep.* **17**, 1683–1698 (2016).
16. Eising, E. et al. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol. Psychiatry.* ***https://doi.org/10.1038/s41380-018-0020-x*** (2018).
17. Estruch, S. B., Graham, S. A., Deriziotis, P. & Fisher, S. E. The languagerelated transcription factor FOXP2 is post-translationally modified with small ubiquitin-like modifiers. *Sci. Rep.* **6**, 20911 (2016).
18. Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–523 (2001).
19. Graham, S. A. & Fisher, S. E. Understanding language from a genomic perspective. *Annu. Rev. Genet.* **49**, 131–160 (2015).
20. Morgan, A., Fisher, S.E., Scheffer, I. & Hildebrand, M. FOXP2-related Speech and Language Disorders. in *GeneReviews(R)* (eds Pagon, R.A. et al.) (Seattle, WA, 2017).
21. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
22. Yuen, R. K. et al. Genome-wide characteristics of de novo mutations in autism. *Genom. Med.* **1**, 16027 (2016).
23. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
24. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum. Mutat.* **36**, 928–930 (2015).
25. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
26. Lu, H. C. et al. Disruption of the ATXN1-CIC complex causes a spectrum of neurobehavioral phenotypes in mice and humans. *Nat. Genet.* **49**, 527–536 (2017).
27. Liu, X., Li, M., Xia, X., Li, X. & Chen, Z. Mechanism of chromatin remodelling revealed by the Snf2-nucleosome structure. *Nature* **544**, 440–445 (2017).
28. Hodges, H. C. et al. Dominant-negative SMARCA4 mutants alter the accessibility landscape of tissue-unrestricted enhancers. *Nat. Struct. Mol. Biol.* **25**, 61–72 (2018).

29. Sifrim, A. et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).
30. Lelieveld, S. H. et al. Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorder-associated genes. *Am. J. Hum. Genet.* **101**, 478–484 (2017).
31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
33. Neveling, K. et al. A post-hoc comparison of the utility of Sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Hum. Mutat.* **34**, 1721–1726 (2013).
34. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
35. Hollink, I. et al. Acute myeloid leukaemia in a case with Tatton–Brown–Rahman syndrome: the peculiar DNMT3A R882 mutation. *J. Med. Genet.* **54**, 805–808 (2017).
36. Gibson, W. T. et al. Mutations in EZH2 cause Weaver syndrome. *Am. J. Hum. Genet.* **90**, 110–118 (2012).
37. Cohen, A. S. et al. A novel mutation in EED associated with overgrowth. *J. Hum. Genet.* **60**, 339–342 (2015).
38. Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
39. Thevenon, J. et al. Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. *Clin. Genet.* **89**, 700–707 (2016).
40. Louie, R. J. et al. Novel pathogenic variants in FOXP3 in fetuses with echogenic bowel and skin desquamation identified by ultrasound. *Am. J. Med. Genet.* A **173**, 1219–1225 (2017).
41. Hempel, M. et al. De novo mutations in CHAMP1 cause intellectual disability with severe speech impairment. *Am. J. Hum. Genet.* **97**, 493–500 (2015).
42. Popp, B. et al. Do the exome: a case of Williams–Beuren syndrome with severe epilepsy due to a truncating de novo variant in GABRA1. *Eur. J. Med. Genet.* **59**, 549–553 (2016).
43. Tanaka, A. J. et al. Mutations in SPATA5 are associated with microcephaly, intellectual disability, seizures, and hearing loss. *Am. J. Hum. Genet.* **97**, 457–464 (2015).
44. Kremer, L. S. et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
45. Low, J. K. et al. CHD4 is a peripheral component of the nucleosome remodeling and deacetylase complex. *J. Biol. Chem.* **291**, 15853–15866 (2016).
46. Taguchi, H., Horikoshi, N., Arimura, Y. & Kurumizaka, H. A method for evaluating nucleosome stability with a protein-binding fluorescent dye. *Methods* **70**, 119–126 (2014).
47. Deriziotis, P. et al. De novo TBR1 mutations in sporadic autism disrupt protein functions. *Nat. Commun.* **5**, 4954 (2014).
48. Krieger, E., Koraimann, G. & Vriend, G. Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. *Proteins* **47**, 393–402 (2002).
49. Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 29 (1990).

Supplementary Information



```
sp|Q12873|CHD3_HUMAN      TQPRFITATGGTLHMYQLEGLNWLRFSWAQGTDTILADEMGLGKTIQTIVFLYSLYKEGH 783
sp|Q14839|CHD4_HUMAN      RQPEYLDATGGTLHPYQMEGLNWLRFSWAQGTDTILADEMGLGKTVQTAVFLYSLYKEGH 773
sp|O16102|CHD3_DROME      DQPVFLKEAGLKLHPFQIEGVSWLRYSWGQGIPTILADEMGLGKTIQTVVFLYSLFKEGH 314
sp|Q22516|CHD3_CAEEL      VQPDFISETGGNLHPYQLEGINWLRHCWSNGTDAILADEMGLGKTVQSLTFLYTLMKEGH 663
sp|P22082|SNF2_YEAST      KQPS--ILVGGTLKDYQIKGLQWMVSLFNNHLNGILADEMGLGKTIQTISLLTYLYEMKN 814
                           **     .* .*: :*::*:.*:   : :    ***********:*:  :*  * :  :
                                                                  I

sp|Q12873|CHD3_HUMAN      TKGPFLVSAPLSTIINWEREFQMWAPKFYVVTYTGDKDSRAIIRENEFSFEDNAIKGGKK 843
sp|Q14839|CHD4_HUMAN      SKGPFLVSAPLSTIINWEREFEMWAPDMYVVTYVGDKDSRAIIRENEFSFEDNAIRGGKK 833
sp|O16102|CHD3_DROME      CRGPFLISVPLSTLTNWERELELWAPELYCVTYVGGKTARAVIRKHEISFEEVTTKTM-- 372
sp|Q22516|CHD3_CAEEL      TKGPFLIAAPLSTIINWEREAELWCPDFYVVTYVGDRESRMVIREHEFSFVDGAVRGGPK 723
sp|P22082|SNF2_YEAST      IRGPYLVIVPLSTLSNWSSEFAKWAPTLRTISFKGSPNERKAK----------------- 857
                           :**:*: .****: **. *   *.* :  ::: *.   *
                                  Ia

sp|Q12873|CHD3_HUMAN      AFKMKREAQVKFHVLLTSYELITIDQAALGSIRWACLVVDEAHRLKNNQSKFFRVL-NGY 902
sp|Q14839|CHD4_HUMAN      ASRMKKEASVKFHVLLTSYELITIDMAILGSIDWACLIVDEAHRLKNNQSKFFRVL-NGY 892
sp|O16102|CHD3_DROME      ---RENQTQYKFNVMLTSYEFISVDAAFLGCIDWAALVVDEAHRLRSNQSKFFRIL-SKY 428
sp|Q22516|CHD3_CAEEL      VSKIKTLENLKFHVLLTSYECINMDKAILSSIDWAALVVDEAHRLKNNQSTFFKNL-REY 782
sp|P22082|SNF2_YEAST      ---QAKIRAGEFDVVLTTFEYIIKERALLSKVKWVHMIIDEGHRMKNAQSKLSLTLNTHY 914
                                    :*.*:**::* *  : * *. : *. :::**.**::. **.:   *   *
                                                               II

sp|Q12873|CHD3_HUMAN      KIDHKLLLTGTPLQNNLEELFHLLNFLTPERFNNLEGFLEEFADI--------------- 947
sp|Q14839|CHD4_HUMAN      SLQHKLLLTGTPLQNNLEELFHLLNFLTPERFHNLEGFLEEFADI--------------- 937
sp|O16102|CHD3_DROME      RIGYKLLLTGTPLQNNLEELFHLLNFLSSGKFNDLQTFQAEFTDV--------------- 473
sp|Q22516|CHD3_CAEEL      NIQYRVLLTGTPLQNNLEELFHLLNFLAPDRFNQLESFTAEFSEI--------------- 827
sp|P22082|SNF2_YEAST      HADYRLILTGTPLQNNLPELWALLNFVLPKIFNSVKSFDEWFNTPFANTGGQDKIELSEE 974
                             ::::********** **: ****:    *:.:: *   *
                                    III

sp|Q12873|CHD3_HUMAN      SKEDQIKKLHDLLGPHMLRRLKADVFKNMPAKTELIVRVELSPMQKKYYKYILTRNFEAL 1007
sp|Q14839|CHD4_HUMAN      AKEDQIKKLHDMLGPHMLRRLKADVFKNMPSKTELIVRVELSPMQKKYYKYILTRNFEAL 997
sp|O16102|CHD3_DROME      SKEEQVKRLHEILEPHMLRRLKADVLKSMPPKSEFIVRVELSSMQKKFYKHILTKNFKAL 533
sp|Q22516|CHD3_CAEEL      SKEDQIEKLHNLLGPHMLRRLKADVLTGMPSKQELIVRVELSAMQKKYYKNILTRNFDAL 887
sp|P22082|SNF2_YEAST      ETLLVIRRLHKVLRPFLLRRLKKDVEKELPDKVEKVVKCKMSALQQIMYQQMLKYRRLFI 1034
                           .   :.:**.:* *.:***** **.  :* * * :*: ::* :*:  *: :*. .   :
                                            IV

sp|Q12873|CHD3_HUMAN      NSRGGG---NQVSLLNIMMDLKKCCNHPYLFPVAAMESPKLPSGAYEGGALIKSSGKLML 1064
sp|Q14839|CHD4_HUMAN      NARGGG---NQVSLLNVVMDLKKCCNHPYLFPVAAMEAPKMPNGMYDGSALIRASGKLLL 1054
sp|O16102|CHD3_DROME      NQKGGG---RVCSLLNIMMDLRKCCNHPYLFPSAAEEATISPSGLYEMSSLTKASGKLDL 590
sp|Q22516|CHD3_CAEEL      NVKNGG---TQMSLINIIMELKKCCNHPYLFMKACLEAPKLKNGMYEGSALIKNAGKFVL 944
sp|P22082|SNF2_YEAST      GDQNNKKMVGLRGFNNQIMQLKKICNHPFVFEEVEDQIN--P-TRETNDDIWRVAGKFEL 1091
                          . :..       .: * :*:*:* ****::*  .  :           . : : :**: *

sp|Q12873|CHD3_HUMAN      LQKMLRKLKEQGHRVLIFSQMTKMLDLLEDFLDYEGYKYERIDGGITGALRQEAIDRFNA 1124
sp|Q14839|CHD4_HUMAN      LQKMLKNLKEGGHRVLIFSQMTKMLDLLEDFLEHEGYKYERIDGGITGNMRQEAIDRFNA 1114
sp|O16102|CHD3_DROME      LSKMLKQLKADNHRVLLFSQMTKMLNVLEHFLEGEGYQYDRIDGSIKGDLRQKAIDRFND 650
sp|Q22516|CHD3_CAEEL      LQKMLRKLKDGGHRVLIFSQMTMMLDILEDFCDVEGYKYERIDGSITGQQRQDAIDRYNA 1004
sp|P22082|SNF2_YEAST      LDRILPKLKATGHRVLIFFQMTQIMDIMEDFLRYINIKYLRLDGHTKSDERSELLRLFNA 1151
                          *.::* :**  .****:* *** :::::*.*    . :* *:**  ..  *.. :  :*

sp|Q12873|CHD3_HUMAN      PGAQQFCFLLSTRAGGLGINLATADTVIIFDSDWNPHNDIQAFSRAHRIGQANKVMIYRF 1184
sp|Q14839|CHD4_HUMAN      PGAQQFCFLLSTRAGGLGINLATADTVIIYDSDWNPHNDIQAFSRAHRIGQNKKVMIYRF 1174
sp|O16102|CHD3_DROME      PVSEHFVFLLSTRAGGLGINLATADTVIIFDSDWNPHNDVQAFSRAHRMGQKKKVMIYRF 710
sp|Q22516|CHD3_CAEEL      PGAKQFVFLLSTRAGGLGINLATADTVIIYDSDWNPHNDIQAFSRAHRLGQKHKVMIYRF 1064
sp|P22082|SNF2_YEAST      PDSEYLCFILSTRAGGLGLNLQTADTVIIFDTDWNPHQDLQAQDRAHRIGQKNEVRILRL 1211
                          * :: : *:*********:** *******:*:*****:*:** .****:** ::* * *:
                                         V                                 VI
sp|Q12873|CHD3_HUMAN      VTRASVEERITQVAKRKMMLTHLVVRPGLGSKAG-SMSKQELDDILKFGTEELFKDENEG 1243
sp|Q14839|CHD4_HUMAN      VTRASVEERITQVAKKKMMLTHLVVRPGLGSKTG-SMSKQELDDILKFGTEELFKDEATD 1233
sp|O16102|CHD3_DROME      VTHNSVEERIMQVAKHKMMLTHLVVRPGMGGMTT-NFSKDELEDILRFGTEDLFKDGK-- 767
sp|Q22516|CHD3_CAEEL      VTKGSVEERITSVAKKKMLLTHLVVRAGLGAKDGKSMSKTELDDVLRWGTEELFKEEEAP 1124
sp|P22082|SNF2_YEAST      ITTNSVEEVILERAYKKLDIDGKVIQAGKFDNKSTSEEQEALLRSLLDAEEERRKKRESG 1271
                          :*  **** * . * :*: :   *:: *       . .:  *   *  . *:  *.
```

Figure S1: Conservation of mutated amino acids and clustering around conserved SNF2-motifs

Alignment of amino acids 724-1243 of the CHD3 protein with the Swiss-Prot sequences of human CHD4 (Q14839), CHD3 in drosophila melanogaster (O16102), CHD3 in C. Elegans (Q22516) and SNF2 in yeast (P22082). Missense mutations (affected amino acid residues) in our CHD3 cohort are shown in red, while published mutations in CHD4 are shown in orange[1,2]. The majority of missense mutations affect highly conserved amino acid residues. The missense mutations clearly cluster in or around the known conserved SNF2 motifs (motif I, Ia, II, III, IV, V and VI; in figure depicted by boxes with the respective motif number) of the helicase domain. One missense mutation in CHD3 (p.W1158R) affects the same residue as a previously published mutation in CHD4: p.W1148L1.

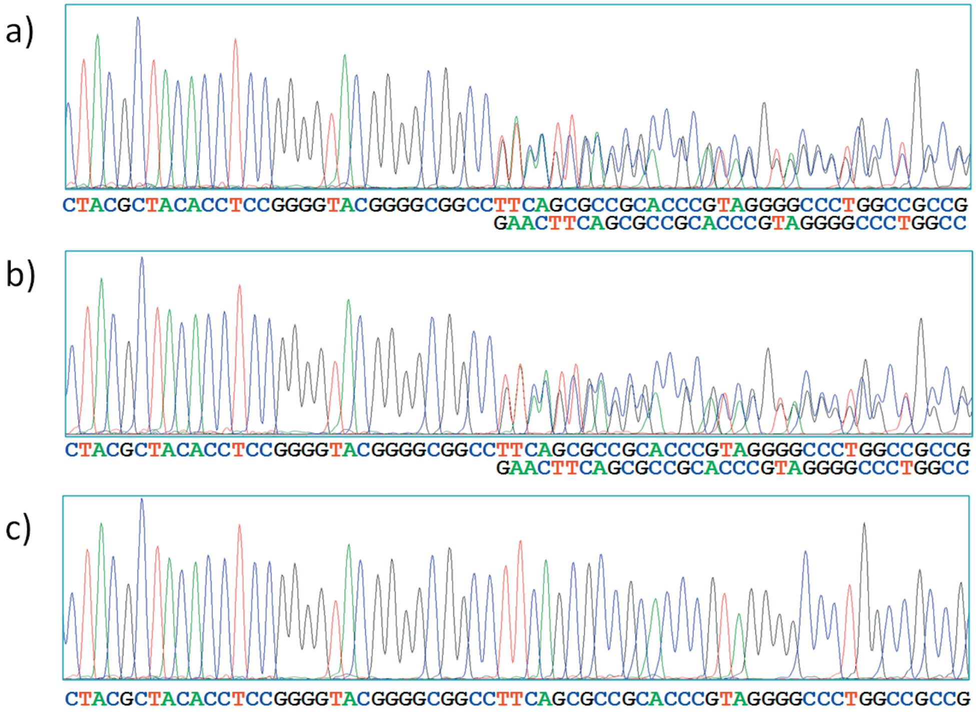


Figure S2: RNA analysis of the c.5802_5803insGAAC mutation (p.(Phe1935Glufs*108))

To study the effects of the frameshift mutation in the penultimate exon of the CHD3 gene [c.5802_5803insGAAC (p.(Phe1935Glufs*108))] in mRNA from individual 35, lymphoblastoid cell lines were generated from peripheral blood cells by Epstein-Barr virus transformation following standard procedures. To check for the occurrence of nonsense mediated decay, mRNA was isolated from cells that were cultured in the presence (a) and absence (b) of cyclohexamide. A negative control was also included (c). CHD3 mRNA was analyzed by the synthesis of cDNA and Sanger sequencing according to standard protocols. Sanger analysis shows no difference between the RNA analysis from untreated cells and the cells that were treated with cyclohexamide. These data indicate that the alternative transcript that is a result of the frameshift mutation is not degraded by nonsense mediated decay. In conclusion, in this individual (individual 35) two transcripts are present: the wild-type transcript and the transcript with the frameshift, that leads to a stop codon after 108 amino acids.

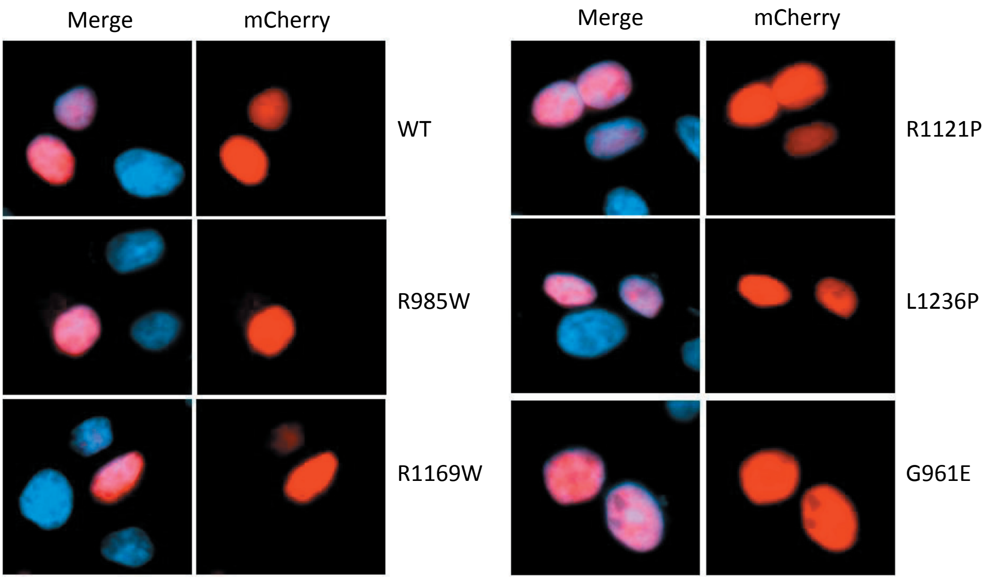


Figure S3: localization of wildtype and mutant CHD3

Fluorescence microscopy images of HEK293 cells transfected with wild-type and synthetic CHD3 variants fused to mCherry (shown in red). Nuclei were stained with Hoechst 33342 (blue). The localization of CHD3 is nuclear in all pictures (wild-type and mutations), no difference in subcellular localization was seen between wild-type and mutants.

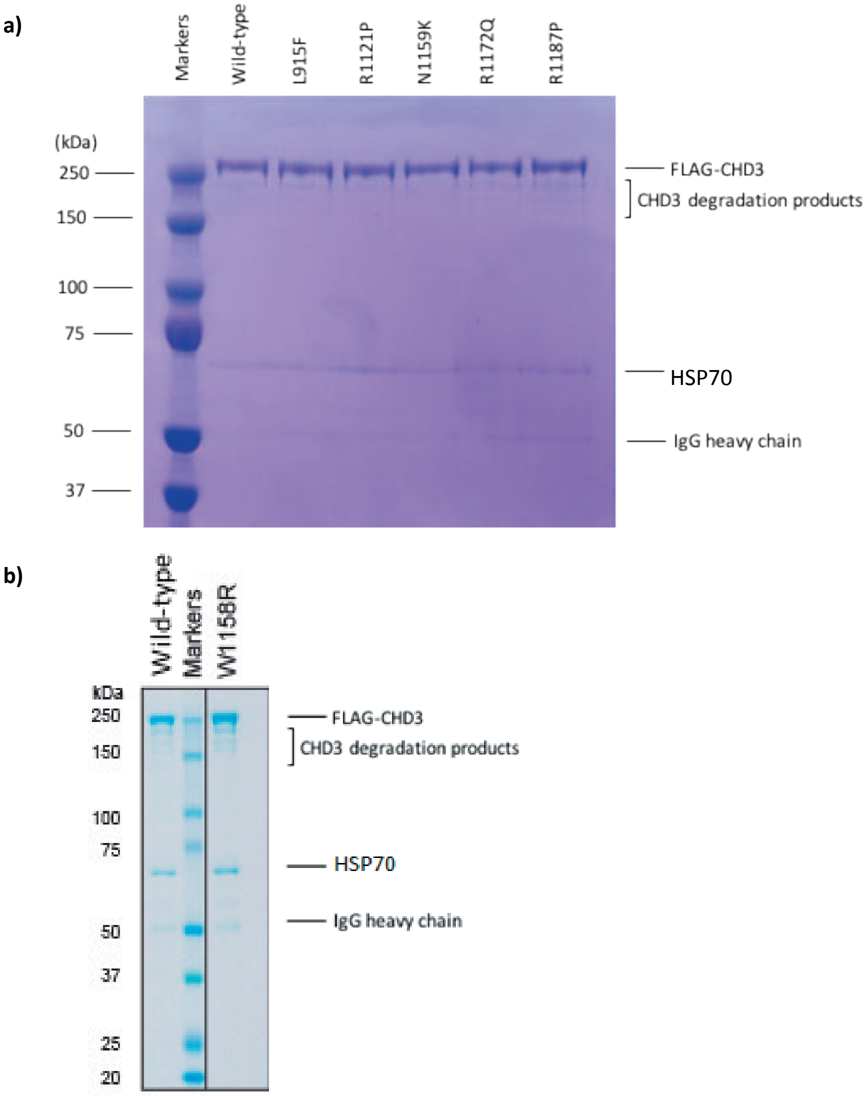


Figure S4: Purification of wild-type and mutant CHD3

Purified wild-type CHD3 and mutant CHD3. 500 ng of the purified proteins were analyzed by SDS-PAGE and stained with Coomassie Brilliant Blue. Proteins were confirmed by mass spectrometry by the NIEHS Mass Spectrometry Research and Support Group.

a) Lane 1: protein marker; lane 2: wild-type FLAG-CHD3 (~230 kDa); lane 3: p.Leu915Phe; lane 4: p.Arg1121Pro; lane 5: p.Asn1159Lys; lane 6: p.Arg1172Gln; and lane 7: p.Arg1187Pro.

b) Lane 1: wild-type FLAG-CHD3 (~230 kDa); lane 2: protein marker and lane 3: p.Trp1158Arg.

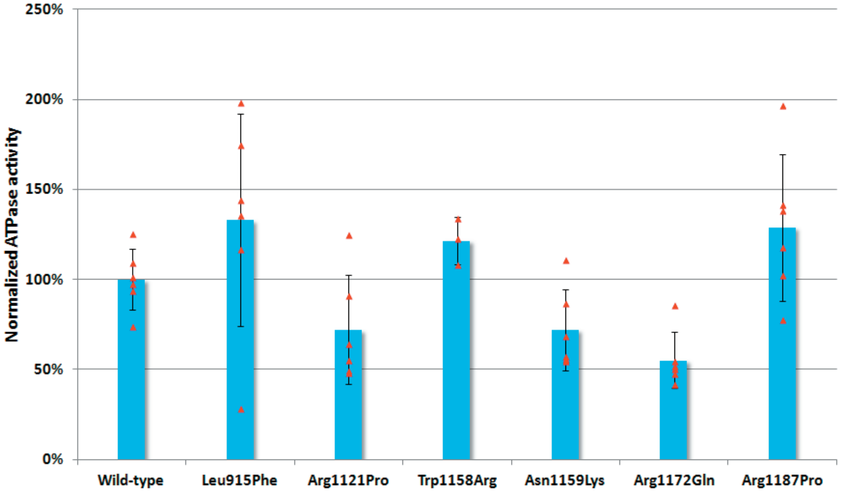


Figure S5: ATPase activity in the absence of DNA substrate

ATPase activities were measured in the absence of DNA substrates. Released phosphate was separated from unhydrolyzed ATP by thin layer chromatography, and detected by exposure to a phosphorimager. The experimental values (percentage hydrolyzed ATP) for the different mutant conditions were normalized to values for the wild-type condition within the experiment, to derive a normalized ATPase activity. Wild-type data points depicted are representative, for individual wild-type replicate values for all experiments please see Supplementary Data 2. All other raw values for these experiments can also be found in Supplementary Data 2. Three independent experiments from two individual purifications (Wild-type, p.Leu915Phe, p.Arg1121Pro, p.Asn1159Lys, p.Arg1172Gln, and p.Arg1187Pro) (N=6) or one purification (p.Trp1158Arg) (N=3) were performed. The experimental data are presented as means +/- standard deviation, individual data points are shown as red triangles.

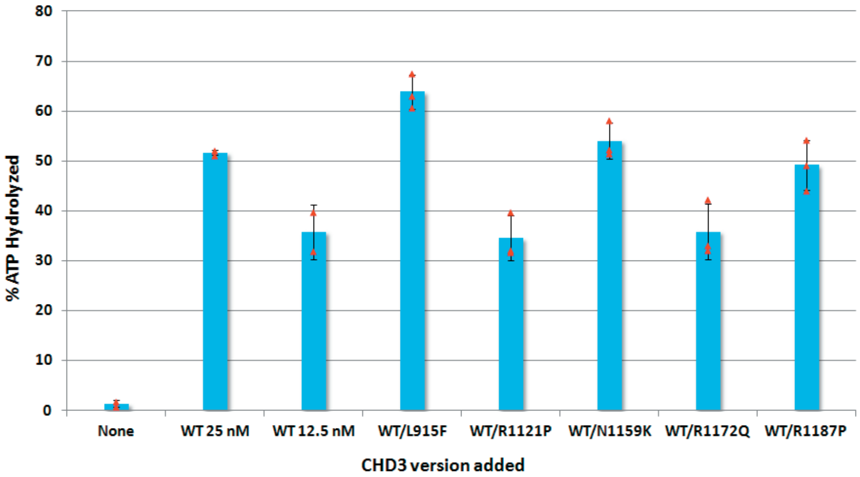


Figure S6: Assay with mixing of wildtype and mutant CHD3

Thin layer chromatography was used to detect nucleosome-dependent ATPase activity. Equimolar concentrations of wild-type (12.5 nM) and mutant protein (12.5 nM) were incubated with nucleosome for 40 minutes. Error bars indicate standard deviations, individual data points are shown as red triangles (N=3 for mixed WT/mutant conditions, N=2 for other conditions). Raw values for these experiments can be found in Supplementary Data 2.

Table S1: Primer sequences

Primers used for cloning of FLAG-CHD3 expression constructs		
	Fw (5’ to 3’)	Rv (5’ to 3’)
	gaaaacctgtattttcagggcaaggcggcagacactgtgatcc	gggtccctgaaagaggacttcaaggtcgtctatacagatcacctcccc
Primers used to create CHD3 mutant constructs for ATPase and remodeling assays		
	Fw (5’ to 3’)	Rv (5’ to 3’)
L915F	ctgacaggaaccccatttcagaataatctggagga	tcctccagattattctgaaatggggttcctgtcag
R1121P	ccaggagcattaaacggatcgatggcctcct	aggaggccatcgatccgtttaatgctcctgg
W1158R	aggaacccccataatgacatccagg	gtcagaatcaaagatgatgacagtgtcagc
N1159K	gtcatcatctttgattctgactggaaaccccataatgacat	atgtcattatggggtttccagtcagaatcaaagatgatgac
R1172Q	gccgggctcatcagattggccaggc	gcctggccaatctgatgagcccggc
R1187P	Cttccactgacgcgggagtcacaaaccgg	ccggtttgtgactcccgcgtcagtggaag
Primers used to create CHD3 mutant constructs for immunofluorescence		
	Fw (5’ to 3’)	Rv (5’ to 3’)
R985W	gccaagacagagctcatcgtttgggtggagcta	tagctccacccaaacgatgagctctgtcttggc
R1169W	atccaggcctttagctgggctcatcggattg	caatccgatgagcccagctaaaggcctggat
R1121P	aggaggccatcgatccgtttaatgctcctgg	ccaggagcattaaacggatcgatggcctcct
L1236P	caaatttggcactgaagagccattcaaggatgaaaacgagg	cctcgttttcatccttgaatggctcttcagtgccaaatttg
G961E	tgcatgatttgctggagccacacatgctgcg	cgcagcatgtgtggctccagcaaatcatgca

Supplementary Data

Supplementary Data 1 and 2 are available online via doi:10.1038/s41467-018-06014-6

Supplementary Note 1: Three-dimensional modeling and mutation analysis for de novo CHD3 mutations

Methods

For this analysis, we used the CHD3 de novo mutations found in our cohort and the following sequence of interest:

```
>CHD3_
MKAADTVILWARSKNDQLRISFPPGLCWGDRMPDKDDIRLLPSALGVKKRKRGPKKQKEN
KPGKPRKRKKRDSEEEFGSERDEYREKSESGGSEYGTGPGRKRRRKHREKKEKKTKRRKK
GEGDGGQKQVEQKSSATLLLTWGLEDVEHVFSEEDYHTLTNYKAFSQFMRPLIAKKNPKI
PMSKMMTILGAKWREFSANNPFKGSAAAVAAAAAAAAAAVAEQVSAAVSSATPIAPSGPP
ALPPPPAADIQPPPIRRAKTKEGKGPGHKRRSKSPRVPDGRKKLRGKKMAPLKIKLGLLG
GKRKKGGSYVFQSDEGPEPEAEESDLDSGSVHSASGRPDGPVRTKKLKRGRPGRKKKKVL
GCPAVAGEEEVDGYETDHQDYCEVCQQGGEIILCDTCPRAYHLVCLDPELDRAPEGKWSC
PHCEKEGVQWEAKEEEEEYEEEGEEEGEKEEEDDHMEYCRVCKDGGELLCCDACISSYHI
HCLNPPLPDIPNGEWLCPRCTCPVLKGRVQKILHWRWGEPPVAVPAPQQADGNPDVPPPR
PLQGRSEREFFVKWVGLSYWHCSWAKELQLEIFHLVMYRNYQRKNDMDEPPPLDYGSGED
DGKSDKRKVKDPHYAEMEEKYYRFGIKPEWMTVHRIINHSVDKKGNYHYLVKWRDLPYDQ
STWEEDEMNIPEYEEHKQSYWRHRELIMGEDPAQPRKYKKKKKELQGDGPPSSPTNDPTV
KYETQPRFITATGGTLHMYQLEGLNWLRFSWAQGTDTILADEMGLGKTIQTIVFLYSLYK
EGHTKGPFLVSAPLSTIINWEREFQMWAPKFYVVTYTGDKDSRAIIRENEFSFEDNAIKG
GKKAFKMKREAQVKFHVLLTSYELITIDQAALGSIRWACLVVDEAHRLKNNQSKFFRVLN
GYKIDHKLLLTGTPLQNNLEELFHLLNFLTPERFNNLEGFLEEFADISKEDQIKKLHDLL
GPHMLRRLKADVFKNMPAKTELIVRVELSPMQKKYYKYILTRNFEALNSRGGGNQVSLLN
IMMDLKKCCNHPYLFPVAAMESPKLPSGAYEGGALIKSSGKLMLLQKMLRKLKEQGHRVL
IFSQMTKMLDLLEDFLDYEGYKYERIDGGITGALRQEAIDRFNAPGAQQFCFLLSTRAGG
LGINLATADTVIIFDSDWNPHNDIQAFSRAHRIGQANKVMIYRFVTRASVEERITQVAKR
KMMLTHLVVRPGLGSKAGSMSKQELDDILKFGTEELFKDENEGENKEEDSSVIHYDNEAI
ARLLDRNQDATEDTDVQNMNEYLSSFKVAQYVVREEDKIEEIEREIIKQEENVDPDYWEK
LLRHHYEQQQEDLARNLGKGKRVRKQVNYNDAAQEDQDNQSEYSVGSEEEDEDFDERPEG
RRQSKRQLRNEKDKPLPPLLARVGGNIEVLGFNTRQRKAFLNAVMRWGMPPQDAFTTQWL
VRDLRGKTEKEFKAYVSLFMRHLCEPGADGSETFADGVPREGLSRQQVLTRIGVMSLVKK
KVQEFEHINGRWSMPELMPDPSADSKRSSRASSPTKTSPTTPEASATNSPCTSKPATPAP
SEKGEGIRTPLEKEEAENQEEKPEKNSRIGEKMETEADAPSPAPSLGERLEPRKIPLEDE
VPGVPGEMEPEPGYRGDREKSATESTPGERGEEKPLDGQEHRERPEGETGDLGKREDVKG
DRELRPGPRDEPRSNGRREEKTEKPRFMFNIADGGFTELHTLWQNEERAAISSGKLNEIW
HRRHDYWLLAGIVLHGYARWQDIQNDAQFAIINEPFKTEANKGNFLEMKNKFLARRFKLL
EQALVIEEQLRRAAYLNLSQEPAHPAMALHARFAEAECLAESHQHLSKESLAGNKPANAV
LHKVLNQLEELLSDMKADVTRLPATLSRIPPIAARLQMSERSILSRLASKGTEPHPTPAY
PPGPYATPPGYGAAFSAAPVGALAAAGANYSQMPAGSFITAATNGPPVLVKKEKEMVGAL
VSDGLDRKEPRAGEVICIDD
```

Mutation numbering corresponds with transcript variant 1, NM_001005273.

As no experimentally solved 3D-structure of CHD3 exists, we performed homology modeling using the modeling option with standard parameters in the YASARA[3] & WHAT IF[4] twinset. Several models of the ATPase/helicase domain were created. The best scoring model was based on template PDB-file 5JXR (M. thermophilia MtISWI, sequence identity 41% over the aligned residues), which shows a closed conformation of the ATPase/helicase domain. We also studied the model based on PDB-file 3MWY (yeast Chd1, sequence identity 45%) which shows a more open conformation but contains an ATP substitute. Both templates represent an auto-inhibited form, however, it is impossible to say which of these models best represents the real biological form of CHD3, since movement of the domains is probably important for correct functioning of the protein. Therefore, both models were studied.

Results

Introduction of models and mutations

The complete CHD3 protein (2000 amino acids in isoform 1) is even bigger than the modeled domains shown here. Model 1 (closed conformation, based on 5JXR) contains residues 500-1290, while model 2 (open conformation, based on 3MWY) contains residues 445-1413.

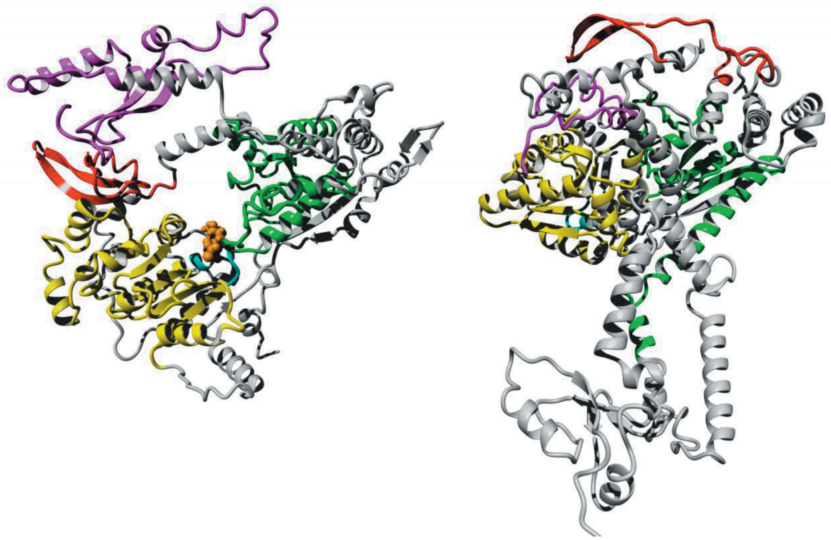


Figure SN1: Overview of the two CHD3- models used in this study. Model 1 on the left represents the opened conformation with ATP (orange) bound. The model on the right (model 2) represents a more closed conformation, in which the two ATP-domains are interacting.

Domains indicated are: chromo domain 1 (494-595; magenta), chromo domain 2 (631-673; red), helicase ATP binding domain (yellow), helicase C-terminal domain (green), ATP binding residues (761-768; cyan). Grey residues do not belong to an indicated domain.

Figure SN1 shows that the position of the domains relative to each other can possibly change depending on removal of inhibition and complex formation (with ATP, but probably also with other molecules, co-factors and DNA). In the open conformation, a wide gap exists between the Helicase ATP-binding domain (yellow) and Helicase C-terminal domain (green), while in the closed conformation no gap is seen. The existence of at least two different conformations indicates that it is necessary to study the effects of the mutations in these two different models.

It is unclear what triggers the conformational change. The authors of the original articles that belong to the templates[5,6] speculate that the Chromodomains are important for auto-inhibition and differentiation between naked DNA and DNA with nucleosomes. If that is true, both models here represent an autoinhibited state. However, one of them contains ATP and the other one shows more interaction within the ATP binding domains.
Also, it might very well be possible that the closed conformation contains ATP while the open conformation does not (the opposite of what is shown here). This is one of the limitations of modeling; substrates present or missing in the template will also be present or missing in the model.
The missense mutations were mapped on the models as is shown in figure SN2.

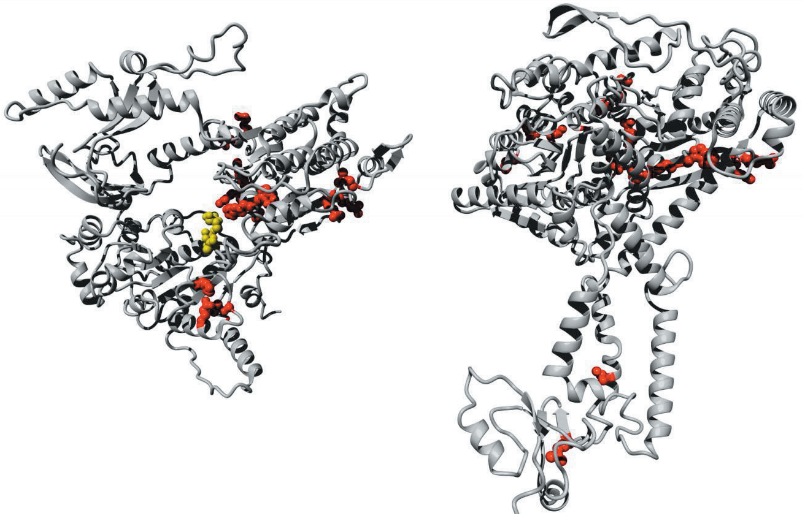


Figure SN2: shows the position of all missense mutations studied here in both conformations. The positions of the mutated residues are indicated in red, the sidechains of these residues are shown as red balls. The ATP molecule is shown in yellow.

We performed a detailed analysis of the three-dimensional modeling of the mutations. As the CHD3 mutations are found largely clustered in the conserved motifs characteristic of the SF2 superfamily of helicases/translocases, we combined the 3D modeling analysis with information from the literature on these conserved motifs. A summary of this analysis is provided below.

Table SN1: His886Arg, Leu915Phe, Glu921Lys and Gly961Glu

	Model 1 (open state, with ATP)	Model 2 (closed confirmation)	Conserved motifs?
His886Arg	Located on interaction surface of Helicase ATP-binding domain. Might be responsible for correct interactions to facilitate ATP-binding. Mutation into an Arginine, which is bigger and positively charged, might affect the function of the protein domain.	Residue also located on surface, but does not interact with other half of Helicase domain. However, still in ATP-binding domain, and responsible for stabilizing interactions, which might be lost due to the mutation.	His886 is part of the Walker B motif in motif II. This motif coordinates the catalytic Mg++ involved in ATP hydrolysis.
Leu915Phe	Leu915 residue is semiburied, and in close contact with His886 and Glu921. It seems to make hydrophobic interactions that stabilize ATP-binding domain. A bigger residue like Phe will not fit and destabilizes the protein.	Leu915 is close to Glu921 and Gly961 (in this model His886 is a bit further away). Leu915 makes still hydrophobic interactions, probably important for interaction surface with other helicase domain. A Phe residue will not fit here without causing reorganisation of surrounding residues.	Leu915 is located in conserved motif III.
Glu921Lys	This residue is surrounded by His886, Gly961 and Leu915. It can be found at the surface but makes interactions that will be lost when mutated into Lys with opposite charge.	This residue is semiburied and makes hydrogenbonds and saltbridges, thereby stabilizing the domain. Close to Leu915 and His886. Mutation into Lys will destabilize the area since Lys carries an opposite charge and has a different shape.	Glu921 is located just next to motif III.
Gly961Glu	This residue clusters with His886, Leu915 and Glu921; although these other three are closer together. Gly is small and flexible, and located close to a Proline at the end of a helix. Mutation into Glu will introduce a much bigger and less flexible residue, this will cause a structural change that might affect the interaction surface and protein function.	Gly961 is still close to Leu915. It is also more clearly located on the surface that interacts with the other half of the helicase domain. Mutation might affect the local structure.	Gly961 is part of conserved motif IV.

Conclusion

These four residues are the only four mutated residues in the Helicase ATP-binding domain. In model 1 it is clearly visible that these four residues cluster together and mutation of one of them could affect the position of the others. Clustering is not so clear in the closed model (model 2), but the residues still appear in the same area. The two Helicase lobes are closer together in model 1, and therefore a close interaction with other mutated residues becomes visible as well. For example, in this model Asp1120 and Arg1121 are close to Gly961 and Leu915. Figure SN3 and SN4 below illustrate this clustering.

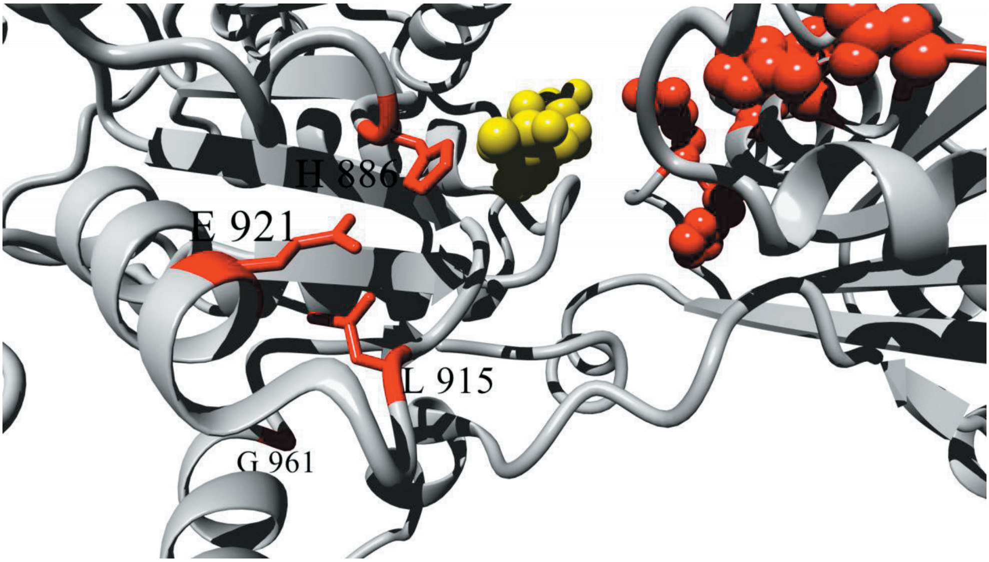


Figure SN3: ATP bound model (model 1) based on 3MWY. In this model, residue H886, E921, L915 and G961 appear close together and might be responsible for the correct structure to bind ATP and to interact with the other half of the ATP binding domain (shown on the right with a subset of the other mutations shown as red balls) It seems possible that mutation of any of these residues would affect the correct conformation of this domain, which might affect ATP binding and/or correct interaction with the other half.

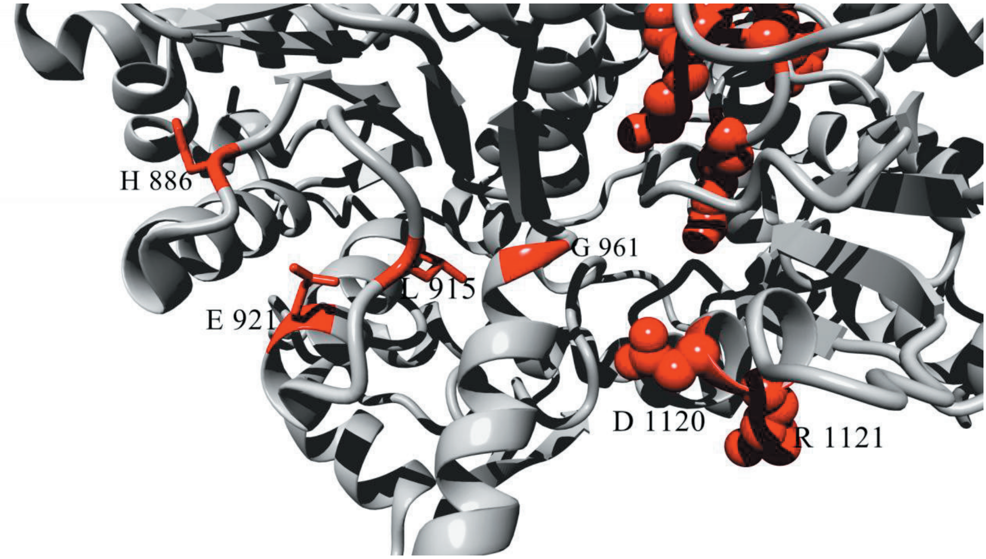


Figure SN4: The ATP-free model (model 2) based on 5JXR shows the position of the residues mentioned above. Clustering in this model seems less obvious, but the residues are still located in the same domain. In this closed conformation, without ATP, it becomes clear how closely the mutations located in the two different helicase lobes could be located to each other (see the position of the labeled D1120 and R1121, other mutations in the other domain are shown as red balls).

Table SN2: Arg985Trp, Arg985Gln, Arg1187Pro and Leu1236Pro

	Model 1 (open state, with ATP)	Model 2 (closed confirmation)	Conserved motifs?
Arg985Trp / Arg985Gln	Arg985 is located on the surface of the Helicase C-terminal domain. It is not closely located to the ATP-binding or interaction surface. It makes a saltbridge and could be responsible for overall stability of this part of the protein. It clusters with Arg1187. Mutation of Arg into Trp might cause folding problems. Mutation into Glu would be easier, but stabilizing saltbridges would still be lost.	In this model the last C-terminal tail has a completely different conformation, which might be caused by the crystallization process. Originally, this model is a dimer with the last C-terminal tail swapped, and as a result Leu1236 is not close to Arg985 and Arg1187. Also, this model does not contain the C-terminal bridge (see Leu1326Pro), and Arg only seems to add some stable interactions to the area, contributing to general stable protein folding. Mutation into either Gln or Trp will affect this folding.	This mutation is located outside the canonical helicase motifs.
Arg1187Pro	This residue clusters with Arg985 on the surface of the protein's C-terminal helicase domain. It might mediate interaction with a regulatory unit. The residue makes a saltbridge, which will be lost due to the mutation. Mutation into Pro will change the backbone conformation. Regardless of interactions with a regulatory unit, this mutation will change local protein structure and might affect stability and function.	This residue clusters with Arg985 on the surface of the protein's C-terminal helicase domain. Mutation into Pro will change the backbone conformation, and will change local protein structure and might affect stability and function.	Arg1187 is part of motif VI, a motif that contains multiple mutations, and contributes to ATP binding and hydrolysis.
Leu1236Pro	In the ATP bound model, the residue is located closely to Arg985 and Arg1187 in a C-terminal bridge that is suggested to be important for regulation[5].	Leu1236 is found in a helix that is used for domain swapping in the crystal. The biological function of this helix is unclear. However, it is known that any mutation into Pro will affect the local backbone structure, because Pro is the only amino acid that will force the backbone into a rigid turn. Regardless of the exact position of the residue, this mutation can affect the protein's structure.	This mutation is located outside the canonical helicase motifs.

Conclusion

Residues Arg985, Arg1187 and maybe Leu1236 are found close together. It is unclear what the function of the last C-terminal tail is, although it has been suggested to have a regulatory effect[5]. In that case, interactions with the tail are important. If this is not the case, both mutations Arg1187Pro and Arg985Trp/Gln are expected to affect the structure and thereby maybe affect the function as well.

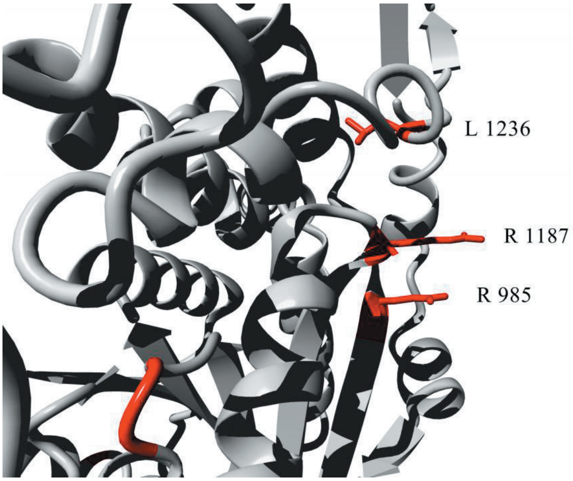


Figure SN5: Positions of residue R985, R1187 and L1236 in the ATP-bound model. The three residues appear close together, although L1236 is officially not part of the helicase C-terminal domain.

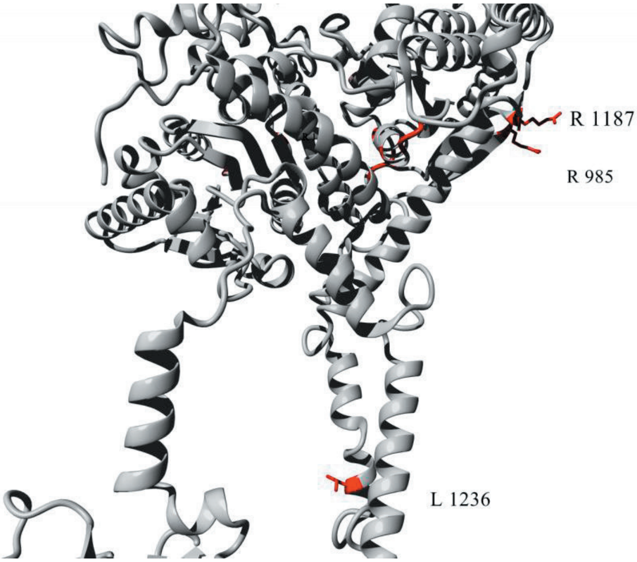


Figure SN6: Positions of the residues R985, R1187 and L1236 in the ATP-free model. In this model, we can see that the L1236 residue is located in a very different position. This is due to the different position of the last C-terminal tail. It is unclear whether this tail has a biologically relevant function.

Table SN3: Asp1120His and Arg1121Pro

	Model 1 (open state, with ATP)	Model 2 (closed confirmation)	Conserved motifs?
Asp1120His	Asp1120 is part of a helix in the C-terminal helicase domain, and clusters with Arg1121. The residues are not especially close to the ATP, although Asp1120 seems to be located on the possible interaction surface. Asp is negatively charged, which is needed to make correct interactions, mutation into His will affect these interactions. Mutation might affect interaction between domains and/or ATP binding.	In this model Asp1120 gets close to the other half of the helicase domain (Helicase ATP-binding domain), for example close to mutated residue Gly961. Asp is negatively charged, which is needed to make correct interactions, mutation into His will affect these interactions. Mutation might affect interaction between domains and/or ATP binding..	Asp1120 is part of helix integral to conserved motif V. Mutations in Motif V in the context of yeast SNF2 abrogate ATP hydrolysis and remodeling activity.[7,8]
Arg1121Pro	This residue is located in the same helix in the C-terminal helicase domain as Asp1120, although its sidechain points in a different direction. The Arg sidechain is positively charged and makes hydrogen bonds and saltbridges. The Arg 1121 residue is probably not involved in ATP binding or interaction with the other domains, but the interactions it makes might be important for a stable structure. Mutation into Pro will surely affect the structure because interactions will be lost, and Pro will change the backbone conformation	This residue is located in the same helix in the C-terminal helicase domain as Asp1120, although its sidechain points in a different direction. The Arg sidechain is positively charged and makes hydrogen bonds and saltbridges. The Arg 1121 residue is probably not involved in ATP binding or interaction with the other domains, but the interactions it makes might be important for a stable structure. Mutation into Pro will surely affect the structure because interactions will be lost, and Pro will change the backbone conformation.	Arg1121 is part of helix integral to conserved motif V. Mutations in Motif V in the context of yeast SNF2 abrogate ATP hydrolysis and remodeling activity.[7,8]

Conclusion

Asp1120 and Arg1121 are part of a helix in the C-terminal helicase domain, integral to conserved motif V. The mutations are expected to affect interaction between domains and/or ATP-binding (Asp1120His), and to affect the stable structure (Arg1121Pro). Mutations in Motif V in the context of yeast SNF2 affect ATP hydrolysis and remodeling activity.

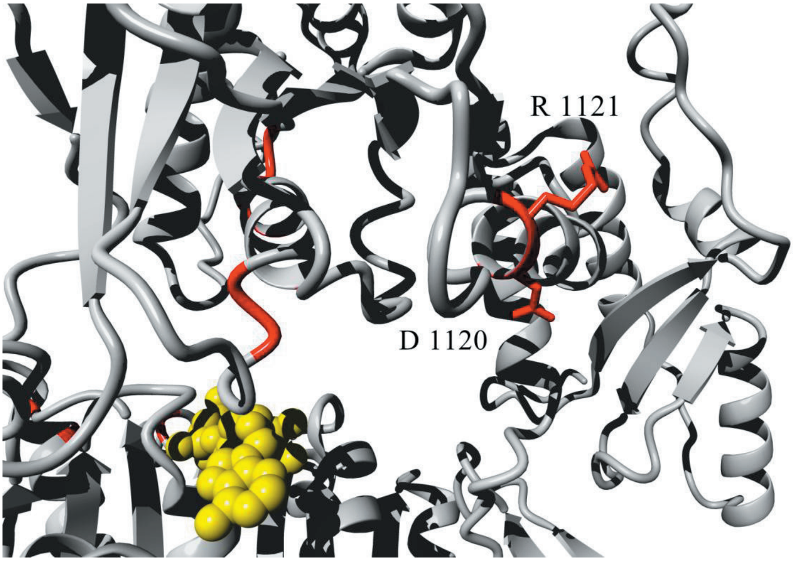


Figure SN7: Positions of R1121 and D1120 in the ATP bound model (model 1). In this model, the residues are not located closely to the ATP, but could contribute to the interaction surface.

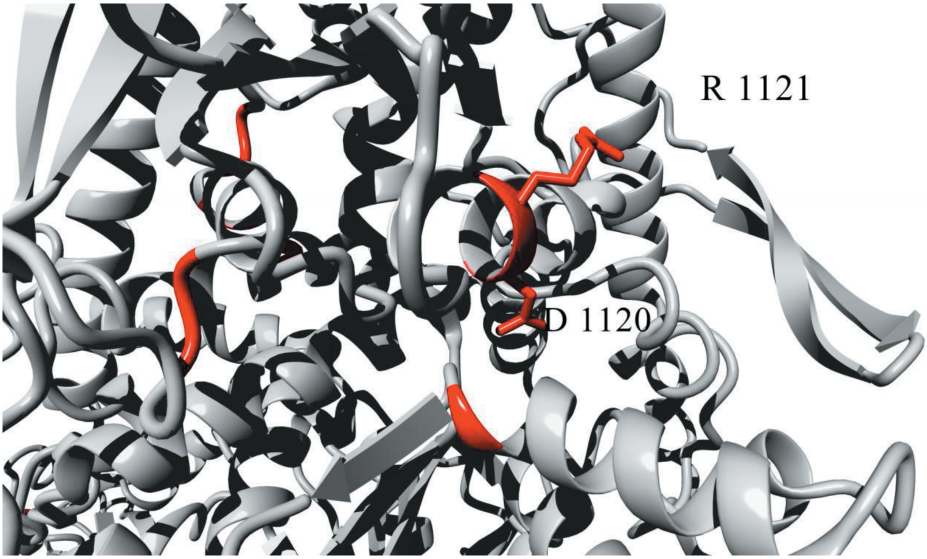


Figure SN8: ATP-free model (model 2), shows the positions of D1120 and R1121. In this case we can see the close proximity of D1120 and G961 (the red residue in the helix right below D1120).

Table SN4: Thr1136Ile, Trp1158Arg, Asn1159Lys, His1161Arg

	Model 1 (open state, with ATP)	Model 2 (closed confirmation)	Conserved motifs?
Thr1136Ile	This residue is located in the core of the protein, making a hydrogenbond with its –OH group and hydrophobic interactions with the methyl in its sidechain. The tight packing does not allow a bigger residue here. Also, the mutation will cause loss of the hydrogenbond and thereby destabilize the local structure. The surrounding residues seem important for correct shape of the interaction site.	This residue is located in the core of the protein, making a hydrogenbond with its –OH group and hydrophobic interactions with the methyl in its sidechain. The tight packing does not allow a bigger residue here. Also, the mutation will cause loss of the hydrogenbond and thereby destabilize the local structure. The surrounding residues seem important for correct shape of the interaction site.	This amino acid is located within Motif V. Mutations in Motif V in the context of yeast SNF2 abrogate ATP hydrolysis and remodeling activity.
Trp1158Arg	This residue is clearly important for the hydrophobic core of the protein. In both models it is (semi) buried and makes hydrophobic interactions. Mutation into anything else will affect the stability and the structure of this domain.	This residue is clearly important for the hydrophobic core of the protein. In both models it is (semi) buried and makes hydrophobic interactions. Mutation into anything else will affect the stability and the structure of this domain.	This Trp residue has recently been shown to be of critical importance in remodeling, by binding nucleosomal DNA in the minor groove. Mutation of this residue impacts remodeling, not ATP hydrolysis[9,10].
Asn1159Lys	This residue is located in the same helix as some of the following mutations (see below) and seems to form the interaction surface with ATP and the other domain. It makes a few hydrogenbonds. Mutation into something larger and positively charged might affect interactions.	This residue is located in the same helix as some of the following mutations(see below). This residue is buried and makes many interactions, some of the interactions are made with residues in a N-terminal helix. This helix might be important for auto-inhibition6. Mutation into something larger and positively charged might affect interactions with surrounding residues or with the inhibition helix.	This amino acid is located adjacent to the Trp residue at position 1158, and might alter the environment of this critical Trp residue.
His1161Arg	This residue follows a similar story to the mutations above. It is located in the same helix as Asn1159 and seems to be important for the surface interactions. Mutation into Arg will change the amino acid properties drastically.	This residue follows a similar story to the mutations above. It is located in the same helix as Asn1159 and seems to be important for the surface interactions. In this model the residue makes interactions with the putative inhibition helix. Mutation into Arg will change the amino acid properties drastically.	This amino acid is located close to the Trp residue at position 1158, and might alter the environment of this critical Trp residue.

Conclusion

Based on recent literature, it is known that the Trp residue at position 1158 is of critical importance in remodeling. The Trp binds nucleosomal DNA in the minor groove, and mutation of this residue will impact chromatin remodeling. The other two residues (1159 and 1161) are probably altering the environment of this critical Trp residue.

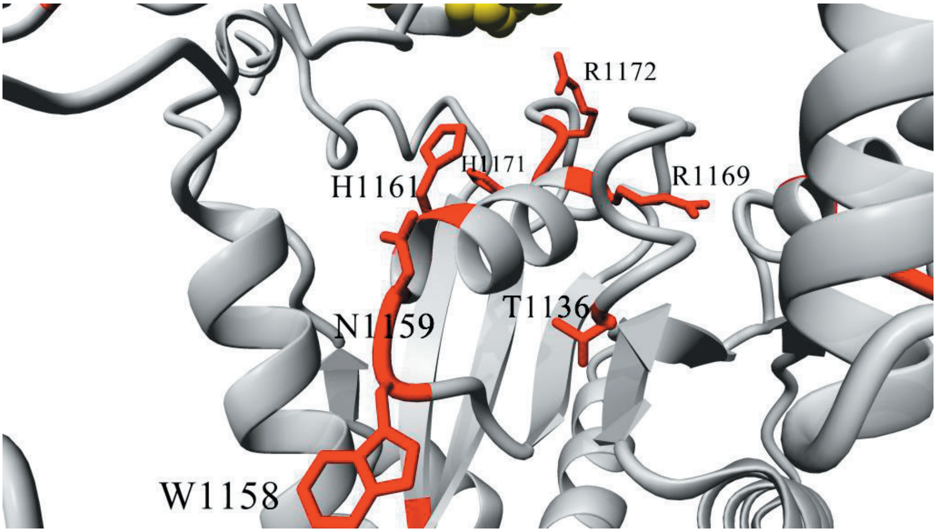


Figure SN9: Overview of the remaining mutations in the core of the protein. Most of these mutations are in the ATP-bound open model located on the surface which becomes buried in the closed model. T1136 is located below this surface, but is required for correct positioning of the other residues.

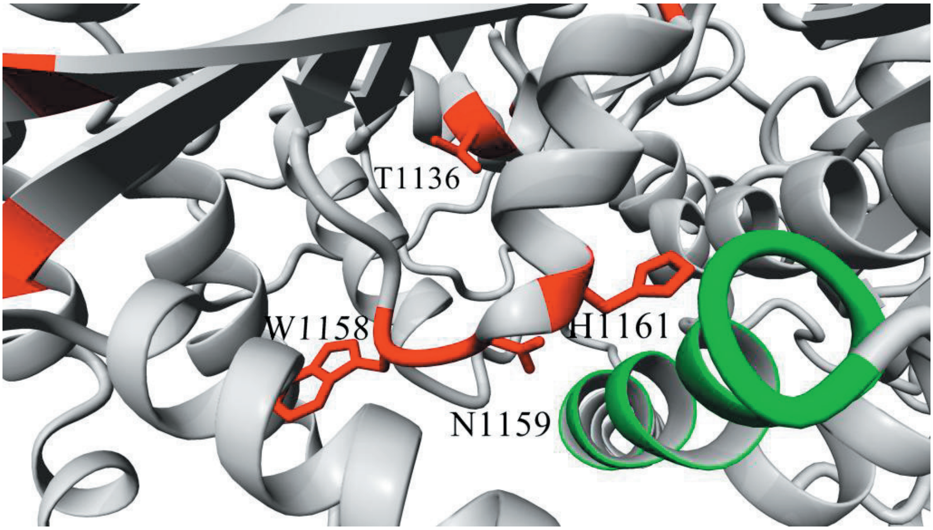


Figure SN10: Mutations T1136,W1158, N1159 and H1161. The W1158 is buried and makes many hydrophobic interactions. N1159 and H1161 become buried in the closed model and seem to interact with residues in a putative inhibition helix (green)[6]. Recent articles show that W1158 is of critical importance in binding nucleosomal DNA[9,10].

Table SN5: Arg1169Trp, His1171Arg and Arg1172Gln

	Model 1 (open state, with ATP)	Model 2 (closed confirmation)	Conserved motifs?
Arg1169Trp	This residue is buried in both models. The residue might be important for correct shaping of the interaction surface.	This residue is buried in both models. In this model the residue is so buried that a bigger Trp will never fit.	This mutation is part of **Motif VI**. Arg1169 is an Arginine finger that is thought to be critical for ATP binding and catalysis.
His1171Arg	This residue is located close to Arg1172 and Arg1169, and also responsible for the correct interaction surface. Mutation of His to Arg introduces a bigger residue with positive charge. This will change the interaction surface.	In this model the residue becomes buried (see figure 11). Mutation of His to Arg introduces a bigger residue with positive charge. This residue will not fit and damages interactions made by His. This mutation will affect protein structure and function.	This mutation is part of **Motif VI**, a critical motif for ATP binding and catalysis.
Arg1172Gln	This mutation might have a similar effect as the mutations above. It is located on the possible interaction surface between the two halves of the helicase domain. Interactions of the Arg will be lost because Gln will not have the same size and charge.	This mutation becomes buried in the closed model. Interactions of the Arg will be lost because Gln will not have the same size and charge.	This mutation is part of Motif VI. Arg1172 is an Arginine finger that is thought to be critical for ATP binding and catalysis.

Conclusion

These three mutations are part of Motif VI, a motif that contributes Arginine fingers (Arg1169 and Arg1172) critical for ATP binding and catalysis. Based on three-dimensional modeling, these three mutations are important for correct shaping of the interaction surface.

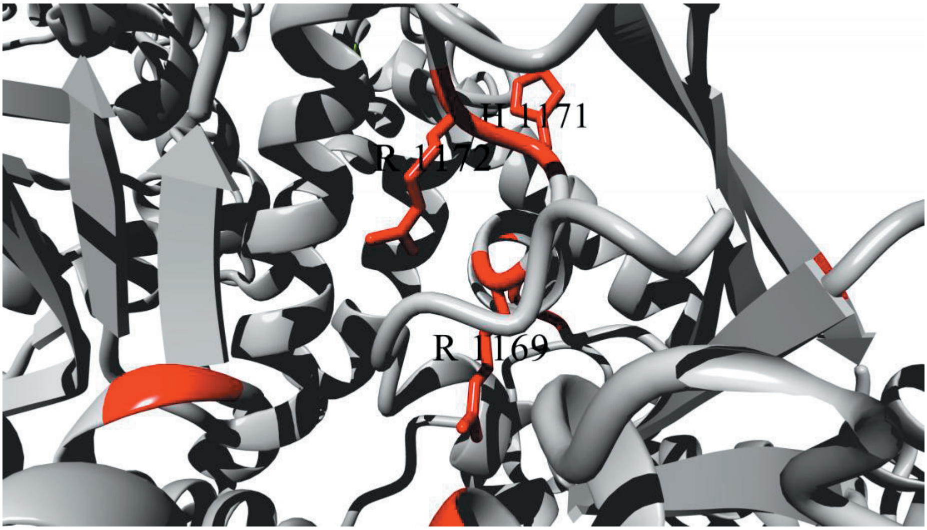


Figure SN11: Residues R1169, H1171 and R1172 in the closed model show that these residues are buried and that any mutation here can damage the protein structure. See Figure 9 for the modeling of these three mutations in the open model (model 1).

Arg1342Gln, Arg1881Leu and Gly1109del

Arg1342Gln

This residue is unfortunately not located in the modeled domain. No information or function is known for this residue or this region and therefore it is difficult to predict the effect of the mutation. Arg and Gln are both hydrophilic, but Arg is postively charged while Gln is neutral. Arg is also is bigger than Gln. Depending on the interactions made by the wild-type, this mutation can be damaging.

Arg1881Leu

This residue is not even close to the modeled helicase domains. Again, it is difficult to predict an effect without structural knowledge. However, the properties of Arg and Leu amino acids are very different. Arg is larger, positively charged and hydrophilic. Leu is smaller, neutral and hydrophobic. The differences in properties between these residues can strongly affect the protein structure.

Gly1109del

This mutation is the only deletion of 1 residue in patients known so far. It deletes a flexible Gly residue in a contact loop. The residue is not located on the interaction surface. Instead, it seems to be important for a particular loop structure that interacts with residues in helix 611-617. Gly itself does not have a sidechain, and therefore is the most flexible residue. The fact that residue 1108 is also a Glyc indicates that this loop might need more flexilbiity. The interaction with the 611-617 helix occurs in both models, although the helix seems much closer to the residue in the closed model.

The effect of this mutation is difficult to predict without knowing the exact function of the loop and the interacting helix. Deletion of 1 residue would result in a shorter loop, but it will not affect folding of the remaining protein. Interaction between the helix and the loop might be required for regulation, but more information is needed.

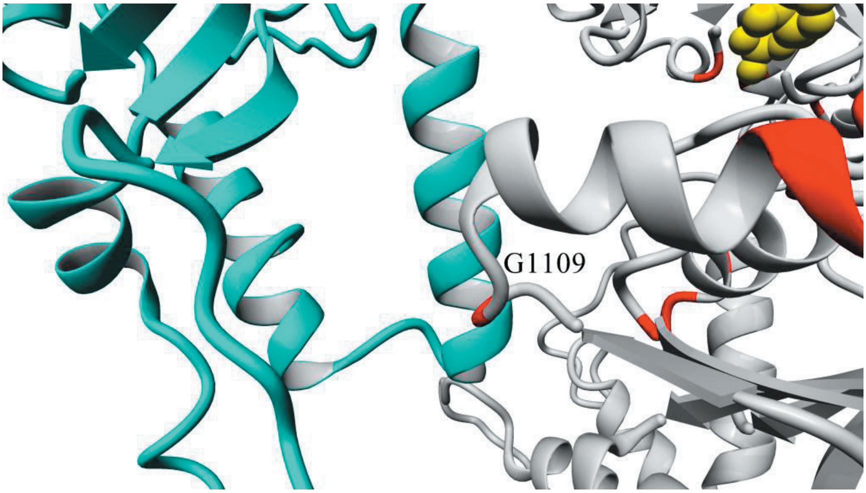


Figure SN12: Position of G1109 in the open structure with ATP. G1109 and its surrounding residues interact with residues 611-617 in the helix (cyan). The function of this interaction is not known but will be affected by the mutation..

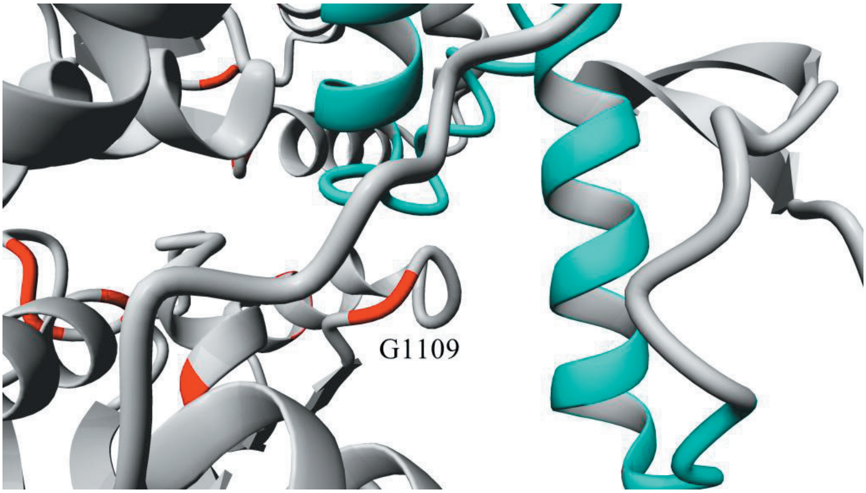


Figure SN13: Position of G1109 in the closed model. The loop that contains 1109 is still close to the helix with residues 611-617. However, the position of the helix (and the domain that contains this helix) has changed and as a result the interaction is less close.

Supplementary References

1. Weiss, K. et al. De Novo Mutations in CHD4, an ATP-Dependent Chromatin Remodeler Gene, Cause an Intellectual Disability Syndrome with Distinctive Dysmorphisms. *Am J Hum Genet* **99**, 934-941 (2016).
2. Sifrim, A. et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat Genet* **48**, 1060-5 (2016).
3. Krieger, E., Koraimann, G. & Vriend, G. Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. *Proteins* **47**, 393-402 (2002).
4. Vriend, G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 8, 52-6, 29 (1990).
5. Hauk, G., McKnight, J.N., Nodelman, I.M. & Bowman, G.D. The chromodomains of the Chd1 chromatin remodeler regulate DNA access to the ATPase motor. *Mol Cell* **39**, 711-23 (2010).
6. Yan, L., Wang, L., Tian, Y., Xia, X. & Chen, Z. Structure and regulation of the chromatin remodeller ISWI. *Nature* **540**, 466-469 (2016).
7. Xia, X., Liu, X., Li, T., Fang, X. & Chen, Z. Structure of chromatin remodeler Swi2/Snf2 in the resting state. *Nat Struct Mol Biol* **23**, 722-9 (2016).
8. Smith, C.L. & Peterson, C.L. A conserved Swi2/Snf2 ATPase motif couples ATP hydrolysis to chromatin remodeling. *Mol Cell Biol* **25**, 5880-92 (2005).
9. Farnung, L., Vos, S.M., Wigge, C. & Cramer, P. Nucleosome-Chd1 structure and implications for chromatin remodelling. *Nature* (2017).
10. Liu, X., Li, M., Xia, X., Li, X. & Chen, Z. Mechanism of chromatin remodelling revealed by the Snf2-nucleosome structure. *Nature* **544**, 440-445 (2017).

5

Chapter 5

A clustering of missense variants in the crucial chromatin modifier WDR5 defines a new neurodevelopmental disorder

Lot Snijders Blok*, Jolijn Verseput*, Dmitrijs Rots, Hanka Venselaar, A. Micheil Innes, Connie Stumpel, Katrin Õunap, Karit Reinson, Eleanor G. Seaby, Shane McKee, Barbara Burton, Katherine Kim, Johanna M. van Hagen, Quinten Waisfisz, Pascal Joset, Katharina Steindl, Anita Rauch, Dong Li, Elaine Zackai, Sarah Sheppard, Beth Keena, Hakon Hakonarson, Andreas Roos, Nicolai Kohlschmidt, Anna Cereda, Maria Iascone, Erika Rebessi, Kristin D. Kernohan, Philippe M. Campeau, Francisca Millan, Jesse A. Taylor, Birgitta Bernhard, Simon E. Fisher, Han G. Brunner, Tjitske Kleefstra

** These authors contributed equally*

Abstract

WDR5 is a broadly studied, highly conserved protein involved in a wide array of biological functions. Among these functions, WDR5 is a part of several protein complexes that affect gene regulation via post-translational modification of histones. We collected data from ten unrelated individuals with six different rare de novo missense variants in WDR5; one identical variant was found in four individuals, and another variant in two individuals. All ten individuals had neurodevelopmental disorders including speech/language delays (N=10), intellectual disability (N=8), epilepsy (N=6) and autism spectrum disorder (N=4). Additional phenotypic features included abnormal growth parameters (N=6), heart anomalies (N=2) and hearing loss (N=2). All six missense variants occurred in regions of the WDR5 locus that are known to be extremely intolerant for variation. Three-dimensional structures indicate that all the residues affected by these variants are located at the surface of one side of the WDR5 protein. It is predicted that five out of the six amino-acid substitutions disrupt interactions of WDR5 with RbBP5 and/or KMT2A/C, as part of the COMPASS family complexes. Thus, we define a new neurodevelopmental disorder associated with missense variants in WDR5 and a broad range of associated features including intellectual disability, speech/language impairments, epilepsy and autism spectrum disorders. This finding highlights the important role of COMPASS family proteins in neurodevelopmental disorders.

Introduction

WDR5 is a small highly conserved protein that is able to interact with a large number of other proteins[1,2]. As a core constituent of many different chromatin-related protein complexes, it controls crucial regulatory processes during development[3,4]. The indispensable function of WDR5 is illustrated by its high evolutionary conservation. Even very basic multicellular organisms such as *Trichoplax adhaerens* have a protein with around 90% similarity to the 334 amino acids of the human orthologue[4,5]. WDR5 is a member of the WD40 repeat family and has seven WD40 domains that each forms a propeller-like wing, resulting in a final barrel-shaped protein[6]. Using two binding sites on opposite sides of the protein, WDR5 can act as an adapter to link different proteins and form protein complexes. Since the protein is highly multifunctional and ubiquitously expressed[7] (data available from https://v19.proteinatlas.org/ENSG00000196363-WDR5), the unavailability of a well-functioning WDR5 could potentially impact myriad downstream processes.

Most of the protein complexes that WDR5 participates in affect gene regulation via post-translational modification of histones. The MLL/SET complexes (also known as COMPASS family complexes) catalyse histone 3 lysine 4 (H3K4) di- and trimethylation[8,9] and the NSL and ATAC complexes are involved in histone acetylation[10,11]. WDR5 can also be part of an embryonic stem cell-specific NuRD complex that combines nucleosome sliding capacities with histone deacetylation activity[12]. In addition to interactions with other proteins, WDR5 is also able to bind to >1000 different endogenous RNAs (including long non-coding RNAs), and binding to certain long non-coding RNAs can be crucial for WDR5 stability and function in cells[13]. A recent study linked WDR5, as part of the COMPASS complex, to a newly discovered genetic compensation mechanism: nonsense-induced transcriptional compensation[14]. In short, this mechanism is triggered by a truncating variant in a gene, and leads to the expression of related genes, thereby compensating for the effects of a deleterious variant[15]. One of the most well-studied aspects of WDR5 function is its role in embryonic stem cell (ESC) self-renewal and maintenance of a pluripotent state[16,17]. Recently, a direct interaction between the p53 protein and WDR5 has been uncovered, in which mouse ESC stem cell fate (the differentiation into neuroectoderm or mesoderm) is regulated in a p53-dependent manner[18]. Thus, WDR5 has already been implicated in multiple different molecular pathways and mechanisms, and this list is growing steadily.

While the biological functions of the WDR5 protein have been studied from numerous angles, there is still little known about the impact of rare germline variants in the gene that encodes it. Using data from large-scale sequencing resources, it is clear that *WDR5* is extremely intolerant for loss-of-function variation: in both the gnomAD database[19] (version 2.1.1; containing sequencing data of 141,456 individuals) and the TOPMED database[20] (containing sequencing data of 62,784 individuals) not a single truncating variant in *WDR5* is reported. Similarly, *WDR5* is also depleted for missense variants. Therefore, the initial finding of a *de novo* missense variant (p.(Thr208Met)) in *WDR5* in a child with a developmental speech disorder[21], prompted us to investigate the effects and possible pathogenicity of rare germline variants in this gene. We collected clinical information on this proband and nine additional individuals with rare *de novo* germline variants in *WDR5*, collated from several clinical exome or genome sequencing studies. We used a range of *in silico* tools and analyses of variants using three-dimensional structures in order to evaluate the likely consequences of the different variants found.

Methods

Study participants and consent

Individuals with *WDR5* variants were identified via matchmaking using GeneMatcher[22], the Dutch genetic diagnostic variant classification database (VKGL database)[23-25], ClinVar[26] and denovo-db[27]. Clinical data and details on variants were collected in a Castor EDC database[28]. Informed consent to share and publish these data was given by all individuals or their legal guardian.

Next generation sequencing and in silico variant analyses

Details on next generation sequencing methods used to identify the *WDR5* variants found in all individuals are included in table S1. Variants were analysed using Alamut Visual 2.10. Conservation was studied using a Clustal[29] alignment of WDR5 amino acid sequences extracted from Uniprot (human, mouse and C. elegans)[30]. To assess the likelihood of pathogenicity, the prediction programs SIFT[31], PolyPhen-2[32] and CADD v1.4[33] were used.

Three-dimensional (3D) protein modelling

The effects of the identified variants on the WDR5 protein and its interaction with other proteins in the COMPASS family complexes were analyzed using YASARA View[34] with FoldX v4.0 plugin[35]. For the WDR5 structure, PDB file 2GNQ was used. PDB files 6KIV and 6KIW[36] were used for the analysis of the core MLL1 and MLL3 complexes, respectively; the 6UH5 file[37] of the yeast COMPASS model was used for the comparison with the human MLL1 complex. To optimize the position of amino-acid sidechains, all the PDB files that were used were corrected by the FoldX repair function using default settings. Different protein structures were aligned with SHEBA procedure[38], as implemented in YASARA.

Results

Identification and clinical characterization of individuals with de novo WDR5 variants

We collected clinical and molecular data on ten unrelated individuals with a *de novo* missense variant in *WDR5*. Six different missense variants were reported in these ten individuals: p.(Ala169Pro), p.(Arg196Cys), p.(Ala201Val), p.(Thr208Met), p.(Asp213Asn) and p.(Lys245Arg). The p.(Thr208Met) variant was found in four unrelated individuals, and the p.(Arg196Cys) variant in two unrelated individuals.

All individuals had neurodevelopmental disorders with a spectrum of overlapping associated features (Figure 1 and table S1). Intellectual disability (ID) was present in eight out of ten individuals, with a severity ranging from moderate ID (IQ 35-50, five individuals) to mild ID (IQ 50-70, three individuals). Of the two remaining individuals, one individual (individual 6) had a borderline level of intellectual functioning (IQ 70-85) and another individual (individual 4) had no intellectual disability. Interestingly, in all ten individuals speech delays were reported. Two individuals (aged 6y and 17y) were non-verbal, and three other individuals had a developmental language disorder diagnosis (mixed expressive/receptive language disorder in two individuals, and expressive language disorder in one). One of these latter three individuals was also reported with verbal dyspraxia. In addition, five individuals were reported to have nasal speech, and one individual was reported with persistent stuttering.

All but one of the individuals with *WDR5* variants had motor development delays, with the age of first steps ranging from 12-40 months. Hypotonia was reported in six individuals. Concerning the behavioural phenotype, four individuals had an autism spectrum disorder (ASD) diagnosis, and two individuals were diagnosed with Attention Deficit Hyperactivity Disorder (ADHD).

Six out of the ten individuals were diagnosed with epilepsy, with different forms of presentations varying from absence seizures to refractory generalized myoclonic epilepsy. In four of these six individuals, the disorder was only present in childhood and medication could be discontinued at a later age. A brain MRI scan was performed for seven individuals, showing different types of abnormalities in three of these scans: mild ventricular dilatation with thinning of the posterior corpus callosum in individual 3, subtle grey matter heterotopias in individual 5, and periventricular gliosis in individual 9.

Individuals with *WDR5* variants showed divergent growth parameters. One individual had macrocephaly (head circumference ≥+2 SD) and another had microcephaly (≤-2 SD). Two individuals had tall stature (≥+2 SD). A weight of +2 SD or more (for height) was seen in four individuals in total, including all three adult individuals in our study. One individual had a low weight (≤-2 SD). No clear correlation between height, weight and head circumference was observed (table S1), with the exception of one individual (individual 6) which showed a striking pattern of progressive overgrowth (height +3 SD, weight +5 SD, head circumference +4 SD at 25 years of age) and one individual (individual 2) with a milder generalized overgrowth phenotype (height +3 SD, weight +1.5 SD, head circumference +1.5 SD).

Different abnormalities of the skeleton and limbs were present in a subset of individuals. Bilateral clubfeet were reported in one individual and another individual had hemihypertrophy of the left leg. One individual was reported with a hemivertebra of L5 and kyphosis (possibly secondary to the hemivertebra), and another individual had scoliosis. Two individuals had single palmar creases. In two individuals, heart abnormalities were reported: cardiac arrhythmias and decompensated heart failure requiring surgery in one individual, and left ventricular non-compaction cardiomyopathy in another individual. Three individuals were reported with frequent infections of the ears and/or airways. One individual had strabismus, another individual had amblyopia and hyperopia with astigmatism.

Significant facial dysmorphisms were noted in only a subset of individuals. When comparing facial features of seven individuals in our cohort, overlapping facial features included a bulbous nasal tip, low-set, posteriorly rotated and/or dysplastic ears, ptosis and thin lip vermilion (Figure 2). Two individuals in our cohort, individual 4 and individual 10, had distinct facial features compared to the others, with severe micrognathia (requiring tracheostomy in one), a small mouth and prominent down-slanting palpebral fissures. These two individuals both had conductive hearing loss too, a feature not reported in any of the other individuals. Clinical features reported in individuals in our cohort are described in more detail in table S1.

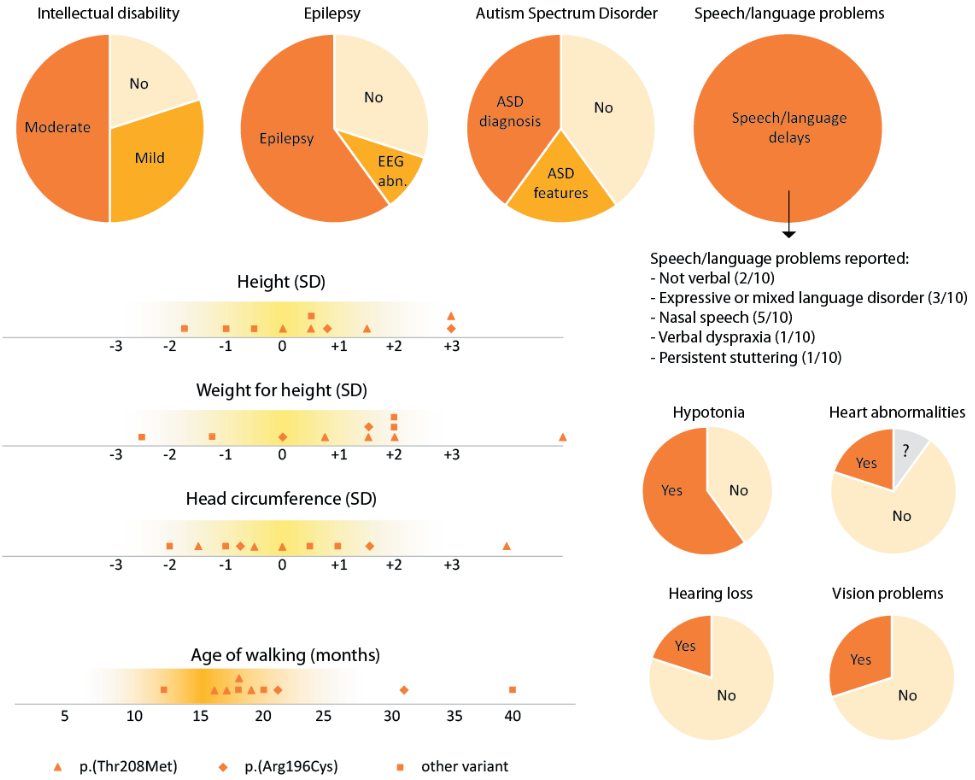


Figure 1: Clinical features reported in individuals with *WDR5* variants

Graphical overview of clinical features reported in ten individuals with WDR5 missense variants. Growth parameters are shown as standard deviations to the mean for a certain age. All graphs include data for all ten individuals (N=10). 'EEG abn.' = EEG abnormalities.

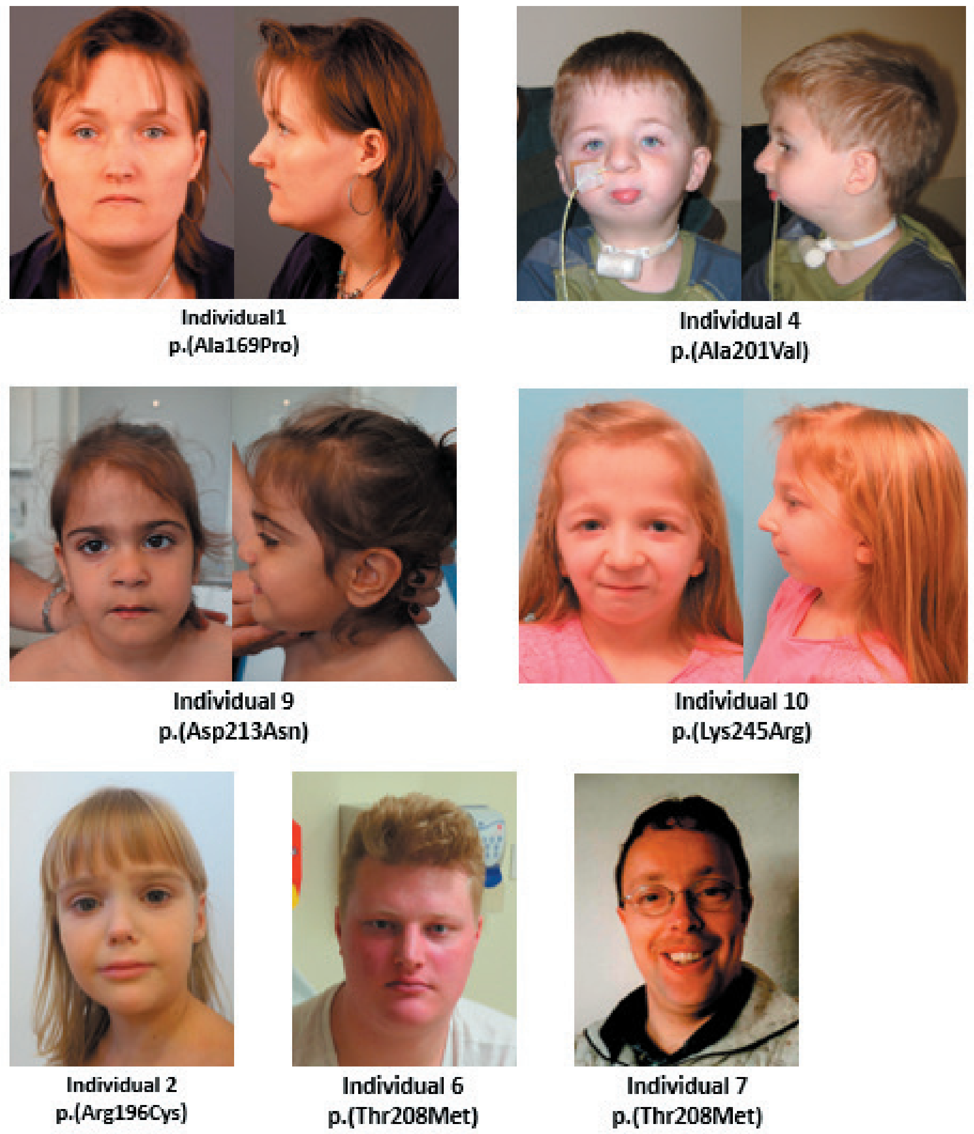


Figure 2: Facial features in individuals with six different *WDR5* variants

Facial images of seven individuals with a heterozygous WDR5 variant. Several overlapping facial features are seen, such as a bulbous nasal tip (individual 2, 4, 9), low-set, posteriorly rotated and/or dysplastic ears (individual 2, 4, 7, 9), ptosis (individual 10) and thin upper lip vermilion (individual 4, 9, 10). In addition to this, individual 4 and individual 10 have severe micrognathia, a small mouth and down-slanting palpebral fissures.

In silico analysis of variants

The missense variants in our cohort are all located within or flanking the fourth and the fifth WD40 domain of WDR5, with each WD40 domain encoding one 'propeller' of the three-dimensional WDR5 structure (Figure 3A and 3B). All missense variants in our study were absent from the gnomAD database[19]. We used *in silico* prediction programs to evaluate likely pathogenicity for the six different missense variants. The resulting scores are summarised in Table 1. All CADD scores were above 22, while SIFT and Polyphen-2 predicted three out of the six variants to be pathogenic: p.(Ala169Pro), p.(Arg196Cys) and p.(Thr208Met).

Three-dimensional structure analysis

Using three-dimensional structure analysis, we determined that the residues affected by all the identified missense variants cluster together on one side of the WDR5 surface, suggesting that these variants may disrupt specific interactions with other proteins (Figure 3B). Different surfaces of WDR5 are known to mediate interactions with different proteins, thereby linking them together. In previous co-precipitation experiments with WDR5 and short fragments of proteins from complexes in which WDR5 is involved, two distinct binding sites were identified: the 'WDR5-interacting' (WIN) site[39-41] and the 'WDR5-binding motif' (WBM) site[41,42], located on opposite sides (often referred to as 'top' and 'bottom', respectively) of the protein. The variants found in our study are not located in the vicinity of these well-studied binding locations.

However, recently published cryo-electron microscopy 3D structures of the core MLL1 and MLL3 complexes revealed a region located between the WIN and WBM binding sites, that is involved in the interaction with RbBP5 and/or histone-lysine methylase (KMT) enzymes in these complexes[36]. Five out of the six variants that we identified map within this RbBP5/KMT interaction region. Based on the 3D structure analysis of the MLL1 and MLL3 complexes, p.(Ala169Pro) and p.(Asp213Asn) are predicted to affect the WDR5 interaction with KMT enzymes, and p.(Ala201Val) and p.(Thr208Met) are predicted to affect the interaction with the RbBP5 enzymes, while p.(Arg196Cys) most likely influences the interaction with both enzymes (Figure 3C-F). The effects of the p.(Lys245Arg) variant cannot be predicted using the currently available 3D structures. A detailed description of the predicted effects of all variants, from the perspective of the structural modelling analyses, is provided in Supplementary Note 1.

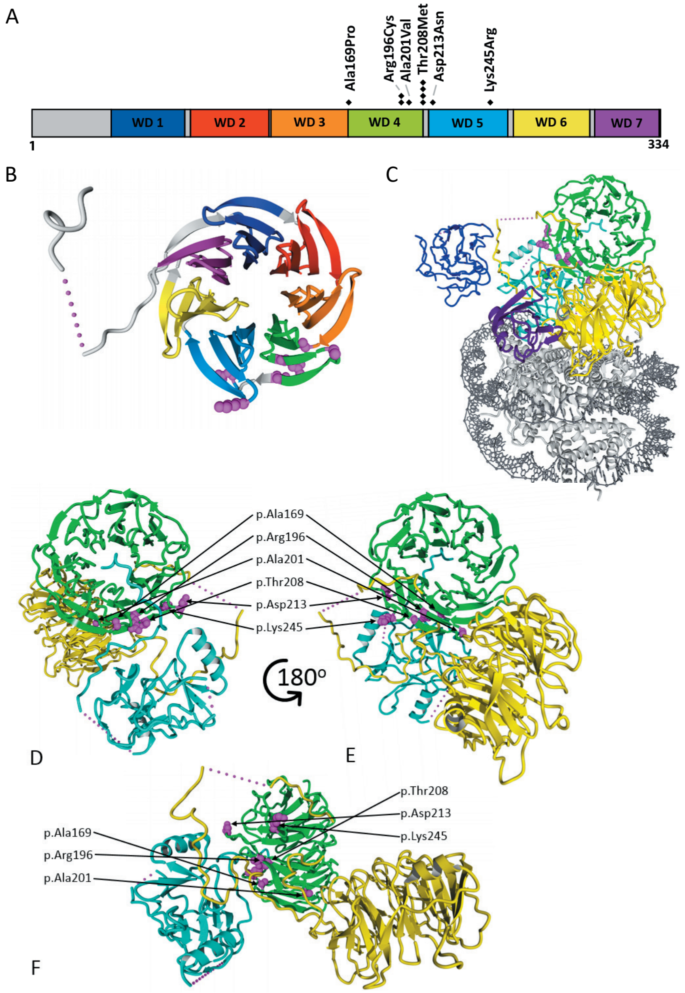


Figure 3: Overview of variants in linear and threedimensional protein structures

a) Linear structure of WDR5 protein (334 amino acids) with the seven different WD40-domains and all identified missense variants shown; in total six different missense variants were found in ten individuals, as one variant (p.(Thr208Met)) was found in four unrelated individuals and another variant (p.(Arg196Cys)) was found in two unrelated individuals.

b) Three-dimensional visualization of WDR5 (PDB 2GNQ), locations of the amino acids involved in missense variants are shown with magenta balls. Colours of the different domains match with the colours used in panel A.

c) WDR5 (green) is shown as part of the core MLL1 complex, with RbBP5 (yellow), ASH2L (blue), DPY30 (purple) and KMT2A (cyan) (PDB:6KIV). The nucleosome is shown in grey. The locations of the amino acids affected in patients identified in this study are shown with balls (magenta).

d) E) F) WDR5 (green, p.33-332) is shown together with RbBP5 (yellow, p.1-380) and KMT2A (cyan, p.3764-3969), as part of the core MLL1 complex (PDB:6KIV). The locations of the amino acids affected in patients identified in this study are shown with balls (magenta), from three different angles: facing the WIN site (D), the WBM site (E) and a side between WIM and WBM sites (F).

Discussion

We identified rare *de novo* missense variants in *WDR5*, a gene that encodes a core chromatin regulator and is extremely intolerant for variation in the population. We clarify the molecular and phenotypic consequences of these rare variants which define a novel Mendelian disorder. By using the GeneMatcher database and international collaborations, we collected clinical data on a cohort of ten individuals with a *de novo* variant in *WDR5*. We studied all variants in three-dimensional conformations of WDR5 in interaction with COMPASS complex family subunits, and showed that the missense variants are predicted to affect important binding sites of WDR5. By combining data from these clinical and *in silico* approaches, we can confidently link *WDR5* to a neurodevelopmental disorder with a broad spectrum of associated features, further confirming that proteins in the COMPASS complex family are important contributors to neurodevelopment.

All individuals in our cohort had a heterozygous *de novo* missense variant in *WDR5*: six different missense variants were found, affecting amino acids on one surface of the WDR5 protein. The recurrence of the p.(Thr208Met) and p.(Arg196Cys) variant in four and two unrelated individuals, respectively, points to the presence of 'hotspots' for recurrent variants in *WDR5*. Three different missense variants in our cohort, p.(Ala169Pro), p.(Arg196Cys) and p.(Thr208Met), were found to locate at adjacent amino positions in the three-dimensional structure of WDR5. The fact that all identified variants are missense variants, the recurrence of specific variants in unrelated individuals and the spatial clustering of the variants on the protein surface of WDR5, all suggest that specific pathogenic mechanisms that might be at play are not just loss-of-function or haploinsufficiency.

To the best of our knowledge, truncating variants (e.g. frameshift or nonsense variants) in *WDR5* have not been identified so far: not in our cohort or any disease cohort in literature, nor in healthy individuals (e.g. in the gnomAD or TOPMED database). According to sequencing data from the gnomAD database, *WDR5* is extremely intolerant for both missense and loss-of-function variation. The gene has a LOEUF score of 0.124, which is well within the first decile of most highly constrained genes against loss-of-function[19]. In contrast to the absence of truncating variants, heterozygous chromosomal microdeletions encompassing the whole *WDR5* gene have been reported; the Decipher database lists eleven heterozygous deletions that include WDR5[43]. This means that haploinsufficiency for *WDR5* is compatible with life, but it is unclear how the loss of WDR5 contributes to specific phenotypes found in individuals with these deletions, as all deletions are larger than 3 Mb and encompass many other genes as well.

Analysis of the missense variants in a three-dimensional structure of WDR5 in the context of the MLL1 or MLL3 complex, revealed that all but one were located at amino-acid positions that are important for binding of WDR5 with other proteins of these complexes. Interestingly, analysis of the intolerance landscape of the *WDR5* gene in the three-dimensional structure of the encoded protein shows that WDR5 is generally intolerant to missense variants, but that residues interacting with other proteins are most intolerant for normal variation (Figure S1). For five out of six missense variants, we predict that the corresponding amino-acid change disrupts the interaction of WDR5 with the MLL1/MLL3 complex subunits RbBP5 and with KMT2A/C simultaneously or separately. WDR5 is a crucial core protein within the COMPASS complex family; it is essential for complex assembly and activity[44,45]. Based on 3D analysis of the MLL1 and MLL3 complexes, it seems that differently composed COMPASS complexes make use of different interaction surfaces of WDR5. Some variants might therefore disrupt interactions in only one specific complex. As the core function of the WDR5 protein seems to be to act as an 'adapter' protein and form links between different molecules, disruption of protein-protein interactions within the complex might have important effects on complex activity.

However, it is important to take into account that the extensive and detailed three-dimensional structures that we used for these analyses are only available for the MLL1/3 complex, and not for all other complexes and interactions in which WDR5 is involved. Therefore, it remains unclear whether the predicted disruptive effects on WDR5 interactions are specific to those with RbBP5/KMT and MLL1/3, or whether interactions with other molecules might also be disturbed. For the p.(Lys245Arg) variant, we were not able to predict a possible pathogenic mechanism analysing available three-dimensional structures. One hypothesis could be that the variant affects a so far uncharacterized interaction site with RbBP5 or KMT2A/C, as a comparison of available three-dimensional structures between human MLL1 and yeast COMPASS complex suggest even more extensive interaction surfaces between WDR5 and histone methylases (Figure S2). Another hypothesis is that the p.(Lys245Arg) variant affects the interaction with other molecules that are not involved in the MLL1/MLL3 complex. Lastly, although the *de novo* occurrence of the p.(Lys245Arg) and the phenotypic similarity to the rest of the cohort suggest pathogenicity, the possibility that this is a benign variant without any effect on WDR5 function cannot be excluded.

The individuals in our study showed a broad range of clinical features: neurodevelopmental phenotypes with several additional abnormalities. All individuals had developmental delays, with mild or moderate intellectual disability being present in the majority. Speech and language problems were a prominent feature, as was epilepsy. Abnormal height, weight and head circumferences were frequently seen, and two individuals had a remarkable distinct facial phenotype with severe micrognathia, a small mouth and down-slanted palpebral fissures. When comparing the phenotypes corresponding to different variants in our study, no clear genotype-phenotype correlation was established. Even in four individuals with the exact same p.(Thr208Met) variant a different clinical presentation was seen: e.g. borderline vs. moderate intellectual disability, normal growth parameters vs. a generalized overgrowth phenotype, and variability in the presence of additional phenotypic features. Altogether, *WDR5*-associated disorder can be characterized as a neurodevelopmental disorder with variable expressivity of associated features. This is in line with other disorders caused by variants in genes encoding COMPASS complex family subunits, such as Kabuki syndrome (caused by variants in *KMT2D* or *KDM6A*)[46], Kleefstra syndrome type 2 (caused by variants in *KMT2C*)[47] and the neurodevelopmental disorder associated with *SETD1A* loss-of-function variants[48-50]. In all these COMPASS complex-associated disorders a variable spectrum of associated features can be present in varying degrees of severity, including intellectual disability, speech and language delays, autism spectrum disorders, epilepsy and abnormal growth parameters[46-50].

This study represents the first characterization of a Mendelian disorder associated with germline variants in WDR5, and was initiated after the report of a *de novo* WDR5 variant in a child with a speech disorder[21]. It is worth mentioning that one additional *de novo* variant in *WDR5* is reported in the literature: a p.(Lys7Gln) variant, found in a child with a conotruncal heart defect with a right aortic arch[51]. This missense variant is located in the N-terminal tail of WDR5, an intrinsically disordered region of the protein (not available for three-dimensional protein structure analysis), which is not involved in the beta-propeller structure of WDR5, and has been shown to be dispensable for COMPASS complex assembly[52]. A study in *Xenopus tropicalis* shows that this p.(Lys7Gln) variant might interfere with the ability of WDR5 to localize to the bases of left-right organizer cilia, independent from the H3K4-methylation-related functions of WDR5[53]. As p.(Lys7Gln) is located in a different region of the WDR5 protein compared to the variants found in our study, and since complete phenotypic details are not available for this individual, it is currently unclear whether this reported individual has the *WDR5*-associated neurodevelopmental disorder presented in this study, or whether this specific variant gives rise to a different disorder with different pathogenic mechanisms.

All in all, with this study WDR5 can be added to the list of genes robustly associated with autosomal dominant neurodevelopmental disorders. All variants found so far are missense variants, and although they are not always predicted to be pathogenic by the commonly used prediction programs, our three-dimensional protein structure analysis showed that it is very likely that five out of the six found missense variants disturb important protein-protein interactions. Future studies are needed to confirm that these *in silico* observations in 3D structures indeed affect protein-protein interactions between WDR5 and RbBP5 and KMT enzymes, and what the downstream consequences are on H3K4 methyltransferase activity. But taken together, with our study we implicate *WDR5* variants in an autosomal dominant disorder associated with intellectual disability, speech and language delays, epilepsy and autism spectrum disorder, thereby highlighting the role of COMPASS complex family protein disruptions in neurodevelopmental disorders.

Acknowledgements

We thank all included individuals and their families for their contribution to this research project.

Funding was provided by the Netherlands Organization for Scientific Research (NWO) Gravitation Grant 24.001.006 to the Language in Interaction Consortium (L.S.B., S.E.F., and H.G.B.), the Max Planck Society (S.E.F.) and the Netherlands Organization for Health Research and Development (ZonMw grant 91718310 to T.K.). The research of A.C., M.I. and E.R. was supported by PROGETTEO GENE (GENE = Genomic analysis Evaluation Network) founded by PROGETTI DI INNOVAZIONE IN AMBITO SANITARIO E SOCIO SANITARIO (BANDO EX DECRETO N. 2713 DEL 28/02/2018). A.R. and N.K. acknowledge funding from the European Regional Development Fund (ERDF). K.Õ. and K.R. were supported by the Estonian Research Council grants PUT355 and PRG471. The Broad Center for Mendelian Genomics (UM1 HG008900) is funded by the National Human Genome Research Institute with supplemental funding provided by the National Heart, Lung, and Blood Institute under the Trans-Omics for Precision Medicine (TOPMed) program and the National Eye Institute.

Individuals 3 and 4 in this study were part of the DDD study cohort. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003]. This study makes use of DECIPHER (http://decipher.sanger.ac.uk), which is funded by Wellcome. See Nature PMID: 25533962 or www.ddduk.org/access.html for full acknowledgement.

References

1. Trievel RC, Shilatifard A. WDR5, a complexed protein. *Nat Struct Mol Biol.* 2009;16(7):678-680.
2. van Nuland R, Smits AH, Pallaki P, Jansen PW, Vermeulen M, Timmers HT. Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Mol Cell Biol.* 2013;33(10):2067-2077.
3. Wysocka J, Swigut T, Milne TA, et al. WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell.* 2005;121(6):859-872.
4. Guarnaccia AD, Tansey WP. Moonlighting with WDR5: A Cellular Multitasker. *J Clin Med.* 2018;7(2).
5. Copley RR. The Unicellular Ancestry of Groucho-Mediated Repression and the Origins of Metazoan Transcription Factors. *Genome Biol Evol.* 2016;8(6):1859-1867.
6. Xu C, Min J. Structure and function of WD40 domain proteins. *Protein Cell.* 2011;2(3):202-214.
7. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. *Science.* 2015;347(6220):1260419.
8. Roguev A, Schaft D, Shevchenko A, et al. The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *EMBO J.* 2001;20(24):7137-7148.
9. Miller T, Krogan NJ, Dover J, et al. COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc Natl Acad Sci U S A.* 2001;98(23):12902-12907.
10. Cai Y, Jin J, Swanson SK, et al. Subunit composition and substrate specificity of a MOF-containing histone acetyltransferase distinct from the male-specific lethal (MSL) complex. *J Biol Chem.* 2010;285(7):4268-4272.
11. Suganuma T, Gutierrez JL, Li B, et al. ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding. *Nat Struct Mol Biol.* 2008;15(4):364-372.
12. Ee LS, McCannell KN, Tang Y, et al. An Embryonic Stem Cell-Specific NuRD Complex Functions through Interaction with WDR5. *Stem Cell Reports.* 2017;8(6):1488-1496.
13. Yang YW, Flynn RA, Chen Y, et al. Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *Elife.* 2014;3:e02046.
14. Ma Z, Zhu P, Shi H, et al. PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. *Nature.* 2019;568(7751):259-263.
15. Wilkinson MF. Genetic paradox explained by nonsense. *Nature.* 2019;568(7751):179-180.
16. Ang YS, Tsai SY, Lee DF, et al. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell.* 2011;145(2):183-197.
17. Li X, Li L, Pandey R, et al. The histone acetyltransferase MOF is a key regulator of the embryonic stem cell core transcriptional network. *Cell Stem Cell.* 2012;11(2):163-178.
18. Li Q, Mao F, Zhou B, et al. p53 Integrates Temporal WDR5 Inputs during Neuroectoderm and Mesoderm Differentiation of Mouse Embryonic Stem Cells. *Cell Rep.* 2020;30(2):465-480 e466.
19. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443.
20. NHLBI UoMa. The NHLBI Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Program. *BRAVO variant browser.* 2018.
21. Eising E, Carrion-Castillo A, Vino A, et al. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol Psychiatry.* 2019;24(7):1065-1078.
22. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat.* 2015;36(10):928-930.
23. van der Velde KJ, Imhann F, Charbon B, et al. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics.* 2019;35(6):1076-1078.
24. Swertz MA, Dijkstra M, Adamusiak T, et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics.* 2010;11 Suppl 12:S12.
25. Swertz MA, Jansen RC. Beyond standardization: dynamic software infrastructures for systems biology. *Nat Rev Genet.* 2007;8(3):235-243.
26. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062-d1067.
27. Scholte E, Van Duijn E, Dijkxhoorn Y, Noens I, Van Berckelaer-Onnes I. Vineland screener 0–6 years: manual of the Dutch adaptation. *PITS, Leiden.* 2008.
28. Castor EDC. Castor Electronic Data Capture. https://castoredc.com. Published 2019. Accessed August 28, 2019.

29. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7:539. doi:10.1038/msb.2011.75. Accessed 2011.
30. Consortium TU. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2018;47(D1):D506-D515.
31. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc.* 2016;11(1):1-9.
32. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7(4):248-249.
33. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-315.
34. Krieger E, Vriend G. YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics.* 2014;30(20):2981-2982.
35. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic acids research.* 2005;33(Web Server issue):W382-W388.
36. Xue H, Yao T, Cao M, et al. Structural basis of nucleosome recognition and modification by MLL methyltransferases. *Nature.* 2019;573(7774):445-449.
37. Hsu PL, Shi H, Leonen C, Kang J, Chatterjee C, Zheng N. Structural Basis of H2B Ubiquitination-Dependent H3K4 Methylation by COMPASS. *Mol Cell.* 2019;76(5):712-723.e714.
38. Jung J, Lee B. Protein structure alignment using environmental profiles. *Protein Engineering, Design and Selection.* 2000;13(8):535-543.
39. Patel A, Dharmarajan V, Cosgrove MS. Structure of WDR5 bound to mixed lineage leukemia protein-1 peptide. *J Biol Chem.* 2008;283(47):32158-32161.
40. Patel A, Vought VE, Dharmarajan V, Cosgrove MS. A conserved arginine-containing motif crucial for the assembly and enzymatic activity of the mixed lineage leukemia protein-1 core complex. *J Biol Chem.* 2008;283(47):32162-32175.
41. Dias J, Van Nguyen N, Georgiev P, et al. Structural analysis of the KANSL1/WDR5/KANSL2 complex reveals that WDR5 is required for efficient assembly and chromatin targeting of the NSL complex. *Genes Dev.* 2014;28(9):929-942.
42. Odho Z, Southall SM, Wilson JR. Characterization of a novel WDR5-binding site that recruits RbBP5 through a conserved motif to enhance methylation of histone H3 lysine 4 by mixed lineage leukemia protein-1. *J Biol Chem.* 2010;285(43):32967-32976.
43. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet.* 2009;84(4):524-533.
44. Dou Y, Milne TA, Ruthenburg AJ, et al. Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nat Struct Mol Biol.* 2006;13(8):713-719.
45. Steward MM, Lee JS, O'Donovan A, Wyatt M, Bernstein BE, Shilatifard A. Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nat Struct Mol Biol.* 2006;13(9):852-854.
46. Adam MP, Hudgins L, Hannibal M. Kabuki Syndrome. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews((R)).* Seattle (WA)1993.
47. Koemans TS, Kleefstra T, Chubak MC, et al. Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder. *PLoS Genet.* 2017;13(10):e1006864.
48. Yu X, Yang L, Li J, et al. De Novo and Inherited SETD1A Variants in Early-onset Epilepsy. *Neurosci Bull.* 2019;35(6):1045-1057.
49. Singh T, Kurki MI, Curtis D, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci.* 2016;19(4):571-577.
50. Kummeling J, Stremmelaar DE, Raun N, et al. Characterization of SETD1A haploinsufficiency in humans and Drosophila defines a novel neurodevelopmental syndrome. *Mol Psychiatry.* 2020.
51. Zaidi S, Choi M, Wakimoto H, et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature.* 2013;498(7453):220-223.
52. Schuetz A, Allali-Hassani A, Martin F, et al. Structural basis for molecular recognition and presentation of histone H3 by WDR5. *EMBO J.* 2006;25(18):4245-4252.
53. Kulkarni SS, Khokha MK. WDR5 regulates left-right patterning via chromatin-dependent and -independent functions. *Development.* 2018;145(23).

Supplementary Information

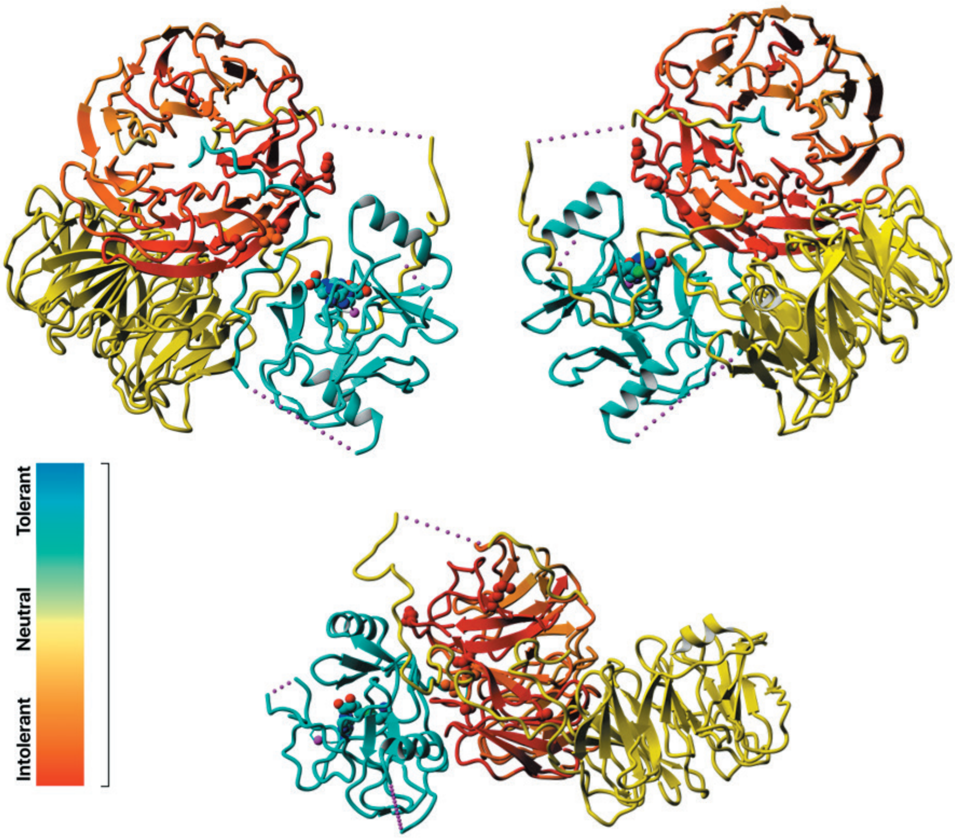


Figure S1: MetaDome intolerance visualization of WDR5

WDR5 is coloured in line with the MetaDome tolerance scale shown. RbBP5 is shown in yellow and KMT2A in cyan (PDB:6KIV). As can be seen in this figure, WDR5 is generally intolerant to missense variants, but WDR5 amino acids that are known to interact with other proteins are most intolerant (darker red).

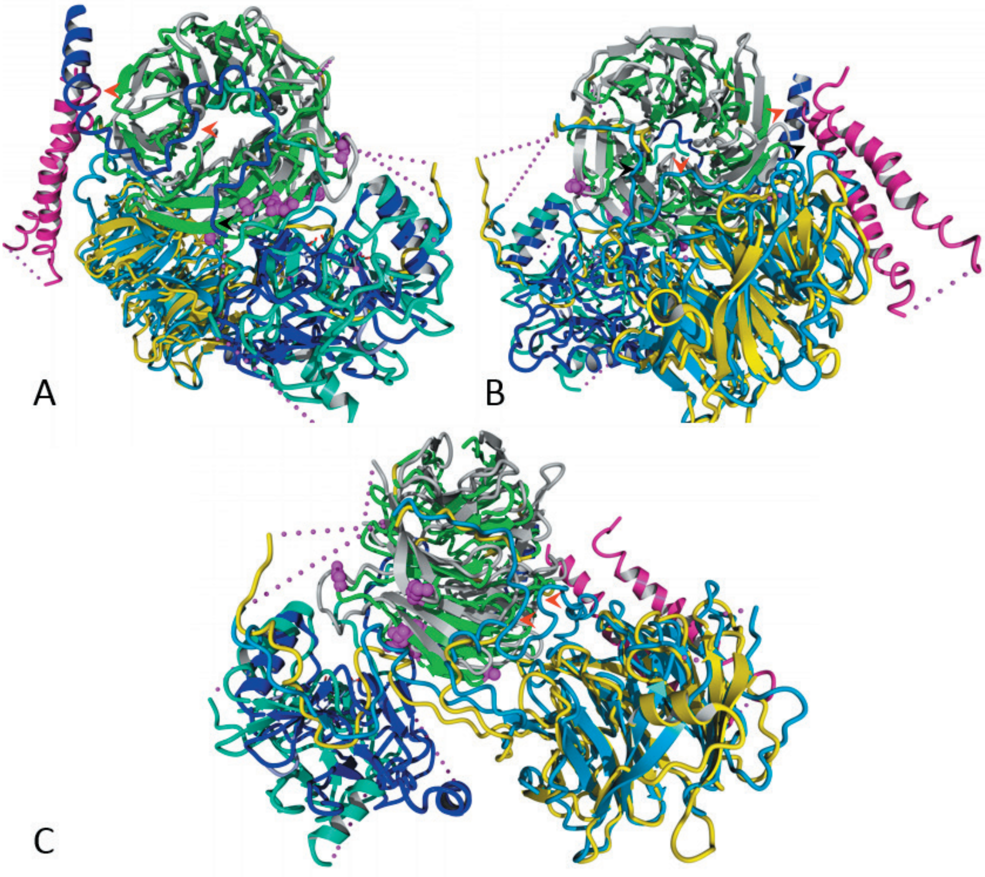


Figure S2: Comparison of the core human MLL1 with the yeast COMPASS complexes

The alignment of human WDR5 in complex with RbBP5 and KMT2A/MLL1 from the core MLL1 complex (PDB:6KIV) with homologues of the yeast COMPASS complex (PDB:6UH5) is shown: WDR5 (green, p.33-332) with its homologue Swd3 (grey, p.16-326); RbBP5 (yellow, p.1-380) with its homologue Swd1 (light blue, p.1-435); KMT2A/MLL1 (cyan, p.3764-3969) with yeast homologue Set1c (dark blue, p.819-999). Additionally, yeast Spp1 (purple) is shown. The Spp1 homologue is not present in human COMPASS family complexes. The locations of the amino acids that are affect in patients identified in this study are shown with balls (magenta). Three different angles are shown: WDR5 faced from the WIN site (A), from the WBM site (B), and from the side between WIN and WBM (C).

The human core COMPASS/COMPASS family complexes (eg., MLL1) are highly conserved and have a structure similar to the yeast COMPASS complex. Because the yeast COMPASS complex proteins in the 3D model are more complete, substantially more extensive interaction of the RbBP5 and KMT2A/MLL1 homologues with WDR5 homologues can be observed (red arrows). Additionally, another interaction site of the WDR5 homologue is observed with a Spp1 protein

These 3D modelling data, in addition to the high conservation level and low tolerance to the missense and LoF variants in the general population, suggest that also human WDR5 may have significantly more extensive interaction surfaces within COMPASS family complexes and other chromatin-remodelling complexes.

Supplementary Note 1: Detailed description and visualization of the predicted effect of identified WDR5 variants

p.(Ala169Pro)

Wild type residue role	Effect of the residue substitution
Ala169 is located in a turn from the third to fourth WDR5 beta-propeller. Despite the fact that the Ala169 is located in close proximity to the KMT2A/MLL1, and KMT2C/MLL3 interaction sites, it does not directly interact with the KMT enzymes.	Change from the alanine to a larger proline at this position is predicted to result in a local backbone change, because of the rigid sidechain of the proline. This change is predicted to disturb the flexibility and local structure of WDR5, which will disrupt the binding to the KMT enzymes.

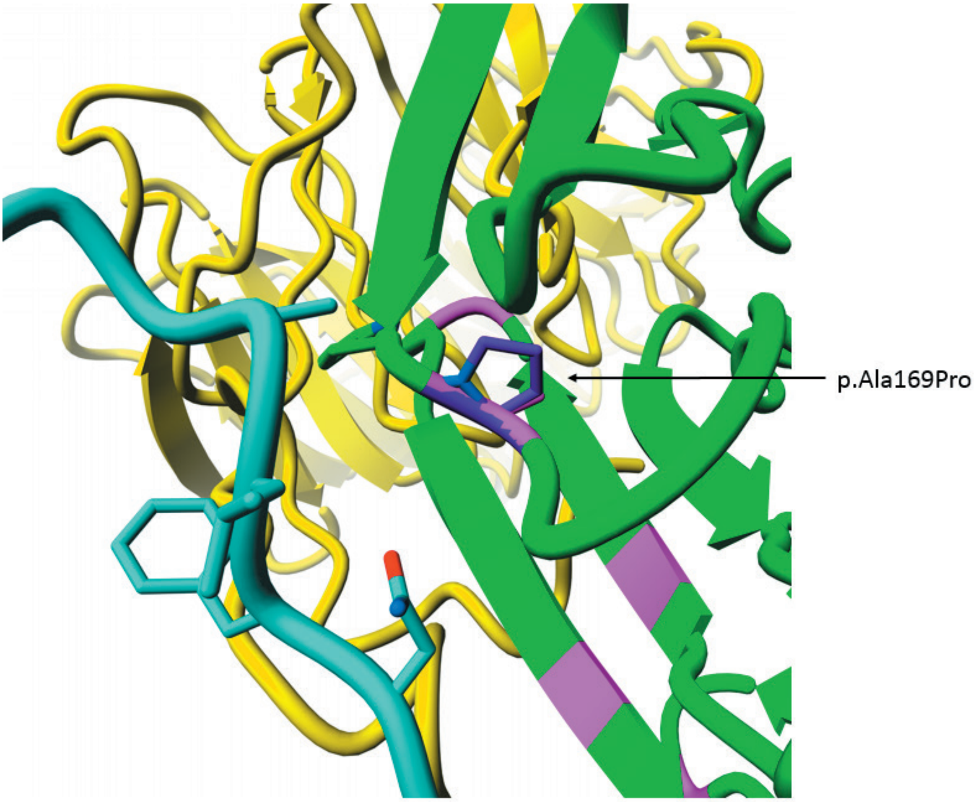


Figure S1: WDR5 (green) interaction with KMT2A/MLL1 (cyan) and RbBP5 (yellow), are shown from the core MLL1 complex (PDB:6KIV). The mutated aminoacid and nearby aminoacids are shown with sticks. The wild type alanine at the position p.169 is colored in magenta and the mutated proline in purple.

p.(Arg196Cys)

Wild type residue role	Effect of the residue substitution
Arg196 is located on the WDR5 lateral surface for interaction with RbBP5 and KMT2A enzymes. Arg196 interacts with Asn3779 in the KMT2A protein and Phe332 in the C-term tail of RbBBP5 but has no visible interactions with the KMT2C protein.	Cysteine is a much smaller residue and does not have a charge. Therefore, a change from the arginine to cysteine at this position would result in a loss of the hydrogen-bond with Asn3779 in KMT2A, as well resulting in an empty pocket between the WDR5, KMT2A and RbBBP5 interaction surfaces, which would lead to a loss of packing interactions and disruption of the interactions between the proteins.

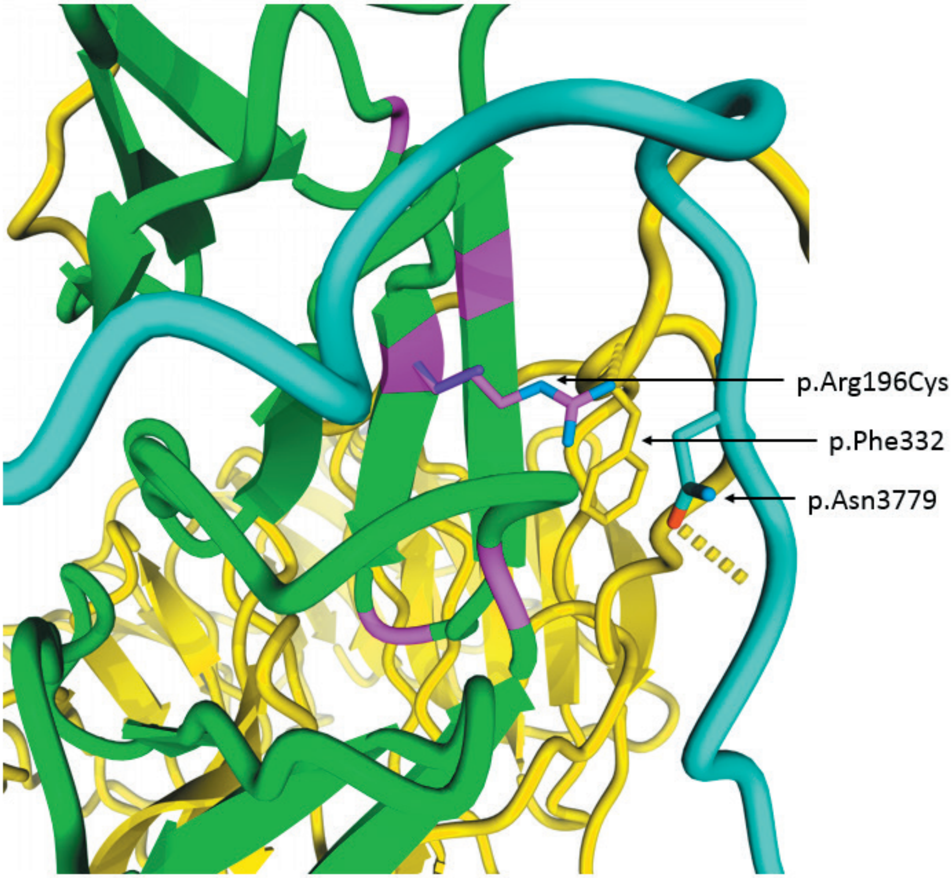


Figure S2: WDR5 (green) interaction with KMT2A/MLL1 (cyan) and RbBP5 (yellow), are shown from the core MLL1 complex (PDB:6KIV). The mutated aminoacid and nearby aminoacids are shown with sticks. The wild type arginine at the position p.196 is colored in magenta and the mutated cysteine in purple.

p.(Ala201Val)

Wild type residue role	Effect of the residue substitution
Ala201 is located on the WBM surface of WDR5 and interacts with RbBP5. It is located in clear proximity to Arg56 and Leu54 in the RbBP5 protein.	Despite the fact that valine is also small and non-polar, it has a bigger sidechain than alanine. Therefore, change to a valine at this position could affect the interaction with RbBP5 because of the change of the interaction surface.

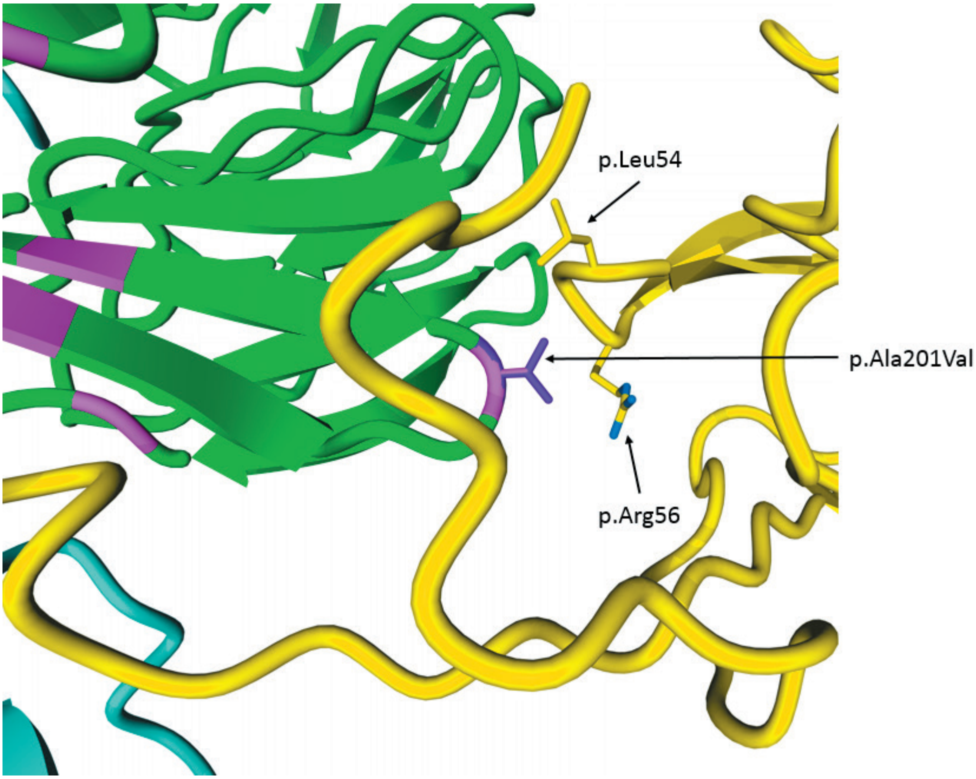


Figure S3: WDR5 (green) interaction with KMT2A/MLL1 (cyan) and RbBP5 (yellow), are shown from the core MLL1 complex (PDB:6KIV). The mutated aminoacid and nearby aminoacids are shown with sticks. The wild type alanine at the position p.201 is colored in magenta and the mutated valine in purple.

p.(Thr208Met)

Wild type residue role	Effect of the residue substitution
Thr208 in WDR5 interacts with several RbBP5 C-term tail amino acids (Ala331, Pro334). Additionally, it makes a hydrogen-bond with a backbone of Ala331 in RbBBP5.	Methionine has a substantially bigger size than threonine and is not able to form the hydrogen-bond with RbBP5 Ala331. Therefore, a change to methionine at this position is expected to disrupt the WDR5 interaction interface with RbBP5.

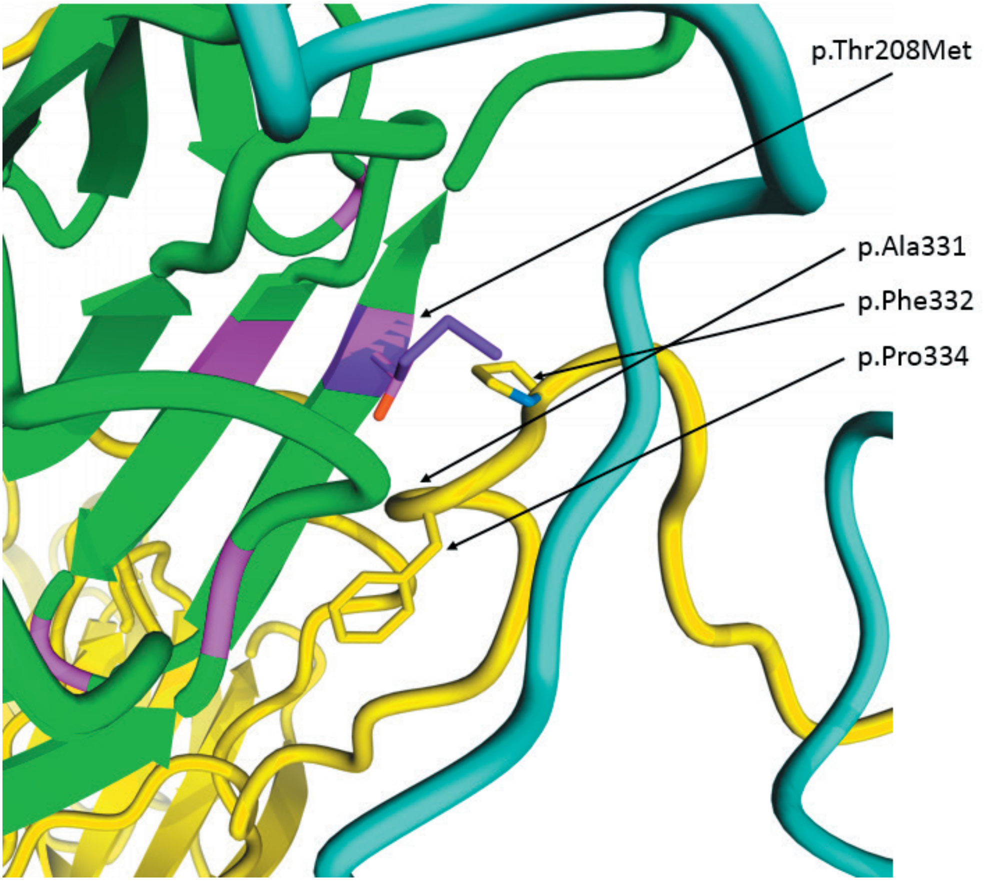


Figure S4: WDR5 (green) interaction with KMT2A/MLL1 (cyan) and RbBP5 (yellow), are shown from the core MLL1 complex (PDB:6KIV). The mutated aminoacid and nearby aminoacids are shown with sticks. The wild type threonine at the position p.208 is colored in magenta and the mutated methionine in purple.

p.(Asp213Asn)

Wild type residue role	Effect of the residue substitution
Asp213 is located in a WDR5 hydrophylic loop, which is involved in the interaction with the KMT enzymes. Although located distantly, it may interact with a positively charged KMT2A Lys3878, because the lysine has a highly flexible sidechain. Additionally, Asp213 forms a hydrogen-bond with Asn235 in WDR5.	Change to the aspartate would result in a similar amino acid with similar size, although the negative charge of the aspartic acid would be lost. The hydrogen bond with WDR5 Asn235 would be lost due to this change, which may disrupt the stability and position of the loop, and, therefore, affect interaction with the KMT enzymes. Additionally, it can lose interactions with positively charged KMT residues. However, the exact effect of the variant is unknown.

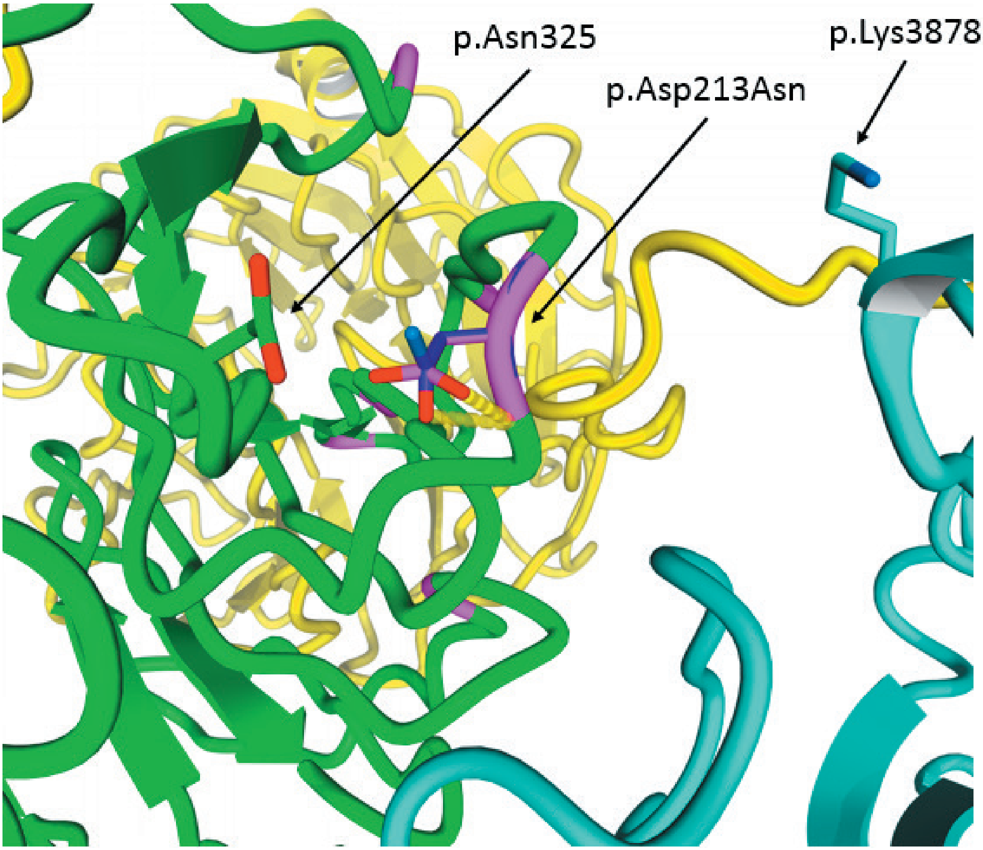


Figure S5A: WDR5 (green) interaction with KMT2A/MLL1 (cyan) and RbBP5 (yellow) are shown from the core MLL1 and MLL3 complexes (PDB:6KIV and 6KIW, respectively). The mutated aminoacid and nearby aminoacids are shown with sticks. The wild type aspartic acid at the position p.213 is colored in magenta and the mutated aspartate in purple.

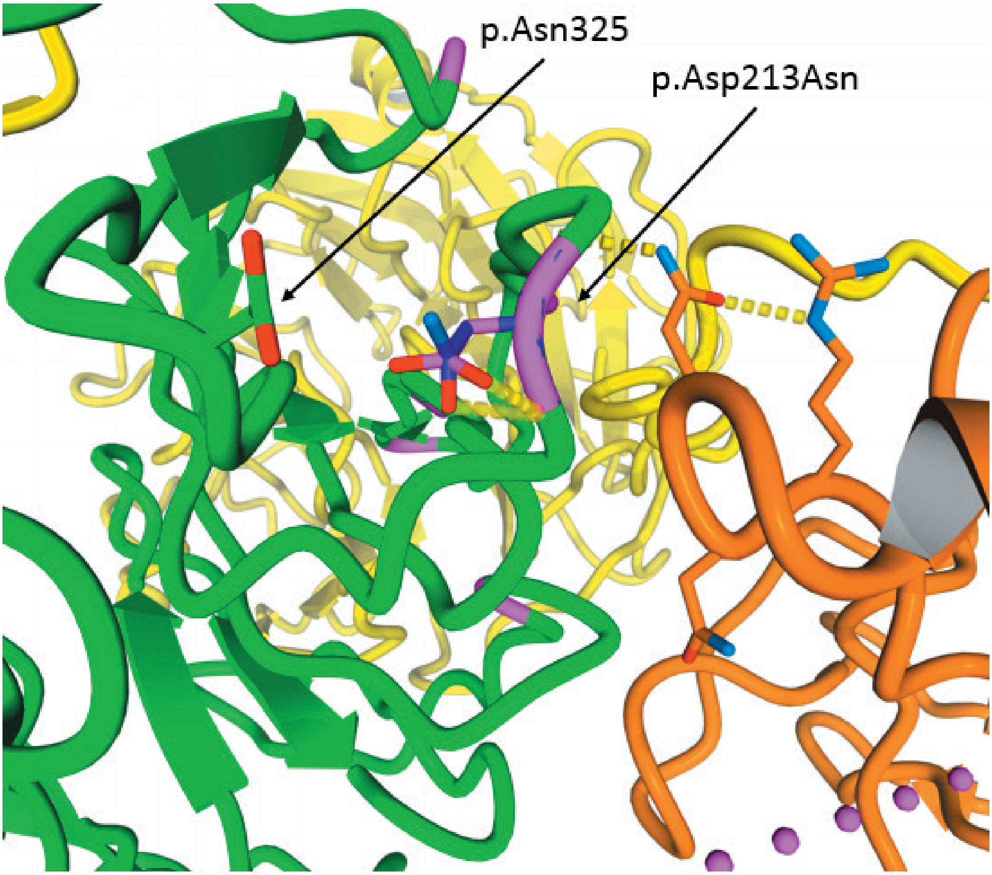


Figure S5B: WDR5 (green) interaction with KMT2C/MLL3 (orange) and RbBP5 (yellow) are shown from the core MLL1 and MLL3 complexes (PDB:6KIV and 6KIW, respectively). The mutated aminoacid and nearby aminoacids are shown with sticks. The wild type aspartic acid at the position p.213 is colored in magenta and the mutated aspartate in purple.

Wild type residue role	Effect of the residue substitution
Lys245 located in a position that is a significant distance from the site of interaction with RbBP5 and KMT enzymes and is not known to be involved in a protein interaction.	A change from lysine to arginine at this position, would result in a similar amino acid by charge and flexibility of the side-chain with minimal effect on protein structure or interactions. Even though arginine is slightly larger, the effect of this variant is not clear.

p.(Lys245Arg)

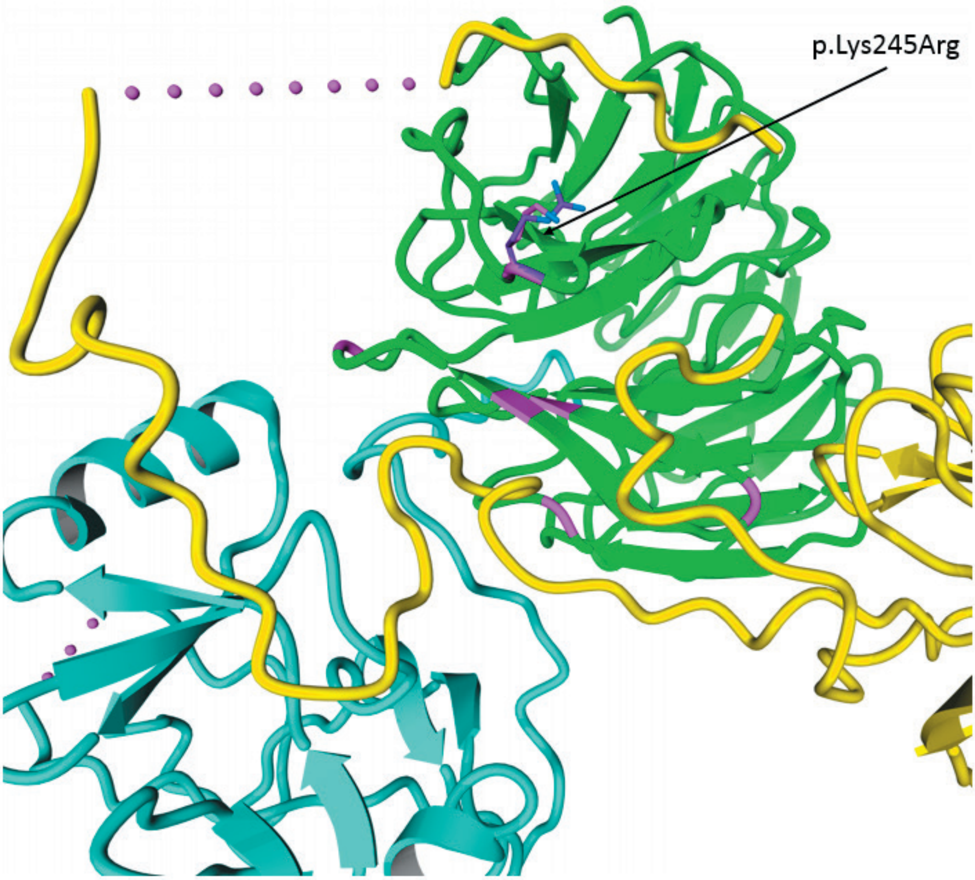


Figure S6: WDR5 (green) interaction with KMT2A/MLL1 (cyan) and RbBP5 (yellow), are shown from the core MLL1 complex (PDB:6KIV). The mutated aminoacid and nearby aminoacids are shown with sticks. The wild type lysine at the position p.245 is colored in magenta and the mutated arginine in purple.

Table S1:

Individual	Individual 1	Individual 2	Individual 3	Individual 4	Individual 5	Individual 6	Individual 7	Individual 8	Individual 9	Individual 10
Molecular characteristics										
cDNA (NM_017588.2)	c.505C>G	c.586C>T	c.586C>T	c.602C>T	c.623C>T	c.623C>T	c.623C>T	c.623C>T	c.637G>A	c.734A>G
Predicted protein effect	p.(Ala169Pro	p.(Arg196Cys)	p.(Arg196Cys)	p.(Ala201Val)	p.(Thr208Met)	p.(Thr208Met)	p.(Thr208Met)	p.(Thr208Met)	p.(Asp213Asn)	p.(Lys245Arg)
PMID with sequencing methods	24123792	25847626	25533962	25533962	26299366 and 29463886	26299366	29681102	30462361	29737001	27616483
Inheritance	De novo	De novo	De novo	De novo	De novo	De novo	De novo	De novo	De novo	De novo
Other SNV/CNV reported (hg19)	No	No	No	chr4:(41,459,887-41,951,637)x3 de novo	No	No	No	No	chr12:(1,944,805-1,985,246)x1	No
CADD score (v1.4)	25,8	34	34	23,1	26,4	26,4	26,4	26,4	22,8	23
SIFT score	deleterious (0)	deleterious (0)	deleterious (0)	tolerated (score 0.1)	deleterious (0)	deleterious (0)	deleterious (0)	deleterious (0)	tolerated (0.12)	tolerated (0.42)
PolyPhen-2	Probably damaging (1.000)	Probably damaging (1.000)	Probably damaging (1.000)	Benign (0.029)	Probably damaging (0.997)	Probably damaging (0.997)	Probably damaging (0.997)	Probably damaging (0.997)	Benign (0.013)	Benign (0.000)
Clinical characterisation										
Gender	Female	Female	Male	Male	Female	Male	Male	Female	Female	Female
Age (at last visit)	36 y	6,5 y	16y	10 y	11 y	25 y	50 y	17 y	3 y	6 y
Height	+0,5 SD	+3 SD	0.75 SD	-1.75 SD	0 SD	+3 SD	+0,5 SD	+1,5 SD	-1 SD	-0,5 SD
Weight (for height)	+2 SD	+1,5 SD	0 SD	-1.25 SD	-0,75 SD	+5 SD	+2 SD	+1,5 SD	+2 SD	-2,5 SD
Head circumference	+0,5 SD	+1,5 SD	-0,75 SD	-1 SD	0 SD	+4 SD	-0,5 SD	-1,5 SD	-2 SD	+1 SD
Developmental delay	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intellectual disability	Mild	Moderate	Moderate	No	Moderate	Borderline	Mild	Moderate	Moderate	Mild
Delayed motor development	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Walking independently	20 months	31 months	21 months	40 months	18 months	16 months	17 months	19 months	18 months	12 months
Speech delays	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Details speech development	unknown	no speech; at the age of 6.5y only two words	speech and language delays, nasal speech, oromotor difficulties, dribbling	Minimal speech due to severe micrognathia/jaw dysplasia.	Verbal dyspraxia, mixed language disorder, nasal speech	Yes, severe mixed language disorder, nasal speech	Persistent stuttering, nasal speech	No speech	Expressive language disorder	Speech delays, nasal speech
Autism spectrum disorder (ASD)	ASD diagnosis	ASD diagnosis	ASD diagnosis	No	No	ASD suspected	No	ASD diagnosis	ASD suspected	No
ADHD	No	No	ADHD diagnosis	No	No	ADHD diagnosis	No	No	ADHD suspected	No
Hypotonia	No	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes
Brain MRI	not performed	normal (1y11m and 6y)	Mild ventricular dilatation with thinning of the posterior corpus callosum (4y)	normal (11y)	Subtle grey matter heterotopias in margins frontal horns (12y)	normal (11y)	not performed	Normal (17y)	Periventricular gliosis right>left (2y3m)	not performed
Epilepsy	Yes, absences since age 4y	No; EEG (7y): epileptic activity during sleep	Yes, absence seizures, resolved in teenage years	No	Yes, seizure disorder since age 4y	Yes, epileptic episodes in childhood. Medication discontinued at age 17y	Convulsions after surgery in early childhood, medication discontinued at age 10y. At age 53y diagnosed with epilepsy again, requiring medication.	Yes, refractory generalized myoclonic epilepsy	No	No
Skeleton/limb abnormalities	Severe syndactyly 2-3	No	No	Hemivertebra L5, kyphosis (possibly secondary to hemivertebra); scoliosis affecting mobility	Bilateral clubfeet	Single palmar creases, 5th digit clinodactyly	Osteoporosis	No	Hemihypertrophy left leg, scoliosis	Small middle phalange and mild clinodactyly of 5th digit bilaterally. Single palmar creases, fetal finger pads. Syndactyly 2-3. In-toeing and femoral anteversion.
Heart abnormalities	No	No	No	No	No	No	Cardiac arrhythmias and decompensated heart failure requiring surgery	Unknown	Left ventricular non compaction cardiomyopathy	No (except for history of patent foramen ovale and pulmonary hypertension in newborn period)

Vision problems	No	No	No	No	No	No	Has glasses, details unknown.	No	Strabismus (right side)	Anisometropic amblyopia right eye, bilateral hyperopia with astigmatism
Hearing loss	No	No	No	Conductive hearing loss, severe meatal stenosis	No	No	No	No	No	Left ear: mild primarily conductive hearing loss. Right ear: mild mixed hearing loss at 250 Hz only.
Facial dysmorphisms	Thin lips and mild prognatism.	High forehead, large eyes, wide eyebrows, large nasal tip, hypoplastic alae nasi, mildly dysplastic ears. Sparse and thin hair.	No	Severe micrognathia and small mouth requiring tracheostomy. Low-set posteriorly rotated ears, tiny meatus bilaterally. Dermatofibrosarcomas.	No	Upslanting palpebral fissures	Low set ears, long philtrum, full lips.	No	Wide-set eyes, synophrys, long eyelashes. Lobeless posteriorly rotated ears, underdeveloped antitragus. Flattened nasal tip, large anteverted nares. Thin lip vermilion.	Micrognathia, small mouth. High palate, dimples on sides of the mouth. Thin upper lip. Ptosis.
Frequent infections	No	No	No	No	Frequent ear infections. Pneumonia in history.	Frequent ear infections.	No	No	No	Recurrent sinopulmonary infections, including numerous episodes of pneumonia.
Other	No	Two small cafe-au-lait spots	No	Ulcerative colitis aged 16; severe anaemia in response to sulphasalazine (resolved on withdrawal)	Gets dehydrated easily., Kidney reflux as a child.	Dry skin, mottled feet. Progressive overgrowth.	No	Primary amenorrhea, constipation, sleep disturbances	Truncal ataxia. Hypertrichosis of the back. Left inguinal hemangioma	Reflux, dysfagia.

6

Chapter 6

Heterozygous variants that disturb the transcriptional repressor activity of FOXP4 cause a developmental disorder with speech/language delays and multiple congenital abnormalities

Lot Snijders Blok, Arianna Vino, Joery den Hoed, Hunter R. Underhill, Danielle Monteil, Hong Li, Francis Jeshira Reynoso Santos, Wendy K. Chung, Michelle D. Amaral, Rhonda E. Schnur, Teresa Santiago-Sim, Yue Si, Han G. Brunner, Tjitske Kleefstra* and Simon E. Fisher*

** These authors contributed equally*

Genetics in Medicine 2020
doi: 10.1038/s41436-020-01016-6

Abstract

Purpose:

Heterozygous pathogenic variants in various *FOXP* genes cause specific developmental disorders. The phenotype associated with heterozygous variants in *FOXP4* has not been previously described.

Methods:

We assembled a cohort of eight individuals with heterozygous and mostly de novo variants in *FOXP4:* seven individuals with six different missense variants and one individual with a frameshift variant. We collected clinical data to delineate the phenotypic spectrum, and used in silico analyses and functional cell-based assays to assess pathogenicity of the variants.

Results:

We collected clinical data for six individuals: five individuals with a missense variant in the forkhead box DNA-binding domain of FOXP4, and one individual with a truncating variant. Overlapping features included speech and language delays, growth abnormalities, congenital diaphragmatic hernia, cervical spine abnormalities, and ptosis. Luciferase assays showed loss-of-function effects for all these variants, and aberrant subcellular localization patterns were seen in a subset. The remaining two missense variants were located outside the functional domains of FOXP4, and showed transcriptional repressor capacities and localization patterns similar to the wild-type protein.

Conclusion:

Collectively, our findings show that heterozygous loss-of-function variants in *FOXP4* are associated with an autosomal dominant neurodevelopmental disorder with speech/ language delays, growth defects, and variable congenital abnormalities.

Introduction

The FOXP subgroup of transcription factors consists of four different proteins: FOXP1, FOXP2, FOXP3, and FOXP4, all with important regulatory functions in developmental processes[1–3]. For three of these FOXP proteins, heterozygous loss-of-function variants have been shown to cause Mendelian disorders, encompassing a broad spectrum of associated phenotypes. Variants in *FOXP1* cause an intellectual disability syndrome with speech delays, autism spectrum disorder, dysmorphisms, and congenital abnormalities in some affected individuals (MIM 613670)[4]; variants in *FOXP2* give rise to a disorder in which childhood apraxia of speech is a prominent feature (MIM 602081)[5]; while variants in *FOXP3* can cause X-linked immunodysregulation, polyendocrinopathy, and enteropathy (MIM 304790)[6].

In contrast to the other *FOXP* genes, *FOXP4* has not yet been convincingly linked to a Mendelian disorder. *FOXP4* is expressed in subsets of cells in a variety of tissues throughout the body, including in the developing brain, lungs, and gut[2,7]. The encoded protein has regulatory roles in the development and maturation of the central nervous system[8,9]. It is coexpressed with FOXP1 and/or FOXP2 in several different brain regions, such as the cortex, cerebellum, and striatum[10], where these transcription factors may heterodimerize, to potentially coregulate downstream targets. The phenotype associated with heterozygous germline *FOXP4* variants remains to be defined. A homozygous loss-of-function variant in *FOXP4* was previously reported in a child with developmental delays, laryngeal hypoplasia, feeding difficulties, and a ventricular septal defect, suggesting autosomal recessive inheritance[11]. However, several different heterozygous de novo *FOXP4* variants of unknown significance have been identified in research cohorts that included individuals with specific disorders (developmental disorders, congenital diaphragmatic hernia, or high myopia)[12–14] and in clinical diagnostic next-generation sequencing laboratories, fitting a possible autosomal dominant disease model.

We aimed to study if heterozygous de novo *FOXP4* variants can cause a specific human disorder by collecting clinical data of individuals with rare coding *FOXP4* variants, characterizing the associated phenotype, and investigating the functional impact of variants using cell-based assays. A better understanding of pathogenicity of different *FOXP4* variants and the associated disease models might directly improve clinical care by facilitating correct classification of variants found in diagnostic and researchbased sequencing studies and providing families with precise recurrent risks. In addition, research on rare *FOXP4* variants and the associated phenotypes expands our knowledge of the key roles that FOXP transcription factors play in human disease.

Materials and methods

Identification and clinical characterization of individuals with FOXP4 variants

We used GeneMatcher[15] and denovo-db[16] to identify individuals with de novo variants in the coding region of *FOXP4* (including canonical splice sites) and individuals with reported *FOXP4* variants of unknown significance in diagnostic next-generation sequencing studies. De-identified clinical data and variant details were collected using Castor EDC[17]. Additional single-nucleotide variants and copynumber variants considered to be possibly pathogenic and/ or to possibly contribute to the phenotype, are listed in Table S1. All variants in this paper are annotated with respect to the NM_001012426.1 transcript (FOXP4 isoform 1).

Cell culture and transfection

HEK293T/17 cells (CRL-11268, ATCC) were cultured in DMEM (Gibco) with 10% fetal bovine serum (Gibco) and Pen/Strep (Thermo Fisher) at 37°C with 5% CO_2. GeneJuice (Merck Millipore) was used for transfection, following the manufacturer's protocol.

DNA constructs and site-directed mutagenesis

Wild-type FOXP4 (NM_138457.2; FOXP4 isoform 2) was amplified from human fetal brain complementary DNA (cDNA) using the primers listed in Table S2. Isoform 2 (NM138457.2; 667 amino acids) is a slightly shorter isoform than isoform 1 (NM_001012426.1; 680 amino acids). For consistency, all variants in this study are annotated using isoform 1. Constructs carrying FOXP4 variants were generated using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent). Primer sequences used for site-directed mutagenesis are provided in Table S2. FOXP4 wild-type and variant cDNAs were subcloned into pYFP and pRluc vectors (Clontech) using BamHI and XbaI restriction sites. All constructs were verified by Sanger sequencing. Plasmid sequences are available upon request.

Luciferase assays

For the luciferase assays, we used a pGL4.23 firefly luciferase reporter vector (Promega), in which the promoter region of *SRPX2* was subcloned as previously described[18]. HEK293T/17 cells were transfected with this firefly reporter construct (9.45 ng), a FOXP4-YFP-expression construct or empty YFPexpression vector (41.36 ng), and a pGL4.74 (hRluc/TK) Renilla reniformis luciferase construct (0.30 ng) 24 hours after seeding in 96-well plates. At 24 hours post-transfection, cells were lysed and luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega) and an Infinite M Plex microplate reader (Tecan). Firefly luciferase activities (experimental condition) were normalized to Renilla luciferase activities (control condition).

Fluorescence imaging of subcellular localization

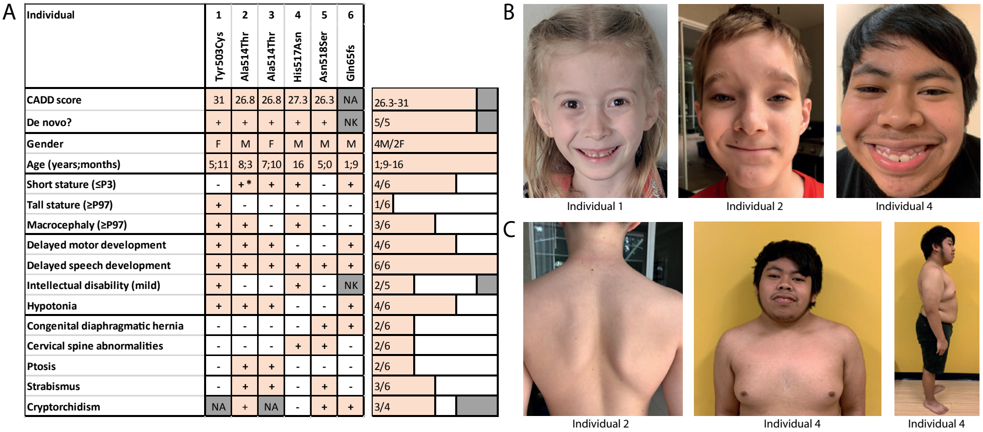
HEK293T/17 cells were grown on coverslips coated with poly- D-lysine (Merck Millipore) in a 24-well plate, and transfected 24 hours after seeding, with 125 ng DNA per well. At 24 hours post-transfection, the cells were fixed with 4% paraformaldehyde (Electron Microscopy

Sciences) in PBS for 15 minutes at room temperature. Hoechst 33342 (Invitrogen) was used for nuclear staining, before mounting with Fluorescence Mounting Medium (Dako).

Bioluminescence resonance energy transfer assays

Bioluminescence resonance energy transfer (BRET) assays were performed as previously described[19]. HEK293T/17 cells were plated in white 96-well plates with transparent bottoms (Greiner) and transfected with equimolar concentrations of YFP and RLuc plasmids. A RLuc-NLS (nuclear localization signal) plasmid was used as a negative control. At 40 hours post-transfection, medium was replaced with DMEM without phenol red and 10% fetal bovine serum (both Invitrogen), supplemented with 60 µM EnduRen Live Cell Substrate (Promega) and incubated for four hours at 37°C. An Infinite F200PRO Microplate reader (TECAN) was used for the measurements using the Blue1 and Green1 filter. Corrected BRET ratios were calculated using 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-NLS plasmid expressed in the control condition.



Individual	1	2	3	4	5	6	
	Tyr503Cys	Ala514Thr	Ala514Thr	His517Asn	Asn518Ser	Gln65fs	
CADD score	31	26.8	26.8	27.3	26.3	NA	26.3-31
De novo?	+	+	+	+	+	NK	5/5
Gender	F	M	F	M	M	M	4M/2F
Age (years;months)	5;11	8;3	7;10	16	5;0	1;9	1;9-16
Short stature (≤P3)	-	+ *	+	+	-	+	4/6
Tall stature (≥P97)	+	-	-	-	-	-	1/6
Macrocephaly (≥P97)	+	+	-	+	-	-	3/6
Delayed motor development	+	+	+	-	-	+	4/6
Delayed speech development	+	+	+	+	+	+	6/6
Intellectual disability (mild)	+	-	-	+	-	NK	2/5
Hypotonia	+	+	+	-	-	+	4/6
Congenital diaphragmatic hernia	-	-	-	-	+	+	2/6
Cervical spine abnormalities	-	-	-	+	+	-	2/6
Ptosis	-	+	+	-	-	-	2/6
Strabismus	-	+	+	-	+	-	3/6
Cryptorchidism	NA	+	NA	-	+	+	3/4

Figure 1: Clinical features and dysmorphisms

(a) Visual overview of clinical features present in six individuals with a heterozygous FOXP4 variant, more details on phenotypes are provided in Table S1. + present, - not present, *NA* not applicable, *NK* not known. *Short stature in history, after growth hormone treatment now normal height. **(b)** Facial phenotype of three individuals with a *FOXP4* variant. Recurrently reported dysmorphisms include tented and/or flared eyebrows, ptosis, small teeth, and gingival hyperplasia. **(c)** Additional abnormalities as noted by physical examination. In individual 2, asymmetric scapulae were reported. Individual 4 presented with a very short stature (<P1) and a short and broad neck.

Statistical analysis

For protein expression experiments, quantified microscopy data, luciferase reporter assays, and BRET assays, statistical analysis was done for each type of assay using one-way analysis of variance (ANOVA) followed by Bonferroni correction for the number of conditions tested. All analyses were performed with GraphPad Prism software.

Ethics statement

All experiments were performed in accordance with relevant guidelines and regulations. All study proceedings involving humans were in compliance with the principles set out in the Declaration or Helsinki. Next-generation sequencing in this study was either performed in a diagnostic setting (with relevant clinical quality accreditations and consent procedures) or in a research setting (University of Alabama at Birmingham Institutional Review Board [IRB-300000328] and Columbia University Irving Medical Center Institutional Review Board [IRB-AAAB2063]). For all individuals in this study, written consent was obtained for publication of the data. For the individuals of which photos are published, specific consent for publication of photos was obtained.

Results

Identification of FOXP4 variants

Using denovo-db[16] and GeneMatcher[15], we aimed to collect data on all reported de novo variants in the coding region of *FOXP4* in research cohorts, as well as all reported *FOXP4* variants of unknown significance in diagnostic sequencing cohorts. Eight unrelated individuals with heterozygous *FOXP4* variants were identified, seven of whom had a de novo missense variant. One individual carried a truncating *FOXP4* variant that was not inherited from the mother; the father was unavailable for testing (Fig. 1a; Table S1). Among the seven individuals with a de novo missense variant, six different variants were found; two unrelated individuals had the same variant (p.Ala514Thr). None of the variants included in our study were present in the gnomAD and dbSNP databases.

Phenotypes of individuals with heterozygous FOXP4 variants

We were able to collect further details on phenotypes for six of the eight individuals with *FOXP4* variants: five individuals with a missense variant in the forkhead box DNA-binding domain (four different variants, one recurrent), and the individual with a heterozygous truncating variant (p.Gln65Serfs*20). A summary of recurrent clinical features for these six individuals can be found in Fig. 1a, with a more detailed overview in Table S1. For the two remaining individuals, both of whom had a missense variant outside the forkhead box domain, we were not able to collect additional information on phenotypes (all available data are included in Table S1).

The six individuals included in our phenotypic comparison comprised four males and two females, with an age range of 1 year 9 months to 16 years. Four individuals had a short stature (≤P3), one of these four reached a normal height after treatment with growth hormone. One individual had a tall stature. Macrocephaly (head circumference ≥ P97) was seen in three out of six individuals. Weights were generally normal for height, although one individual (individual 3) had a low weight (≤P3).

Developmental delays were observed in all six individuals. While only four of six individuals showed delayed motor development, speech/language development was delayed in all of them. All six individuals had received speech therapy and two individuals had a formal diagnosis of expressive language disorder. Despite having shown prominent speech delays in infancy, for three of six individuals (aged 5–16 years) current speech is described as normal with full and complex sentences. Two individuals had a mild intellectual disability, three individuals had no intellectual disability, and for one individual this was unknown. Infant hypotonia was seen in four of the six individuals.

Different types of congenital abnormalities were present in different individuals. Interestingly, congenital diaphragmatic hernia was present in two individuals. Vertebral abnormalities were present in two individuals: one individual had abnormalities of the craniocervical junction and malformations of several arches of C1, C2, and C3 vertebrae (details in Table S1) and in the other individual vertebra C1 was fused to the skull. An additional individual had uneven scapulae, but normal spine films (Fig. 1c). Pectus excavatum was reported in another individual. Two individuals had ptosis (requiring surgery in one individual), and strabismus was reported in three individuals. Cryptorchidism was present in three of four males. In addition to congenital abnormalities, overlapping facial features were reported in several individuals, which included tented and/or flared eyebrows, small teeth, and gingival hyperplasia (Fig. 1b).

In silico variant analysis

We used an array of computational tools to predict the functional effects of all missense variants that were found, including the two missense variants for which no additional information on phenotypes was available. Four of the six different missense variants clustered in the DNA-binding forkhead box domain of the encoded FOXP4 protein, while the remaining two were located outside known functional domains (Fig. 2a). The cross-species conservation of the amino acid sequences in the affected regions is shown in Fig. 2b. The mutated amino acid sites are invariant across all the species that we analyzed, with the sole exception of the Serine 273 residue, which is less conserved. For all missense variants, CADD, PolyPhen, and SIFT scores were derived, all of which predicted pathogenicity for the four forkhead box domain variants (Fig. 1a and Table S1).

As no three-dimensional protein structure is available for FOXP4, we used the SWISS-MODEL Homology Modeling online tool[20] to create a homology model of the forkhead box domain structure of FOXP4 (amino acids 456–542) based on a FOXP1 template model. We then visualized the threedimensional location of the four different missense variants mapping to this functional domain (Fig. 2c). Three of the four missense variants (p.Ala514Thr, p.His517Asn, p.Asn518Ser) are located in the third helix of the DNA-binding domain (the recognition helix), and the fourth variant (p.Tyr503Cys) maps to the hinge loop region.

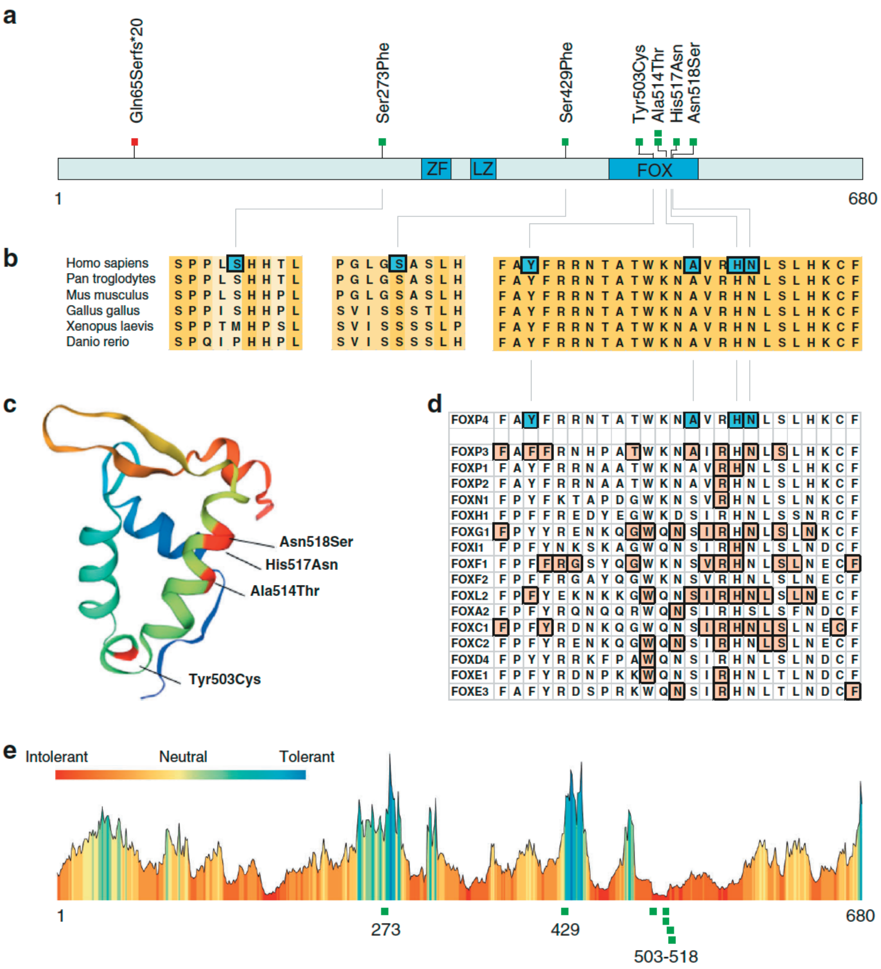
FOXP4 belongs to the large family of FOX transcription factor proteins, defined by the presence of the distinctive highly conserved forkhead box domain. For at least 16 FOX proteins, missense variants in this characteristic DNA-binding domain have already been linked to Mendelian disorders in humans[21,22]. We therefore assessed whether the potentially pathogenic missense variants that we identified in the forkhead box domain of FOXP4 were comparable with the known pathogenic missense variants in these other FOX transcription factors (Fig. 2d). Indeed missense variants in the FOXP4 DNA-binding domain matched well to the known pathogenic missense variants in other FOX transcription factors (Fig. 2d).

We went on to use the MetaDome web tool[23] to visualize all six different *FOXP4* missense variants in the tolerance landscape of the gene (Fig. 2e), which shows regional tolerance for genetic variation based on a missense over synonymous variant count ratio using data from the gnomAD database[24]. This showed us that the four missense variants that cluster in the FOX domain are located in a region of high intolerance (low missense over synonymous variant count ratio), while the two that map elsewhere are located in more tolerant regions of the protein (see Table S1). It is interesting to note that the gnomAD Z-score for missense variants in *FOXP4* as a whole is not particularly high (1.95)[24], indicating that the complete coding region of the *FOXP4* gene is not extremely intolerant for missense variation overall. This finding is in line with the results from the MetaDome analysis, which show that only a few small regions of *FOXP4* show high intolerance for missense variants, including the part of the forkhead box domain in which our four different missense variants are located.

Effects of variants on localization and transcriptional repression activity

Functional assays in HEK293T/17 cells were performed for all the seven different FOXP4 variants that were identified (six missense variants and one variant causing an early frameshift). We used overexpression constructs of FOXP4 (isoform 2) with an N-terminal YFP-tag to assess the subcellular localization of the respective mutant FOXP4 proteins. While all experiments were performed with isoform 2 FOXP4 proteins (consisting of 667 amino acids), all variants in this study are annotated in isoform 1 for consistency of the interpretation. Immunoblotting indicated that the wild-type and mutant proteins were expressed at the expected size and at comparable levels (Fig. S1). In assessments of subcellular localization using fluorescence imaging, wildtype FOXP4 showed nuclear localization, as did mutant proteins with the two missense variants mapping outside the known functional domains (p.Ser273Phe and p.Ser429Phe; Fig. 3a). Three of the four different missense variants in the forkhead box domain led to aberrant localization of the mutant protein: p.Tyr503Cys and p.His517Asn showed cytoplasmic expression with aggregates, and for the p.Asn518Ser variant a nuclear granular pattern was seen. Overexpression of the truncated protein yielded by the frameshift variant (p.Gln65Serfs*20) led to diffuse mislocalization in the cytoplasm, although the protein was still present in the nucleus.

We used luciferase assays to assess the capacity of FOXP4 to repress an *SRPX2*-derived promoter element. Wild-type FOXP4 showed significant repression of reporter gene expression compared with a control construct (Fig. 3b). The four FOXP4 proteins with amino acid substitutions in the forkhead box domain (p.Tyr503Cys, p.Ala514Thr, p.His517Asn, and p.Asn518Ser) all showed a loss of this transcriptional repressor activity, significantly different from wild-type FOXP4. Loss of function was also seen for the truncated FOXP4 protein (p.Gln65Serfs*20), consistent with the lack of a DNA-binding domain. For proteins with the two remaining missense variants (p.Ser273Phe and p.Ser429Phe), both located outside known functional domains of FOXP4, repression capacities were no different from the wild-type protein. In summary, the localization and luciferase assays pointed to pathogenicity for the four missense variants in the forkhead box domain, with a loss-of-function mechanism, in contrast to the two missense variants located elsewhere in the protein, which did not differ from wild-type in these experiments.



Homo sapiens	S P P L S H H T L	P G L G S A S L H	F A Y F R R N T A T W K N A V R H N L S L H K C F
Pan troglodytes	S P P L S H H T L	P G L G S A S L H	F A Y F R R N T A T W K N A V R H N L S L H K C F
Mus musculus	S P P L S H H P L	P G L G S A S L H	F A Y F R R N T A T W K N A V R H N L S L H K C F
Gallus gallus	S P P I S H H P L	S V I S S S T L H	F A Y F R R N T A T W K N A V R H N L S L H K C F
Xenopus laevis	S P P T M H P S L	S V I S S S S L P	F A Y F R R N T A T W K N A V R H N L S L H K C F
Danio rerio	S P Q I P H H P L	S V I S S S S L H	F A Y F R R N T A T W K N A V R H N L S L H K C F

FOXP4	F A Y F R R N T A T W K N A V R H N L S L H K C F
FOXP3	F A F F R N H P A T W K N A I R H N L S L H K C F
FOXP1	F A Y F R R N A A T W K N A V R H N L S L H K C F
FOXP2	F A Y F R R N A A T W K N A V R H N L S L H K C F
FOXN1	F P Y F K T A P D G W K N S V R H N L S L N K C F
FOXH1	F P F F R E D Y E G W K D S I R H N L S S N R C F
FOXG1	F P Y Y R E N K Q G W Q N S I R H N L S L N K C F
FOXI1	F P F Y N K S K A G W Q N S I R H N L S L N D C F
FOXF1	F P F F R G S Y Q G W K N S V R H N L S L N E C F
FOXF2	F P F F R G A Y Q G W K N S V R H N L S L N E C F
FOXL2	F P F Y E K N K K G W Q N S I R H N L S L N E C F
FOXA2	F P F Y R Q N Q Q R W Q N S I R H S L S F N D C F
FOXC1	F P F Y R D N K Q G W Q N S I R H N L S L N E C F
FOXC2	F P F Y R E N K Q G W Q N S I R H N L S L N E C F
FOXD4	F P Y Y R R K F P A W Q N S I R H N L S L N D C F
FOXE1	F P F Y R D N P K K W Q N S I R H N L T L N D C F
FOXE3	F A F Y R D S P R K W Q N S I R H N L T L N D C F

Figure 2: In silico analyses of heterozygous *FOXP4* variants

(a) Linear representation of the FOXP4 protein (Q8IVH2–1) with the identified variants and functional domains annotated: *FOX* forkhead box domain, *LZ* leucine zipper, *ZF* zinc finger. **(b)** Conservation of FOXP4 across different species, with the amino acids affected by missense variants indicated. Species include *Homo sapiens* (UniProt sequence Q8IVH2), *Pan troglodytes* (A0A2J8NZN5), *Mus musculus* (Q9DBY0), *Gallus gallus* (A0A3Q2U1E5), *Xenopus laevis* (Q4VYR7), and *Danio rerio* (B3DJK9). Regions shown span amino acids 269–277, 425–433, and 501–525 of FOXP4 isoform 1 (Q8IVH2). **(c)** Visualization of missense variants in the FOX domain in a three-dimensional structure. A homology model for the FOX domain of FOXP4 (amino acids 456–542) was built based on template structure 2kiu.1.A (FOXP1 monomer), using the SWISS-MODEL Homology Modeling online tool.[20] **(d)** Alignment of missense variants in a subset of the FOX domain with pathogenic missense variants in other FOX proteins. An alignment was made of the Pfam Forkhead domain (PF00250) using Clustal Omega Multiple Sequence Alignment[44] of all FOX proteins with missense variants present in HGMD database 2019.3. Only missense variants labeled as pathogenic were included for this analysis. **(e)** Tolerance landscape of FOXP4 protein visualized via the MetaDome web server.[23] A tolerance landscape is computed based on single-nucleotide variants in the gnomAD database, and shows per amino acid position the missense over synonymous ratio in a sliding window of 21 residues. Green and blue peaks represent regions tolerant to missense variation; red valleys show intolerant regions. The missense variants in the FOX domain are located in extremely intolerant regions of FOXP4, while the two remaining missense variants are located in extremely tolerant regions.

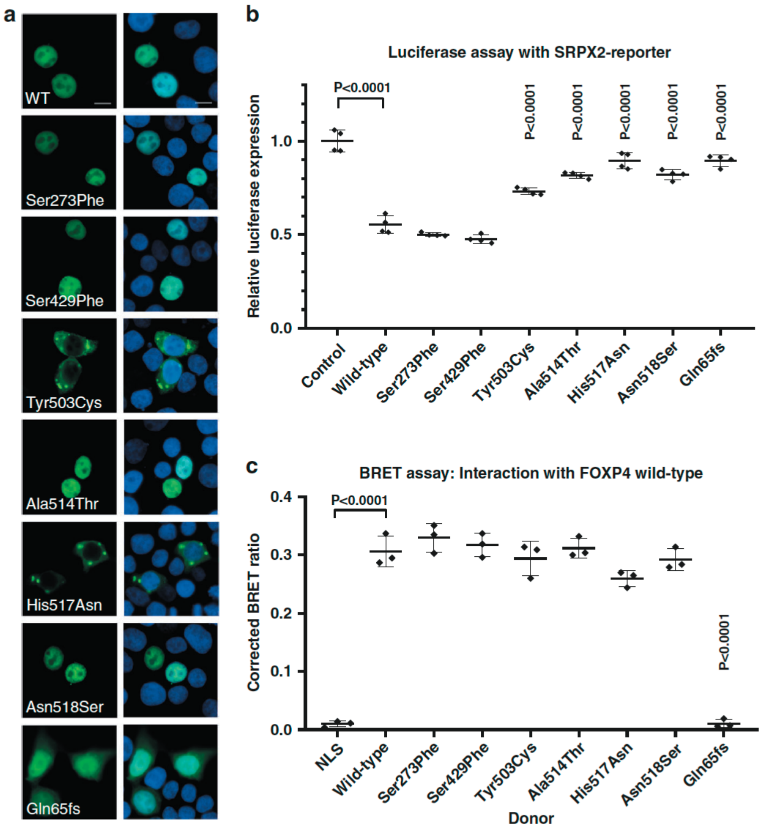


Figure 3: Functional assays to assess pathogenicity

(a) Direct fluorescence imaging of HEK293T/17 cells expressing YFP-FOXP4 fusion proteins (green) with the different FOXP4 variants found in our cohort. Nuclei are stained with Hoechst 33342. Scale bar = 10 µm. **(b)** Results of luciferase assays with FOXP4-YFP constructs and the SRPX2-reporter construct. Values are expressed relative to the control construct and represent the mean ± SD of four independent experiments, each performed in triplicate. *P* values were calculated using one-way analysis of variance (ANOVA) with Bonferroni correction. **(c)** Results of bioluminescence resonance energy transfer (BRET) assays to measure dimerization capacity of mutant FOXP4 constructs (donor) with wild-type (WT) FOXP4 (acceptor). Values represent the corrected mean BRET ratio ± SD of three independent experiments performed in triplicate. *P* values were calculated using one-way ANOVA with Bonferroni correction.In panel A, B and C, 'Gln65fs' is used as a short description for the variant 'p.Gln65Serfs*20'.

As FOXP4 is known to be able to dimerize with itself and/or other FOXP proteins, mediated by a conserved leucine zipper motif[25], we also assessed the effects of variants on dimerization capacities using BRET assays. In these assays, wild-type and variant versions of FOXP4 with an N-terminal Renilla luciferase tag (donor) were coexpressed with wild-type FOXP4 with an N-terminal YFP-tag (acceptor). The corrected BRET ratios of all FOXP4 proteins with missense variants were no different from those of wild-type FOXP4, indicating intact dimerization capacities for all these proteins (Fig. 3c). The truncated version of FOXP4 showed a complete loss of dimerization capacity, similar to the negative rLuc-NLS control construct (Fig. 3c).

Discussion

To characterize the clinical and molecular consequences of heterozygous *FOXP4* variants identified in several nextgeneration sequencing cohorts, we collected data on individuals with rare and possibly pathogenic variants in this gene. We identified seven individuals with a de novo missense variant (six different variants, since one was found independently in two unrelated cases). Using luciferase assays, we showed that four of the six different missense variants had loss-of-function effects on transcription repressor activity of the encoded FOXP4 protein. Notably, these four disruptive missense variants were all located in the forkhead box DNAbinding domain, a key functional motif of the protein. There was also one individual with a frameshift variant of unknown parental origin. The transcript with the frameshift variant will most likely undergo nonsense-mediated decay (NMD), leading to *FOXP4* haploinsufficiency in this individual, and our cell-based experiments indicate that any truncated protein resulting from NMD escape would lack repressor activity. Based on our findings we conclude that heterozygous *FOXP4* variants can cause a neurodevelopmental disorder, with prominent speech/language problems, short stature, macrocephaly, overlapping dysmorphisms, congenital diaphragmatic hernia, and cervical vertebral abnormalities.

Four of the six different missense variants were clustered in the DNA-binding domain of the encoded protein, at positions that are highly conserved across species, and also across different members of the FOX transcription factor family (Fig. 2b, d). Three of these four missense variants (p.Ala514Thr, p.His517Asn, and p.Asn518Ser) map within the third helix of the DNA-binding domain (Fig. 2c), also known as the recognition helix since it mediates sequence specific interaction with nucleotides in the major groove of the DNA of downstream targets[26]. Many different missense variants in FOX family proteins at similar positions in the DNA-binding domain have already been shown to be pathogenic (Fig. 2d). Using direct immunofluorescence, we showed that three of the four missense variants located in the DNA-binding domain of FOXP4 led to an aberrant subcellular localization of the protein. Although the precise mechanism by which these variants affect the localization pattern is not known, these results match well with previous observations of aberrant localization patterns associated with missense variants in the FOX domain of *FOXP1*[27] and *FOXP2*[28], and in more distantly related forkhead genes such as *FOXC1*[29] and *FOXC2*[30]. We used luciferase assays with an SRPX2-derived promoter sequence as a reporter to demonstrate that each variant yielded a loss of transcriptional repression activity for the respective FOXP4 protein. The fourth DNA-binding domain variant, p.Tyr503Cys, is located in the hinge loop region of this motif. Previous studies reported that a variant of the conserved tyrosine residue at the equivalent position in FOXP2 (p.Tyr540Phe in FOXP2 isoform 1; NP_055306.1) disrupted DNA binding, and also had effects on dimerization capability[31]. In our assays, the p.Tyr503Cys variant of FOXP4 significantly disrupted transcription factor capacities to a similar degree to the other DNA-binding domain variants, consistent with loss of function, but no effect on dimerization with FOXP4 wildtype protein was observed. All in all, the observations in functional studies for the different forkhead box DNAbinding domain missense variants all point to a loss-offunction effect, which is in line with existing literature about other *FOX*-associated disorders.

Two missense variants (p.Ser273Phe and p.Ser429Phe) did not show any difference compared with wild-type FOXP4 in functional assays. For these variants, the functions of the regions and amino acids involved is not known. The p.Ser273Phe variant was found in a large exome sequencing study in children with developmental disorders[14] but no additional information on phenotype could be collected for this individual. The p.Ser429Phe variant was found in a small trio exome sequencing cohort, in a young child with high myopia[13]. This individual also carried a hemizygous missense variant in *CACNA1F* (NP_005174.2: p.[Arg1060Trp]), which has already been described as a pathogenic variant causing X-linked congenital stationary night blindness (MIM 300071), possibly explaining the phenotype in this individual. Taking all data into account, these two missense variants might very well be benign variants, although pathogenicity cannot be completely excluded based on our assays.

The remaining variant in our study was a frameshift variant, for which the transcript will most likely undergo NMD, resulting in haploinsufficiency for FOXP4. If the transcript with the variant would still escape NMD, a truncated and dysfunctional version of FOXP4 would be expressed that has an aberrant subcellular localization pattern, does not show transcriptional repressor capacities in *SRPX2*-reporter luciferase assays, and is unable to dimerize. *FOXP4* is known to be extremely intolerant of loss-of-function variation, with a probability of loss-of-function intolerance (pLI) score of 0.98 based on sequencing data from 141,456 individuals, providing independent evidence that *FOXP4* haploinsufficiency is pathogenic[24]. The Decipher database contains seven microdeletions encompassing *FOXP4*, but as these are all large deletions (2.41 to 4.57 Mb in size) it is hard to draw conclusions about the contribution of *FOXP4* haploinsufficiency to the corresponding phenotypes. But interestingly, in the literature one individual has been reported with developmental delays, laryngeal hypoplasia and a ventricular septal defect, and a homozygous truncating *FOXP4* variant: c.815del; p.(Leu272Profs*95)[11]. Both parents were shown to be heterozygous for this variant, suggesting autosomal recessive inheritance, but no further clinical details were reported on the parents or other family relatives. As the individual with the heterozygous frameshift variant in our cohort had a phenotype entirely in line with the individuals with the likely pathogenic forkhead box domain missense variants: a congenital diaphragmatic hernia, short stature, developmental delays, hypotonia, and cryptorchidism, we assume that the *FOXP4* variant is causative. Although for our missense variants we cannot exclude a possible dominant-negative mechanism in addition to the loss-of-function effects, in which FOXP4 proteins with these variants would interfere with wild-type FOXP protein functions via their intact dimerization capacities, we propose that truncating variants in FOXP4 can be pathogenic in a heterozygous state.

FOXP4 was first characterized by Lu et al.[2] and Teufel et al.[32] and shown to be expressed in a range of tissues, including heart, brain, lung, liver, kidney, and testis. Importantly, *FOXP4* is not only expressed in adult tissue, but also during different stages of development of, e.g., the heart, lungs, gut, and skeleton, where it has been shown to play important functional roles[2,32–35]. This widespread expression pattern, in combination with the large number of transcriptional targets and protein–protein interactions known for FOXP transcription factors,[10,36–38] could potentially yield a large variety of downstream consequences when FOXP4 functions are compromised. It is thus not surprising that we found a broad range of associated phenotypes in individuals with likely pathogenic *FOXP4* variants, including growth deficits, developmental delays and a spectrum of associated congenital abnormalities. Although caution is warranted given the limited cohort size of our study, variants in *FOXP4* seem to be associated with certain phenotypic features (e.g., vertebral abnormalities and congenital diaphragmatic hernia) that appear distinct from those observed in individuals carrying variants in *FOXP1* or *FOXP2*. Congenital anomalies are not a common finding in individuals with pathogenic *FOXP2* variants[39], and in *FOXP1*-associated disorder different abnormalities are recurrently reported, such as congenital heart defects or kidney abnormalities[40,41] (Table S3). Variants in *FOXP3* are not associated with a neurodevelopmental disorder phenotype, and were thus not included in this phenotypic comparison. Future studies will establish how the distinctive FOXP expression patterns, together with differences in profiles of cofactors and downstream targets in the relevant tissues, contribute to the different phenotypes associated with haploinsufficency of each transcription factor.

Of note, in *FOXP1*- and *FOXP2*-related disorders, expressive speech problems are a prominent feature[40], and the contributions of these regulatory factors to the development and function of relevant neural circuits are extensively studied[10,42]. A recent study linking FoxP1/2/4 functions to vocal learning in songbirds suggested that FOXP4 should also be considered as a candidate for involvement in vocal disorders[43]. Indeed, all individuals with likely pathogenic FOXP4 variants in our study had delayed speech/language development, with expressive problems prominently present. As FOXP1, FOXP2, and FOXP4 show partially overlapping coexpression in various different regions of the developing brain[10], further research is needed to delineate if loss-offunction of FOXP4 directly impairs speech/language development, or whether secondary disruption of FOXP1 and/or FOXP2 function via heterodimerization with dysfunctional FOXP4 could play a role as well.

In conclusion, through clinical characterization and functional assays, we implicate heterozygous *FOXP4* variants in a neurodevelopmental disorder with mild developmental delays, most prominently in the speech/language domain. The disorder shows variable expressivity: a broad spectrum of associated features is present in a subset of individuals and includes short stature, macrocephaly, congenital diaphragmatic hernia, vertebral abnormalities, ptosis, and cryptorchidism. As several congenital abnormalities are recurrently observed in our patients with likely pathogenic variants, and developmental delays can be mild, the possibility of *FOXP4* involvement should not only be considered in individuals with neurodevelopmental disorders but also in cohorts of individuals with multiple congenital abnormalities, in particular, congenital diaphragmatic hernia and/or vertebral abnormalities.

Acknowledgements

We thank all included individuals and their families for their contribution, and we specifically thank the parents of individual 1 for contacting us about the FOXP4 variant found in their child, which was the starting point of this research project. Funding was provided by the Netherlands Organization for Scientific Research (NWO) Gravitation Grant 24.001.006 to the Language in Interaction Consortium (L.S.B., S.E.F., and H.G.B.), the Netherlands Organization for Health Research and Development (ZonMw grant 91718310 to T.K.), the Max Planck Society (S.E.F.), and by P01HD068250 to W.K.C.

References

1. Shu W, Yang H, Zhang L, Lu MM, Morrisey EE. Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors. *J Biol Chem*. **276**, 27488–27497 (2001).
2. Lu MM, Li S, Yang H, Morrisey EE. Foxp4: a novel member of the Foxp subfamily of winged-helix genes coexpressed with Foxp1 and Foxp2 in pulmonary and gut tissues. *Mech Dev*. **119**, S197–S202 (2002).
3. Hori S, Nomura T, Sakaguchi S. Control of regulatory T cell development by the transcription factor Foxp3. *Science*. **299**, 1057–1061 (2003).
4. Hamdan FF, Daoud H, Rochefort D, et al. De novo mutations in FOXP1 in cases with intellectual disability, autism, and language impairment. *Am J Hum Genet*. **87**, 671–678 (2010).
5. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkheaddomain gene is mutated in a severe speech and language disorder. *Nature*. **413**, 519–523 (2001).
6. Chatila TA, Blaeser F, Ho N, et al. JM2, encoding a fork head-related protein, is mutated in X-linked autoimmunity-allergic disregulation syndrome. *J Clin Invest*. **106**, R75–R81 (2000).
7. Takahashi K, Liu FC, Hirokawa K, Takahashi H. Expression of Foxp4 in the developing and adult rat forebrain. *J Neurosci Res*. **86**, 3106–3116 (2008).
8. Rousso DL, Pearson CA, Gaber ZB, et al. Foxp-mediated suppression of Ncadherin regulates neuroepithelial character and progenitor maintenance in the CNS. *Neuron*. **74**, 314–330 (2012).
9. Tam WY, Leung CK, Tong KK, Kwan KM. Foxp4 is essential in maintenance of Purkinje cell dendritic arborization in the mouse cerebellum. *Neuroscience*. **172**, 562–571 (2011).
10. Co M, Anderson AG, Konopka G. FOXP transcription factors in vertebrate brain development, function, and disorders. *Wiley Interdiscip Rev Dev Biol*. **9**, e375 (2020).
11. Charng WL, Karaca E, Coban Akdemir Z, et al. Exome sequencing in mostly consanguineous Arab families with neurologic disease provides a high potential molecular diagnosis rate. BMC Med Genomics. 2016;**9**:42.
12. Longoni M, High FA, Qi H, et al. Genome-wide enrichment of damaging de novo variants in patients with isolated and complex congenital diaphragmatic hernia. *Hum Genet*. **136**, 679–691 (2017).
13. Jin ZB, Wu J, Huang XF, et al. Trio-based exome sequencing arrests de novo mutations in early-onset high myopia. *Proc Natl Acad Sci USA*. **114**, 4219–4224 (2017).
14. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*. **542**, 433–438 (2017).
15. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat*. **36**, 928–930 (2015).
16. denovo-db. http://denovo-db.gs.washington.edu. Accessed 2 February 2020.
17. Castor EDC. Castor electronic data capture. https://castoredc.com. 2019. Accessed 28 August 2019.
18. Estruch SB, Graham SA, Deriziotis P, Fisher SE. The language-related transcription factor FOXP2 is post-translationally modified with small ubiquitin-like modifiers. *Sci Rep*. **6**, 20911 (2016).
19. Deriziotis P, Graham SA, Estruch SB, Fisher SE. Investigating protein–protein interactions in live cells using bioluminescence resonance energy transfer. *J Vis Exp*. **87**, 51438 (2014).
20. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46**, W296–W303 (2018).
21. Golson ML, Kaestner KH. Fox transcription factors: from development to disease. *Development*. **143**, 4558–4570 (2016).
22. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat*. **21**, 577–581 (2003).
23. Wiel L, Baakman C, Gilissen D, Veltman JA, Vriend G, Gilissen C. MetaDome: pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. *Hum Mutat*. **40**, 1030–1038 (2019).
24. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**, 434–443 (2020).
25. Li S, Weidenfeld J, Morrisey EE. Transcriptional and DNA binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. *Mol Cell Biol*. **24**, 809–822 (2004).
26. Stroud JC, Wu Y, Bates DL, et al. Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*. **14**, 159–166 (2006).
27. Sollis E, Graham SA, Vino A, et al. Identification and functional characterization of de novo FOXP1 variants provides novel insights into the etiology of neurodevelopmental disorder. *Hum Mol Genet*. **25**, 546–557 (2016).

28. Vernes SC, Nicod J, Elahi FM, et al. Functional genetic analysis of mutations implicated in a human speech and language disorder. *Hum Mol Genet*. **15**, 3154–3167 (2006).
29. Saleem RA, Banerjee-Basu S, Berry FB, Baxevanis AD, Walter MA. Structural and functional analyses of disease-causing missense mutations in the forkhead domain of FOXC1. *Hum Mol Genet*. **12**, 2993–3005 (2003).
30. Berry FB, Tamimi Y, Carle MV, Lehmann OJ, Walter MA. The establishment of a predictive mutational model of the forkhead domain through the analyses of FOXC2 missense mutations identified in patients with hereditary lymphedema with distichiasis. *Hum Mol Genet*. **14**, 2619–2627 (2005).
31. Perumal K, Dirr HW, Fanucchi S. A single amino acid in the hinge loop region of the FOXP forkhead domain is significant for dimerisation. *Protein J.* **34**, 111–121 (2015).
32. Teufel A, Wong EA, Mukhopadhyay M, Malik N, Westphal H. FoxP4, a novel forkhead transcription factor. *Biochim Biophys Acta*. **1627**, 147–152 (2003).
33. Li S, Zhou D, Lu MM, Morrisey EE. Advanced cardiac morphogenesis does not require heart tube fusion. *Science*. **305**, 1619–1622 (2004).
34. Zhao H, Zhou W, Yao Z, et al. Foxp1/2/4 regulate endochondral ossification as a suppresser complex. *Dev Biol*. **398**, 242–254 (2015).
35. Li S, Morley M, Lu M, et al. Foxp transcription factors suppress a nonpulmonary gene expression program to permit proper lung development. *Dev Biol*. **416**, 338–346 (2016).
36. Oswald F, Kloble P, Ruland A, et al. The FOXP2-driven network in developmental disorders and neurodegeneration. *Front Cell Neurosci*. **11**, 212 (2017).
37. Estruch SB, Graham SA, Quevedo M, et al. Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. *Hum Mol Genet*. **27**, 1212–1227 (2018).
38. Vernes SC, Spiteri E, Nicod J, et al. High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. *Am J Hum Genet*. **81**, 1232–1250 (2007).
39. Reuter MS, Riess A, Moog U, et al. FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. *J Med Genet*. **54**, 64–72 (2017).
40. Siper PM, De Rubeis S, Trelles MDP, et al. Prospective investigation of FOXP1 syndrome. *Mol Autism*. **8**, 57 (2017).
41. Bekheirnia MR, Bekheirnia N, Bainbridge MN, et al. Whole-exome sequencing in the molecular diagnosis of individuals with congenital anomalies of the kidney and urinary tract and identification of a new causative gene. *Genet Med*. **19**, 412–420 (2017).
42. den Hoed J, Fisher SE. Genetic pathways involved in human speech disorders. *Curr Opin Genet Dev*. **65**, 103–111 (2020).
43. Norton P, Barschke P, Scharff C, Mendoza E. Differential song deficits after lentivirus-mediated knockdown of FoxP1, FoxP2, or FoxP4 in area X of juvenile zebra finches. *J Neurosci*. **39**, 9782–9796 (2019).
44. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of highquality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. **7**, 539 (2011).

Supplementary Materials and Methods

Immunoblotting

Whole-cell lysates were collected 24 hours post-transfection, as previously described1. The membrane was probed with 1:8000 mouse anti-EGFP (Clontech) in 1% milk in PBS-T overnight, followed by incubation with 1:3000 HRP-conjugated goat anti-mouse (Bio-Rad) in 1% milk at room temperature for 1h. After visualization using the Novex ECL Chemiluminescent Substrate Reagent Kit (Invitrogen) and the ChemiDoc XRS+ System (Bio-Rad), the blot was stripped for 20 minutes in Re-blot Plus Strong stripping solution (Millipore) and blocked in 5% milk in PBS-T for one hour. This was followed by incubation with 1:10,000 mouse anti-beta-actin (Sigma) for 1.5 hour, and incubation with 1:3000 HRP-conjugated goat anti-mouse (Bio-Rad) for one hour.

Table S1:

	Missense variants							Truncating variant
			Individual 1	**Individual 2**	**Individual 3**	**Individual 4**	**Individual 5**	**Individual 6**
Baseline details								
Type of sequencing performed	research WES	research WES	diagnostic WES	diagnostic WES	diagnostic WES	diagnostic WES	research WES	research WGS (but variant confirmed in CLIA lab environment)
Variant already published (PubMedID)	yes (PMID 28135719)	yes (PMID 28373534)	no	no	no	no	yes (PMID 28303347)	reported in ClinVar
Variant details ***Variant annotation (isoform 1; NM_001012426.1)***								
cDNA	c.818C>T	c.1286C>T	c.1508A>G	c.1540G>A	c.1540G>A	c.1549C>A	c.1553A>G	c.193del
protein effect	p.(Ser273Phe)	p.(Ser429Phe)	p.(Tyr503Cys)	p.(Ala514Thr)	p.(Ala514Thr)	p.(His517Asn)	p.(Asn518Ser)	p.(Gln65Serfs*20)
Other variant characteristics								
gDNA location (hg19)	g.41555196	g.41557837	g.41559032	g.41562611	g.41562611	g.41562620	g.41562624	g.41533691
heterozygous / homozygous	heterozygous	heterozygous	heterozygous	heterozygous	heterozygous	heterozygous	heterozygous	heterozygous
inheritance	de novo	de novo	de novo	de novo	de novo	de novo; confirmed	de novo	Not inherited from mother; status in father unknown
other SNVs of interest?	not known	hemizygous pathogenic variant in CACNA1F (NP_004174.2; p.(Arg1060Trp))	no	no	no	no	no	no
other CNVs of interest?	not known	not known	Absence of heterozygosity involving approximately 4.1% of the genome.	no	no	no	no	no

Pathogenicity predictions								
CADD	22.8	25.1	31	26.8	26.8	27.3	26.3	NA
PolyPhen	0.257	0.608	0.997	0.989	0.989	0.969	1	NA
SIFT	0.03	0.01	0	0	0	0	0	NA
Regional Tolerance (MetaDome)	1	0.97	0.12	0.06	0.06	0.15	0.21	NA
Phenotype details General								
Gender	female	male	female	male	female	male	male	male
Growth parameters								
Age at last examination			5y11m	8y3m	7y10m	16y	5y0m	1y9m
Weight			P65	P53	<P1	P90	P35	10.6kg (P50-P75 for height)
Height			P97	P39 (short stature in history; received growth hormone for multiple years)	<P1	<P1	P25	79cm (<P3 for age at WHO chart)
Head circumference			>P97	P99	P25-P50	>P98	P50-P75	35cm at birth (P50-P75)
Pregnancy & birth								
Pregnancy			normal	IUGR	IUGR. Partial placenta abruption at 18 weeks	normal	Hernia Diaphragmatica.	questionable vanishing twin (2 gestational sacs at first ultrasound). Mother was taking methimazole PTP prior to pregnancy for hyperthyreoidism, switched to PTU during pregnancy. Mother daily smoker. Polyhydramnios.
Gestational age at birth			39w	38w	38w	39w	38-42w	37w;6d

Birth weight	4050g	2985g	2350g	2722g	2630g	3100g
Development & Behaviour						
Motor development	delayed	delayed motor development. Required AFOs and PT/OT for ambulation.	yes,mildly delayed. Received PT and OT	normal motor development	normal, not delayed	delayed. gross motor delay, impaired functional mobility.
Walking independently	26m	20m	15m	14m	not known	unknown
Speech development	speech delays	speech delays	delayed speech/ language development	speech delays	delayed speech/ language development	delayed
First words	26m	12m	12m	26m	unknown	unknown
Speech therapy in history	yes	yes, still in speech therapy	yes	yes	yes	yes
Formal language disorder	Yes, expressive language disorder.	Yes, expressive language disorder.	unknown	not currently	unknown	no
Speech/language abnormalities	Had one word (Dad) at 26 months of age. Articulation/intelligibility problems. At 3 years of age, her speech rapidly accelerated.	He was delayed in expressive language and had minimal speech until 12 months but was signing well and had advanced receptive language starting at 9 months. He was in speech therapy from 12 months until the end of 3rd grade (for articulation and apraxia).	expressive language delays. Auditory processing disorder.	had speech delays, first words at age 26 months.	speech/language development delays, received speech therapy. Now speaks in full sentences	May 2019: receptive and expressive language skills below average when compared to age matched peers.
Current speech	She is now speaking in full and complex sentences.	Still has a slight lisp.	Mainly expressive language difficulty	normal	full sentences	July 2019: patient doesn't identify between common nounse but does utilize the point gesture for requesting. Has about 20 words.

Intellectual disability	Mild ID	no	no	Mild ID	no	not known
IQ (if known)	not known	not known	not known	69 (WISC-V)	not known	not known
Behavioural/ psychiatric problems?	no	no	social pragmatic disorder	no	no	does not understand no.
ASD?	no	no	no	no	no	no
ADHD/ADD?	no	no	ADHD	no	no	no
Neurology						
Hypotonia	yes, present at birth. Now mild.	Significant hypotonia as an infant. Resolution in early childhood (around age 3-3.5y)	Infant hypotonia	no	no	yes (july 2019). Shoulders and shoulder girdle definite decreased muscle tone.
Epilepsy	no (but history of febrile seizures)	no	no	no	no	no
EEG	normal	not performed	not performed	normal	not performed	unknown
Brain MRI	normal (age 26m)	not performed	yes (at age 7y): prominent pituitary with convex superior border. Narrowing of the isthmus of the corpus callosum, which can be seen with callosal dysgenesis.	normal	not performed	yes (at age 8 months), showed prominent CSF spaces, may be related to previous ECMO treatment with possible mild periventricular leukomalacia
Other	Sensory concerns with certain textures, receives food therapy.	no	Frequent headache	no	no	Failure to thrive in infancy. Trunk and pelvic weakness, impaired balance and coordination, limited activity tolerance.
Other						
Congenital diaphragmatic hernia	no	no	no	no	yes (left)	yes

Skeletal abnormalities		Long, thin fingers (could be familial). Pectus excavatum.	Short stature, received growth hormone for multiple years. Uneven scapula (L higer than R), normal spine films.	Short stature, delayed bone age. Hypermobile joints with associated pain. Left talocalcaneal coalition.	cervical spine abnormalities: hypoplasia of the clivus, mild platybasia, retroflexion of the dens, incomplete fusion of anterior arch of C1, absence of posterior arch of C1 and posterior elements of C2, hypoplasia of posterior elements of C3	C1 vertebra was fused to skull	unknown
Vision	severe myopia	normal	Ptosis (surgery at age 9m, revision needed in next few years). Myopia, astigmatism, exotropia.	Ptosis (right side) and strabismus. Wears glasses.	normal	strabismus	normal
Hearing		normal	normal	normal	Mild hearing loss left ear	normal	normal
Cardiac abnormalities		no	no (normal ultrasound)	no (normal ultrasound)	unknown (no ultrasound made)	no (unknown if ultrasound was made)	no (unknown if ultrasound was made)
Gastrointestinal abnormalities		no	no	no	no	no	episodic vomiting (copious clear vomiting), gastroparesis; intestinal pseudoobstruction; mild malnutrition, on total parenteral nutrition (TPN); dysphagia.
Urogenital abnormalities		no	cryptorchidism	no	no	cryptorchidism (left)	Cryptorchidism. Renal ultrasound: calculus in mid-portion of left kidney.
Endorcrine/ immunological/ metabolic abnormalities		no	As part of evaluation for failure to thrive/short stature: normal thyroid and immune workup	Normal endocrine workup (evaluated for short stature)	no	no	no

Facial dysmorphisms	Narrow palpebral fissures, widening of inner canthi. Intercanthal folds present with mild reddish discoloration under the eyes. Long nose with flattened nasal root. Long philtrum. Low anterior hairline.	Tented eyebrows, shortened eyelids with bilateral ptosis. Crossbite, high arched palate, small teeth. Dolicocephaly. Abnormal posterior hairline. Narrowed facies, prominent and high forehead.	ptosis (right), long palpebral fissures and eyelashes. Posteriorly rotated ears, overfolded helix. Primary teeth hypoplasia.	short neck, mildly coarsened facies,fleshy earlobes, gingival hyperplasia	no	flat nasal bridge.
Skin/hair/nail abnormalities	no	palmar hidrosis, hirsutism	no	no	no	no
Other	Respiratory illnesses require a prolonged recovery.	Failure to thrive	Reported a history of pancreatitis at the age of 7 years.	gingival hyperplasia	plagiocephaly (required helmet)	Failure to thrive in infancy. Hypertension. Wide set hypoplastic nipples. Multiple hospitalizations for fever with central venous line. Extensive infectious workup performed with no source of fever identified.
Family members?	older sister who did not talk until age 2y, walked at 15m. Rapid catch-up, currently in normal classes and no therapies needed.	Two younger brothers with FPIES, Pat great uncle with Crohns	11y old brother had seizure-like activity, further details not known	13 year old full brother is healthy. Parents are healthy.	not affected	not known

Table S2: Primer sequences

Primer	Sequence (5’-3’)
FOXP4 cloning with BamHI restriction site (forward)	aggatcctggtggaatctgcctcggagac
FOXP4 cloning with XbaI restriction site (reverse)	ctctagattaggacagttcttctcccggca
Site-directed mutagenesis Y503C (forward)	caccaggatgttcgcctgtttccgcagaaacactg
Site-directed mutagenesis Y503C (reverse)	cagtgtttctgcggaaacaggcgaacatcctggtg
Site-directed mutagenesis N518S (forward)	acgccgtgcgccacagcctcagcc
Site-directed mutagenesis N518S (reverse)	ggctgaggctgtggcgcacggcgt
Site-directed mutagenesis S429F (forward)	ccctggcctgggctttgcctccctg
Site-directed mutagenesis S429F (reverse)	cagggaggcaaagcccaggccaggg
Site-directed mutagenesis Q65Sfs*20 (forward)	gagcctgttgctgctgaagtgcagcagctc
Site-directed mutagenesis Q65Sfs*20 (reverse)	gagctgctgcacttcagcagcaacaggctc
Site-directed mutagenesis S273F (forward)	gtctcaccccccctcttccaccataccctgc
Site-directed mutagenesis S273F (reverse)	gcagggtatggtggaagagggggggtgagac
Site-directed mutagenesis A514T (forward)	gccacctggaagaacaccgtgcgccac
Site-directed mutagenesis A514T (reverse)	gtggcgcacggtgttcttccaggtggc
Site-directed mutagenesis H517N (forward)	gaacgccgtgcgcaacaacctcagcct
Site-directed mutagenesis H517N (reverse)	aggctgaggttgttgcgcacggcgttc

Table S3: Comparison of phenotypes associated with variants in *FOXP1*, *FOXP2* and *FOXP4*

	FOXP4	FOXP1	FOXP2
Short stature (≤P3)	+	+	-
Tall stature (≥P97)	+	-	+
Macrocephaly (≥P97)	+	+	+
Delayed motor development	+	+	+
Delayed speech development	+	+	+
Intellectual disability	+	+	-
Hypotonia	+	+	-
Congenital diaphragmatic hernia	+	-	-
Cervical spine abnormalities	+	-	-
Ptosis	+	+	-
Strabismus	+	+	+
Cryptorchidism	+	+	-
Kidney abnormalities	-	+	-
Genital abnormalities	-	+	-
Congenital heart defect	-	+	-
Joint contractures/arthrogryposis	-	+	-

The phenotypic features of individuals with FOXP4 variants in our study, in comparison with phenotypes reported in individuals with pathogenic *FOXP1* or *FOXP2* variants. For the analysis of common *FOXP1*- and *FOXP2*-associated features, the following cohort studies were used: Bekheirnia et al.2017[2], Le Fevre et al. 2013[3], Reuter et al. 2017[4], Siper et al. 2017[5], Sollis et al. 2016[6] and Sollis et al. 2017[7]. For *FOXP1* and *FOXP2*, a ‘+’ was scored if this feature was reported in two unrelated individuals in the studies mentioned. As a result, some phenotypes annotated with a ‘+’ are only present in a small subset of individuals with *FOXP1*- or *FOXP2*-associated disorder.

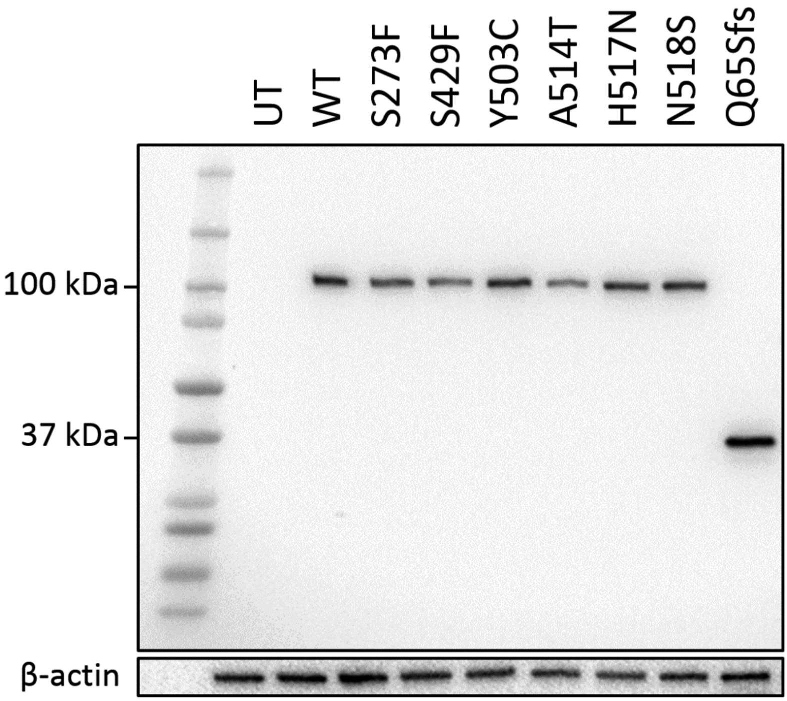


Figure S1: Immunoblot analysis of overexpression constructs

Western Blot of HEK293T/17 cells in which FOXP4-YFP constructs used for functional assays were overexpressed. UT = untransfected, WT= wild-type. Expected molecular weight of wild-type FOXP4-YFP and FOXP4-YFP with missense variants: ~102 kDa. Expected molecular weight of Q65Sfs*20 variant: ~38kDa. All different expressed YFP-fusion proteins are present at the expected molecular weights. The immunoblot was stripped and re-probed with beta-actin as a loading control.

Supplementary References

1. Snijders Blok L, Kleefstra T, Venselaar H, et al. De Novo Variants Disturbing the Transactivation Capacity of POU3F3 Cause a Characteristic Neurodevelopmental Disorder. *Am J Hum Genet.* **105**(2):403-412 (2019).
2. Bekheirnia MR, Bekheirnia N, Bainbridge MN, et al. Whole-exome sequencing in the molecular diagnosis of individuals with congenital anomalies of the kidney and urinary tract and identification of a new causative gene. *Genet Med.* **19**(4):412-420 (2017).
3. Le Fevre AK, Taylor S, Malek NH, et al. FOXP1 mutations cause intellectual disability and a recognizable phenotype. *Am J Med Genet A.* **161A**(12):3166-3175 (2013).
4. Reuter MS, Riess A, Moog U, et al. FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. *J Med Genet.* **54**(1):64-72 (2017).
5. Siper PM, De Rubeis S, Trelles MDP, et al. Prospective investigation of FOXP1 syndrome. *Mol Autism.* **8**:57 (2017).
6. Sollis E, Graham SA, Vino A, et al. Identification and functional characterization of de novo FOXP1 variants provides novel insights into the etiology of neurodevelopmental disorder. *Hum Mol Genet.* **25**(3):546-557 (2016).
7. Sollis E, Deriziotis P, Saitsu H, et al. Equivalent missense variant in the FOXP2 and FOXP1 transcription factors causes distinct neurodevelopmental disorders. *Hum Mutat.* **38**(11):1542-1554 (2017).

7

Chapter 7

Speech-language profiles in the context of cognitive and adaptive functioning in SATB2-associated syndrome

L. Snijders Blok*, Y.M. Goosen*, L. van Haaften, K. van Hulst, S.E. Fisher,
H.G. Brunner, J.I.M. Egger, T. Kleefstra

** These authors contributed equally*

A modified version of this chapter is published in:
Genes, Brain and Behavior 2021
doi:10.1111/gbb.12761

Abstract

SATB2-associated syndrome (SAS) is a neurodevelopmental disorder caused by heterozygous pathogenic variants in the *SATB2* gene, and is typically characterized by intellectual disability and severely impaired communication skills. The goal of this study was to contribute to the understanding of speech and language impairments in SAS, in the context of general developmental skills and cognitive and adaptive functioning. We performed detailed oral motor, speech and language profiling in combination with neuropsychological assessments in 23 individuals with a molecularly confirmed SAS diagnosis: 11 primarily verbal individuals and 12 primarily nonverbal individuals, independent of their ages. All individuals had severe receptive language delays. For all verbal individuals, we were able to define underlying speech conditions. While childhood apraxia of speech was most prevalent, oral motor problems appeared frequent as well and were more present in the nonverbal group than in the verbal group. For seven individuals, age-appropriate Wechsler indices could be derived, showing that the level of intellectual functioning of these individuals varied from moderate-mild ID to mild ID-borderline intellectual functioning. Assessments of adaptive functioning with the Vineland Screener showed relatively high scores on the domain ‘daily functioning’ and relatively low scores on the domain ‘communication’ in most individuals. Altogether, this study provides a detailed delineation of oral motor, speech and language skills and neuropsychological functioning in individuals with SAS, and can provide families and caregivers with information to guide diagnosis, management and treatment approaches.

Introduction

The introduction of new DNA sequencing technologies (next generation sequencing) has rapidly improved the identification of genes of which high-penetrance disruptive variants can cause neurodevelopmental disorders. Amongst the most commonly affected genes in neurodevelopmental disorders is *SATB2*[1]. The neurodevelopmental disorder associated with pathogenic variants in this gene is known as *SATB2*-associated syndrome (SAS).

SAS presents with marked craniofacial dysmorphisms, intellectual disability (ID), developmental delay, as well as generally restricted or absent speech and severely impaired communicative skills[2]. Patients with SAS often use communication methods other than (or in addition to) spoken language, such as gestures, sign language and/or augmentative and alternative communication (AAC) devices. In addition to speech problems, other features related to oral motor skills or oral abnormalities are common, including cleft palate, teeth anomalies, drooling, and feeding problems[2].

SAS is caused by heterozygous disruptions of the *SATB2* gene. These are mostly variants with a clear loss-of-function effect (frameshift and nonsense variants), but missense variants, variants predicted to affect splicing and copy number variants are reported as well[3]. The SATB2 protein is a transcription factor with important roles in cortical development[4]. One could hypothesize that loss-of-function of *SATB2* might disproportionately affect the development of higher cognitive functions, such as attention, memory, and executive functioning. While speech problems are prominent in SAS, there is limited information about mechanisms underlying the oral motor, speech and language impairments observed in affected individuals, other than one recent study on the assessment of speech and language phenotypes in a SAS cohort[5]. That study found that individuals with SAS generally show prominent language impairments, childhood apraxia of speech, and various oral motor problems, including hypernasal resonance, pharyngeal phase dysphagia and drooling[5].

The current study aimed to contribute to the understanding of speech and language abnormalities in SAS in the context of general developmental capacities and cognitive and adaptive functioning. The study design included a detailed characterization of oral motor, speech and language profiles combined with neuropsychological testing in 23 individuals with a molecularly confirmed diagnosis of SAS.

Methods

General study design and data collection

Study design

This study has an observational and cross-sectional study design and was approved by the medical research and ethics committee Arnhem-Nijmegen (CMO Arnhem-Nijmegen; study number NL64562.091.18). All study procedures were in line with the principles of the Declaration of Helsinki. Recruitment and inclusion for the study took place between April and November 2019. After inclusion and the informed consent procedure, individuals were invited for two testing visits within the Radboud University Medical Center in Nijmegen, the Netherlands: one visit with one of the two speech language therapists (SLTs), and one visit with a healthcare psychologist. During one of these two visits, a clinical geneticist in training collected details on medical history and growth parameters. In addition to this, parents and/or caregivers were asked to fill in standardized questionnaires about the patient with SAS. Data collection finished in March 2020.

Individuals

Individuals with SAS from the Netherlands and Belgium were recruited via the Dutch SAS family support group or via the Clinical Genetics department where their SAS diagnosis was established. In order to be eligible to participate in the study, individuals had to meet all of the following three criteria: (1) established molecular diagnosis of *SATB2*-associated syndrome, (2) age of at least two years old at time of testing, and (3) raised in a Dutch-speaking family with Dutch as first language. There was one exclusion criterion: Individuals with SAS who also had another molecular diagnosis that likely contributed to their developmental phenotype were excluded from participation. In total, 23 individuals were included for participation in the study.

General data collection

Data on developmental and medical history were collected via medical file notes and a standardized medical history during one of the visits. Growth parameters were measured during the visit or, if this was not possible, derived from recent measurements in another context. All official molecular test reports with the *SATB2* diagnosis were collected, and variant details were converted into standardized nomenclature using hg19 as a reference genome, and NM_001172509.1 (*SATB2* isoform 1) as the standard transcript. All data were de-identified and stored in a secure and study-specific Castor EDC database[6].

Speech and language profiling

Communication measures

Contingent upon use of words and dominant communication mode, individuals were categorized as primarily nonverbal (an expressive vocabulary of no more than 10 words, communicating nonverbal more than verbal) or verbal (an expressive vocabulary of more than 10 words with speaking as the primary mode of communication).

The Communication Function Classification System (CFCS)[7] was used to rate overall communication abilities. The CFCS is a validated discriminative tool that allows clinicians and parents to categorize children's communication skills into five mutually exclusive levels (CFCS I-V) of everyday communicative function with sending and receiving messages via any modality (e.g., spoken language, sign language, speech-generating electronic devices) with familiar and unfamiliar communication partners.

Utilised forms of augmentative and alternative communication (AAC) were recorded and categorized in (a) unaided – no-tech (gestures, manual signs, facial expressions, vocalizations, verbalizations, body language), (b) aided – low-/light-tech (pictures, objects, photographs, writing, communication boards/books), and (c) aided – high-tech (speech generating devices (SGD), single-message devices and recordable/digitized devices, AAC software that enables dynamic symbol/language representation and that is used with some form of technology hardware such as computer, tablet, or smartphone)[8].

Language measures

Receptive vocabulary was assessed in most individuals with the Dutch version of the Peabody Picture Vocabulary test-III[9], yielding a vocabulary quotient. The Schlichting tests for Language Comprehension and Language Production[10,11] were used to measure receptive and expressive language skills. These norm-based standard scores or *Q* scores have a mean score of 100 (SD 15), with a score of 85-115 representing average range performance.

When administration of the Schlichting tests was not possible due limited language and/or understanding, the Dutch Nonspeech Test (NNST)[12] was used. This test comprises a receptive scale and an expressive scale. Scores on both scales were expressed in percentile scores, with a mean score of 50.

Subtests of the Dutch version of the Clinical Evaluation of Language Fundamentals (CELF)[13] were used instead of the Schlichting tests when individuals had a sufficient level of language. The subtests 'concepts and following directions', 'expressive vocabulary', 'recalling sentences', and 'formulating sentences' were administered.

The Q scores and percentile scores of all the language assessments were interpreted as mild (1-1.5 SD below mean), moderate (1.5-2 SD below mean) and severe (>2 SD below mean).

Speech measures

Where children had sufficient speech, a conversational sample was obtained. The observed speech symptoms provided a basis to form a clinical impression of characteristics of different speech disorders, including a phonological delay or disorder, childhood apraxia of speech (CAS), dysarthria or an articulation deficit. Speech characteristics were analysed using Dodd's Model for Differential Diagnosis[14] and protocols for the classification of dysarthria[15].

The intelligibility of speech was measured in primarily verbal individuals using the Dutch version of the Intelligibility in Context Scale (ICS)[16]. This seven-item questionnaire rates the degree to which the patient's speech is understood by different communication partners (parents/life partners, immediate family, extended family, friends, acquaintances, teachers/colleagues, strangers) on a five point scale (1=never, 2=rarely, 3=sometimes, 4=usually, 5=always).

Feeding and oral motor evaluation

A specifically designed questionnaire for problems with swallowing related to different consistencies of food was used in all individuals. It also included questions regarding drooling and dental problems. This semi-structured questionnaire is used in earlier studies where it has demonstrated its usefulness and importance to differentiate dysphagia characteristics[17,18]. Problems with only chewing (refers to problems in the oral phase) and chewing and choking (refers to problems in the oropharyngeal phase) were scored with a five-point scale and recoded into two categories (-) no problems or (+) problems to a certain extent (2 = less than once a day, 3= once every day, 4= several times a day, 5= food is not offered).

Structural or functional impairments of the oral region were assessed with the self-composed Oral-facial Motor Assessment for Children (OMAC). This assessment tool examines oral motor function (e.g., face, lips, tongue, velum, jaw), oral-facial structural integrity (e.g., symmetry, lip seal), strength (e.g., eye closure, lip closure, tongue, jaw) and the saliva swallow (e.g., slurping, swallowing on demand) by observation. Problems with the performance or imitation of the items were scored and recoded in the category (-) no problems and (+) problems to a certain extent.

Neuropsychological assessment

Intellectual and cognitive functioning

For the reliable and valid assessment of intellectual functioning, three Dutch-language variants of the Wechsler intelligence scales were used, depending on the age of the individual. The Wechsler Preschool and Primary Scale of Intelligence Third Edition (WPPSI-III-NL[19]) was used for individuals aged between 2;6 and 7;11 years, the Wechsler Intelligence Scale for Children Fifth Edition (WISC-V-NL[20]) for individuals with chronological ages between 8 and 17;11 years, and the Wechsler Adult Intelligence Scale Fourth Edition (WAIS-IV-NL[21]) for individuals of 18 years and older. The WPPSI-III-NL, WISC-V-NL, and WAIS-IV-NL provide a full scale IQ (FSIQ, *M*= 100, *SD*=15), based on the performance on four (age group 2;6-3;11), seven (age group 4-7;11) and 10 subtests, respectively (WISC-V-NL age range 6-16;11, WAIS-IV-NL age range 16-84;11). Raw scores are converted to Wechsler standard scores (range 1-19) which are used to calculate IQ and index scores. In addition to Full Scale IQ, the WPPSI-III-NL provides a Verbal IQ (VIQ), a Performance IQ (PIQ) and a Processing Speed Quotient (PSQ; only for the age group 4-7;11). The WISC-V-NL provides a Verbal Comprehension Index (VCI), Visual Spatial Index (VSI), and indices for Fluid Reasoning (FRI), working memory (WMI) and processing speed (PSI). The WAIS-IV-NL provides indices for Verbal Comprehension (VCI), Perceptual Reasoning (PRI), Working Memory (WMI) and Processing Speed (PSI). When age appropriate testing was not possible due to limited language and/or understanding, the WPPSI-III-NL was administered. Raw scores were converted into developmental age equivalents ranging from 'below 2;7' to 'above 7;10'. Although test administration was performed according to standard procedures, slight alterations were made to compensate for language problems of the individuals. For instance individuals were allowed to respond using Dutch Sign Language and/or using AAC when verbal responses were required, and extra verbal cues and explanation were given to engage individuals further when non-compliant (i.e. 'testing of limits').

Adaptive functioning

Adaptive behaviour has been described as the combination of conceptual, social, and practical skills acquired to function adequately in daily life[22]. The level of adaptive functioning was measured using the Vineland Screener 0-6 years[23], filled out by parents. This questionnaire is a Dutch screener version of the golden standard Vineland Adaptive Behavior Scales[24] and consists of 72 questions, providing a total score and four domain scores: communication, social functioning, daily functioning, and motor skills. Raw scores were converted to developmental age scores (in months), reflecting the level of adaptive functioning[23].

Behavioural problems

The presence of behavioural problems was measured by parent-based reports, using age-specific versions of the Achenbach System of Empirically Based Assessment[25]: the Dutch versions of the Child Behavior Checklist (CBCL/1,5-5[26] and CBCL/6-18[27]) and the proxy version of the Adult Behavior Checklist (ABCL/18-59)[28]. These parent-based questionnaires consist of 100, 113 and 134 items, respectively, and provide a total score for observed behavioural problems, scales for internalizing (i.e., anxiety, depression and withdrawal) and externalizing (i.e., agressive behaviour, conflict with others/social mores) problems, and several syndrome subscales. In this study, only the syndrome scales were included that were present in all three versions: somatic, anxious, withdrawn, attention and aggression problems. Raw scores were converted to standardized T-scores. For the the total score and internalizing and externalizing scales, a score of 64 and higher is considered to be in the clinical range (i.e. consideration of professional help is warranted), for the syndrome scale the cut-off for a score in the clinical range is a T score of 70[26-28].

Results

Individuals and characteristics

In total, 32 individuals were examined for eligibility to participate in the study. Nine were not included, because the parents/caregivers decided not to participate after being informed about study details (n=6), because the child was not raised with Dutch as first language (n=2) or because the *SATB2* disruption was part of a large microdeletion with many other genes possibly affecting neurodevelopment (n=1). A total of 23 individuals started participation in the study, all of whom completed it; 70% of these individuals were male. The age of individuals at inclusion varied from 2;10 years to 40 years old (median age 11;7). Growth parameters and other baseline characteristics are included in table 1.

Details on the *SATB2* variants in the individuals are included in Table S1. In short, the majority of individuals (21/23; 91%) had a heterozygous single nucleotide variant (SNV) affecting *SATB2*; two individuals (9%) had a *de novo* 2q33.1 microdeletion (table 1). Almost all variants (21/23; 91%) were confirmed to be *de novo,* hence not present in blood-derived DNA of either of the two parents of the individuals. Two individuals were siblings and carried the same *de novo* variant, suggesting germline mosaicism in one of the parents. Constitutive mosaicism was not detectable by Sanger sequencing of parental blood samples. In one individual, the *SATB2* variant was found to be a mosaic variant, and present in 32 of 143 exome sequencing reads (~22%). The age at which the molecular diagnosis of *SATB2* was established in each individual varied between 0;5 years and 44;1 years, with a median of 10;10 years (table 1).

Table 1: Patient characteristics

	Nonverbal (n=12)	Verbal (n=11)	Total (n=23)
General			
Gender (% male/% female)	67%/33%	73%27%	70%/30%
Median age at inclusion in y;m (range)	11;6 (2;11-39;4)	11;7 (5;6-40;9)	11;7 (2;11-40;9)
Genetic diagnosis			
- SNV (%)	92%	91%	91%
- nonsense (%)	50%	18%	35%
- frameshift (%)	25%	36%	30%
- missense (%)	8%	27%	17%
- splice (%)	8%	9%	9%
- CNV (%)	8%	9%	9%
Confirmed *de novo* (%)	83%	100%	91%
Mosaic variant in individual (%)	0%	9%	4%
Median age of molecular diagnosis	8;1 (0;5-44;1)	10;10 (4;0-37;7)	10;10 (0;5-44;1)
Growth parameters			
Mean birth weight (SD)	3570g (446)	3485g (626)	3531g (524)
Mean height corrected for age (SD)	-0.3 SD (1.5)	0.8 SD (1.3)	+0.3 SD (1.4)
Mean weight corrected for age (SD)	-0.3 SD (1.3)	-0.5 SD (1.4)	-0.4 SD (1.4)
Mean head circumference corrected for age (SD)	0.0 SD (0.7)	0.2 SD (0.8)	0.0 SD (0.8)
Neuro/development			
Median age of walking in months (range)	23 (18-42)	23.5 (17-36)	23 (17-36)
Gross motor delays (%)	100%	82%	91%
Fine motor delays	100%	100%	100%
Epilepsy (confirmed)	17%	9%	13%
Other			
Cleft palate (%)	50%	18%	35%
Dental problems (%)	83%	91%	87%
Vision problems (%)	50%	36%	43%
Hearing loss (%)	0%	0%	0%

Communication

A summary of results per individual is included in table 2. Verbal communication was primarily used by 11 individuals (47.8%), whereas 12 individuals were nonverbal (52.2%). As a group, individuals with primarily verbal communication and nonverbal individuals were comparable in terms of chronological age: median age of the verbal group was 11;7 years (range 5;6-40;9 years) and that of the nonverbal group was 11;6 years (range 2;11-39;4).

AAC was used by most individuals (n = 20/23; 87.0%). The most commonly used form of AAC was signs (n = 14/23; 60.87%). Signs were used alone or in combination with other forms of unaided or aided AAC, e.g., vocalizations, gestures, objects, pictures/photographs, communication books, AAC software, and speech generating devices.

On the CFCS, all individuals exhibited problems with reliable communication with unfamiliar partners (CFCS level III, IV, or V). Three individuals (13%) were rated level V (seldom effective sender and receiver even with familiar partners), 15 individuals (65%) level IV (sometimes effective sender and receiver with familiar partners), and 5 individuals (22%) level III (effective sender and receiver with familiar partners). In the verbal group, all individuals were rated with level III or IV, and in the nonverbal group all individuals had level IV or V (Figure 1A).

Language

Receptive language abilities were measured in 21 individuals. Two individuals were not assessed because test procedures were not developmentally appropriate or individuals were not able to be tested. All the tested individuals showed severe receptive language deficits when compared to age-related peers, except for one (individual 5) with a mild deficit. Expressive language could be measured in nine verbal individuals. Eight of them had a severe expressive language deficit, and one had a moderate to severe deficit (individual 10) when compared to age-related peers.

Speech

All individuals, except one, showed difficulties with speech production. For the ten remaining verbal individuals a speech diagnosis could be established. The 12 nonverbal individuals did not produce enough verbal utterances to be able to differentiate between speech diagnoses although speech symptoms could be described. Six of these 12 individuals had no verbal utterances. The other six had an expressive vocabulary of less than 10 words; three of those individuals showed symptoms of phonological delay, and one had symptoms of dysarthria. In the verbal group, the most common speech diagnosis was Childhood Apraxia of Speech (CAS) (n=8). Two of these individuals showed symptoms of CAS only, while the other six showed symptoms of CAS combined with additional speech diagnoses; CAS and phonological delay (n=2/6), CAS and dysarthria (n=1/6), CAS and phonological delay and dysarthria (n=3/6). The described symptoms of CAS were: words are pronounced sound by sound, fluency difficulties, problems with 'automation' of words, difficulties with speaking on demand, difficulties with maximum repetition rate or diadochokineses. Dysarthria was characterised by slow speech, low pitch, hypernasality, and difficulties with respiratory and voice coordination. One verbal individual showed symptoms of dysarthria only and another verbal individual showed symptoms of a phonological delay (i.e., delayed and atypical phonological speech-sound processes) in combination with an articulation deficit (phonetic distortion). A single verbal individual had no characteristics of any speech disorder. For eight verbal individuals the ICS questionnaire was completed. The mean intelligibility score was 3.41 (range: 2.1 – 5). These findings indicate that the primarily verbal individuals in our study population are 'sometimes' to 'usually' understood.

Feeding and oral motor evaluation

Feeding and swallowing problems were common in the total group of individuals with 87% affected (n=20), while in the remaining three individuals no feeding problems were mentioned. In the nonverbal group, all individuals had feeding problems. In the verbal group, 80% exhibited feeding problems. For the feeding problems in the nonverbal group, 25% involved swallowing problems in the oral phase (e.g. chewing problems and overstuffing) and 75% involved the oropharyngeal phase (e.g. choking, aspiration). This is in contrast to the swallowing problems in the verbal group where 87.5% suffered from oral phase problems and only 12.5% showed oropharyngeal phase problems.

Almost half (48%, n=11) of all 23 individuals showed problems with saliva control (drooling). In the nonverbal group (n=12) there were more individuals suffering from drooling (67%, n=8) compared with the verbal group (30%, n=3). It was difficult for the individuals to collect saliva consciously and swallow it on request, only three individuals of the total group (all part of the verbal group) were able to slurp saliva and swallow it on demand.

Although data collection was not complete due to limited developmental capacities and cooperation of the individuals, oral motor functioning (movements of the face, lips, tongue, velum, jaw) was problematic for almost all individuals in the total group, except for two individuals in the verbal group. Oral facial structural integrity was normal in only three individuals in the non-verbal group, in contrast to five individuals in the verbal group. Only two individuals in the verbal group were able to generate strength when executing orofacial movements.

Neuropsychological functioning

Observation of behaviour during testing procedures showed clear differences with regard to task understanding and concentration, both of which likely mediated task compliance that was further hampered in case of increased restlessness. We classified the results of formal neuropsychological testing using Wechsler scales in three groups (Figure 1B): a group in which age appropriate adminstration of Wechsler subtasks was possible (twelve individuals), a group with results of non age appropriate administration of WPPSI-III-NL with individuals eight years and older (eight individuals), and a group in which no formal testing was possible (three individuals).

The first group of individuals consisted of twelve individuals (52%) in which standardized Wechsler scores were derived. In four of these individuals, a complete profile could be established, with notable differences in indices. In three individuals, a single Wechsler Index score based on only non-verbal tasks could be calculated. Based upon the different indices, the level of intellectual functioning in these seven individuals could be classified as moderate to mild ID (n=1), mild ID (n=5) and mild ID to below average (n=1) respectively. In the remaining five individuals, the administered single subtasks were not sufficient to extract indices. Looking at the group of twelve individuals with standardized Wechsler scores, six individuals were classified as verbal, and six were classified as nonverbal.

In the second group, consisting of eight individuals (35%), non-age matched Wechsler administration of several subtasks of the WPPSI-II-NL were derived. These eight individuals had a chronological age of 9;3 to 40;9, and age equivalents calculated based on Wechsler subtask scores ranged from <2.7 years to <7.1 years. Of these eight individuals, five were classified as verbal and three as nonverbal (Figure 1B; Table 1).

The last group consists of three individuals (13%), who were non-eligible for testing in either form, due to lack of understanding and cooperation. These three individuals were all classified as non verbal.

As measured by the Vineland Total score (n=19), a distinction between chronological and developmental age ranges was found: 35-489 months versus 12-68 months, respectively (Figure 1E). One individual (individual 20, 24 years old) obtained the maximum score of 68 months on the Vineland Screener (i.e. representing a ceiling effect), resulting in scores which do not reflect the actual (higher) level of adaptive functioning. When excluding this single case, the highest adaptive functioning score is 59 months. Inspection of the normalized age equivalents for the total group of individuals showed distinct differences in the domain profile, where the level of daily functioning appeared to be relatively high and the level of communication skills relatively low compared to the total score (Figure 1D). When distinguishing between verbal and nonverbal individuals, identical patterns were seen across subdomains (Figure 1D). The overall levels of adaptive functioning seem to be higher in the verbal group compared to the nonverbal group (Figure 1C).

When comparing the results of receptive language tests (converted to age equivalents) with the age equivalents matching the Vineland adaptive functioning total score, the results of these language tests seem to align with the estimated level of adaptive functioning (Figure 1F).

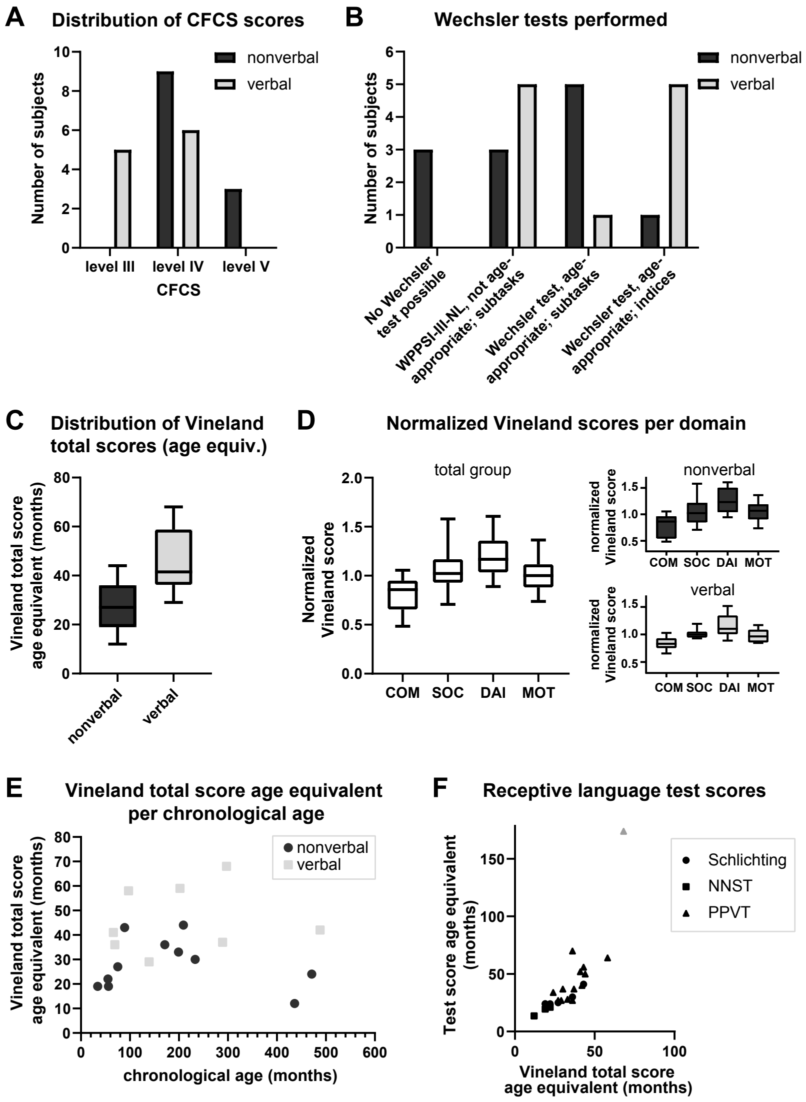


Figure 1: Visual summary of test results

a) Distribution of CFCS test scores in nonverbal and verbal individuals
b) Wechsler tests and subtests performed in nonverbal and verbal individuals.
c) Distribution of age equivalents of Vineland total scores in nonverbal and verbal individuals.
d) Distribution of normalized Vineland scores for each of the four Vineland domains. Individual Vineland scores (age equivalents) per domain were normalized by dividing each score by the total Vineland score (age equivalent) of the same individual: a score of 1.0 indicates that the age equivalent of the domain score is similar to the age equivalent of the total score of this individual.
e) Age equivalents of Vineland total scores (months) obtained in nonverbal and verbal individuals versus chronological age (months).
f) Age equivalents of test scores of three different receptive language (sub)tests, compared to Vineland total score age equivalents. The grey triangle indicates a ceiling effect for the Vineland test score, as the maximum score of 68 months was obtained for this test.

Table 2: Results per individual

Individual	Age inclusion (total in months)	Type of SATB2 variant[a]	Gross motor delays	Fine motor delays	Epilepsy	Vision problems	Cleft palate	Hearing loss	Verbal / Non-verbal	CFCS level	Receptive language impairment	Expressive language impairment	Childhood Apraxia of Speech	Dysarthria	Phonological impairment	Articulation impairment	Feeding difficulties	Problems with oral motor function	Problems with oral facial structural integrity	Problems with oral motor strength	Problems with saliva swallow	Drooling	Wechsler test administered	Wechsler indices[c]	Vineland - Communication skills[d]	Vineland - Social skills[d]	Vineland - Daily functioning skills[d]	Vineland - Motor skills[d]	Vineland - Total score[d]	T-scores in clinical range[e]
1	34	TV	+	+	-	-	-	-	NV	5	Severe	Severe	NA	NA	+	NA	+	NA	NA	NA	NA	+	WPPSI-III-NL		18	30	21	14	19	-
2	55	TV	+	+	-	+	-	-	NV	4	Severe	Severe	NA	NA	NA	NA	+	NA	NA	NA	NA	+	WPPSI-III-NL		19	19	23	30	22	1,2,3,4,5,6
3	56	TV	+	+	-	+	-	-	NV	4	Severe	NA	NA	NA	NA	NA	+	NA	NA	NA	NA	+	WPPSI-III-NL		19	23	18	20	19	1,2,3,4,5
4	64	MD	+	+	-	-	+	-	NV	5	NA	NA	NA	+	NA	NA	+	NA	NA	NA	NA	+	NA		MD	MD	MD	MD	MD	MD
5	66	mTV	+	+	-	-	-	-	V	4	Mild	Severe	+	-	+	-	+	+	NA	NA	NA	+	WPPSI-III-NL	VIQ 55, PIQ 67, PSI 59, FSIQ 55	31	39	46	48	41	-
6	69	MV	+	+	+	-	-	-	V	3	Severe	Severe	+	-	+	-	+	NA	NA	NA	NA	+	WPPSI-III-NL	PIQ 55	29	43	32	40	36	-
7	69	TV	+	+	-	-	-	-	V	3	Severe	Severe	+	-	+	-	+	NA	NA	NA	NA	+	WPPSI-III-NL		MD	MD	MD	MD	MD	-
8	75	TV	+	+	-	-	+	-	NV	4	Severe	Severe	NA	NA	+	NA	+	+	-	+	+	+	WPPSI-III-NL		19	23	35	32	27	-
9	89	MV	+	+	+	+	-	-	NV	4	Severe	NA	NA	+	NA	NA	+	+	NA	NA	NA	-	WPPSI-III-NL	PIQ 55	24	44	53	48	43	-
10	97	MV	-	+	-	-	-	-	V	3	Severe	Severe	-	-	-	-	-	-	-	-	-	-	WISC-V-NL	VCI 70, VSI 67, FRI 69, WMI 59, FSIQ 54	50	61	60	58	58	-
11	111	TV	+	+	-	+	-	-	V	4	Severe	NA	+	-	-	-	-	NA	NA	NA	NA	-	WPPSI-III-NL[b]		MD	MD	MD	MD	MD	MD
12	139	TV	+	+	-	+	-	-	V	3	Severe	NA	-	-	+	+	+	+	-	+	+	-	WPPSI-III-NL[b]		19	28	44	27	29	1,3,5

13	171	SV	+	+	+	-	-	-	NV	4	Severe	NA	NA	NA	+	NA	+	+	-	+	+	-	WPPSI-III-NL[b]		38	48	34	27	36	1,4
14	199	TV	+	+	-	+	+	-	NV	4	Severe	NA	NA	NA	NA	NA	+	+	-	+	+	-	WAIS-IV-NL		16	35	53	30	33	-
15	202	MV	+	+	-	-	-	-	V	4	MD	Severe	+	+	-	-	+	+	-	+	-	-	WISC-V-NL	VCI 68, VSI 45, FRI 61, WMI 55, PSI 56, FSIQ 50	56	61	64	50	59	1,2
16	209	TV	+	+	-	+	+	-	NV	4	Severe	NA	NA	+	NA	NA	+	+	+	NA	NA	+	WPPSI-III-NL[b]		24	44	59	47	44	-
17	233	TV	+	+	-	-	+	-	NV	4	Severe	NA	NA	NA	NA	NA	+	+	+	NA	NA	+	WPPSI-III-NL[b]		16	24	48	33	30	MD
18	256	TV	+	+	-	-	-	-	V	4	Severe	NA	+	+	+	-	-	+	-	+	NA	-	WPPSI-III-NL[b]		MD	MD	MD	MD	MD	MD
19	289	SV	+	+	-	-	+	-	V	4	Severe	NA	+	+	+	-	+	+	+	NA	NA	-	WPPSI-III-NL[b]		28	35	48	37	37	-
20	297	MD	-	+	-	+	+	-	V	4	Severe	NA	-	+	-	-	+	-	-	-	-	-	WAIS-IV-NL	VCI 81, PRI 64, WMI 55, PSI 84, FSIQ 67	70	70	68	58	68	-
21	436	TV	+	+	-	+	-	-	NV	5	Severe	Severe	NA	NA	NA	NA	+	NA	NA	NA	NA	+	NA		11	14	18	11	12	1,2,3,4,5,7
22	471	TV	+	+	-	-	+	-	NV	4	Severe	NA	NA	NA	NA	NA	+	NA	NA	NA	NA	-	NA		23	17	28	29	24	3
23	488	TV	+	+	-	+	-	-	V	3	Severe	NA	+	+	+	-	+	+	+	NA	+	-	WPPSI-III-NL[b]		36	39	57	37	42	-

+ = present, - = not present, NA = not assessed as not developmentally appropriate or not able to be tested, MD = missing data.

[a] TV = truncating variant, MV = missense variant, SV = splice variant, MD = microdeletion, mTV = mosaic truncating variant

[b] Non age-matched administration

[c] VCI=Verbal Comptency Index, VSI=Visual Spatial Index, FRI=Fluid Reasoning Index, PRI=Perceptual Reasoning Index, VIQ=Verbal IQ, PIQ=Performal IQ, WMI=Working Memory Index, PSI=Processing Speed Index, FSIQ=Full Scale IQ

[d] Age equivalent in months

[e] 1=Total, 2=Internalizing, 3=Externalizing, 4=Syndrome scale Attention, 5=Syndrome scale Agression, 6=Syndrome scale Withdrawing, 7=Syndrome scale Somatic.

By reviewing clinical histories with parents and caregivers, sleep disturbances were mentioned in 13 individuals, varying from trouble falling asleep and difficulty staying asleep to increased mobility and/or anxiety. When asked about possible sensory processing problems, these were mentioned in 15 individuals (e.g. high pain threshold, easily overstimulated). Present challenging behaviors were mentioned for five individuals, whereas in a sixth individual these problems had occurred earlier. Regarding psychiatric comorbidity, in three individuals concentration problems were mentioned, and in four individuals autistic traits were mentioned, without meeting formal criteria for a classification of autism spectrum disorder.

Based on CBCL/ABCL total (t-)scores (n=18), behavioural problems within the clinical range were reported in six individuals (33%) of which four are classified as non-verbal and two as verbal. In three individuals (16%, all non-verbal) both internalizing and externalizing problems were reported. In two individuals (11%, one verbal, one non-verbal) only externalizing problems were reported and in one individual (5%, verbal) only internalizing problems. Three individuals (22%, all non-verbal) scored within the clinical range for both attention and aggression problems, of which one (5%) scored within the clinical range on somatic problems, and one on withdrawn behaviour (5%). One individual (5%, non-verbal) scored within the clinical range for attention problems and one (5%, verbal) for aggression problems.

Genotype-phenotype comparison

In terms of genotype-phenotype relations, we looked more specifically to the different types of genetic variants disrupting *SATB2* that were present in our cohort and the associated general developmental and speech-language phenotype. Fourteen individuals had a nonsense or frameshift variant likely causing haploinsufficiency via nonsense-mediated mRNA decay. Two individuals had a variant affecting a canonical splice site and predicted to disrupt correct splicing of the *SATB2* transcript, also likely leading to *SATB2* haploinsufficiency. We thus consider the variants in these 16 individuals to be clear loss-of-function variants. In addition, four individuals had a missense variant in *SATB2*, two individuals had a 2q33.1 microdeletion, and one individual had a mosaic frameshift variant.

Within the group of individuals with a missense variant (n=4), three individuals were classified as primarily verbal (75%) and one individual was primarily nonverbal (25%). In the group of individuals with a loss-of-function single nucleotide variant (n=16), six individuals were classified as primarily verbal (37.5%), and 10 individuals as primarily nonverbal (62.5%). In the group of individuals with a missense variant, the age equivalents of Vineland adaptive functioning total scores ranged from 36 to 59 months (median 39.5), while in the group of individuals with loss of-function variants the range was 12 to 44 months (median 29).

Discussion

With this study we aimed to delineate oral motor, speech, language profiles in the context of cognitive and adaptive functioning in 23 individuals with *SATB2*-associated syndrome, a neurodevelopmental disorder generally characterized by intellectual disability and prominent speech and language problems. We used standardized observations and questionnaires and validated tests to characterize speech/language and oral motor functioning and neuropsychological capacities of primarily verbal (n=11, 47%) and primarily nonverbal (n=12, 52%) individuals with SAS.

Regarding oral motor functioning, almost all individuals (87%) were reported to have feeding problems in addition to speech problems. In the nonverbal group, oropharyngeal problems (chewing with choking) were common (75%), and in the verbal group mainly oral phase problems with chewing and/or overstuffing were seen (87.5%). This finding is in contrast to a recently published study on speech, language and feeding phenotypes in SAS, which reported pharyngeal phase problems in the majority of assessed individuals [5], both non-verbal and verbal. About half of the children in our cohort (48%) suffer from drooling, a problem more present in the nonverbal group. All in all, oral motor problems seem to be a significant problem in the nonverbal group, suggesting that personalized approaches are needed to evaluate and treat oral motor and feeding difficulties.

Using standardized language tests, expressive and receptive language deficits were found in all individuals that could be assessed for such abilities. Almost all had severe receptive language delays, but further discrimination of the individual levels was hampered by floor effects reached using these tests. Age equivalents of receptive language scores correspond to age equivalents of total Vineland scores, suggesting that the Vineland screener is a useful instrument to give an indication of receptive language in clinical practice, with further studies needed to gain insight in the underlying (shared) theoretical constructs. In 10/11 verbal individuals, differentiation of speech symptoms led to diagnoses of speech-related disorders. While childhood apraxia of speech was most common, other diagnoses included phonological delays, dysarthria, and articulation impairment. For *SATB2*-associated disorder, a previous study reported a diagnosis of childhood apraxia of speech in all 40 individuals with enough verbal ability in their SAS cohort [5]. While one might hypothesize that individuals with the same genetic disorder have similar speech and language phenotypes, the results of our detailed diagnostic speech profiling show that even with the same genetic syndrome, divergent speech problems may occur. Subgroups with childhood apraxia of speech, phonological delay, dysarthria and articulation impairment are thought to represent different underlying deficits [29], although such problems might sometimes co-occur. These underlying deficits can be described in terms of problems with phonological encoding, speech motor planning, speech motor programming and speech motor execution[30]. The results of the current study show the need for detailed personalized speech and language assessments in each individual with SAS, since distinct speech problems will benefit from different approaches to intervention.

It is currently unclear which processes underly the absence of speech as a primary mode of communication in the nonverbal group. Based on the results of cognitive, language and oral motor assessments, we would expect these individuals to be able to develop a certain level of speech. As a result, the absence of speech is possibly the result of neurobiological mechanisms involved in the speech process, or behavioural characteristics, but is not simply secondary to cognition, language or oral motor impairments. It is hard to generate hypotheses regarding the specific speech process involved, as we identified several different processes in the verbal group that contributed to impaired speech development. Possibly, the verbal versus nonverbal distinction in SAS is mainly caused by severity of impairments in one or more of the speech processes. It is interesting to note in this context that observation of individuals during assessments showed limited levels of initiation of communication, in addition to lower levels of frustration than would be expected based on the severely limited communication in most individuals.

Generally it is difficult to assess the IQ levels in individuals with ID by using conventional methods that are based on the normal population. Therefore we converted the Wechsler based test scores in corresponding developmental age equivalents, in order to derive useful scores for all or several indices for a subset of individuals. Nonetheless, for the vast majority of the cohort we were able to obtain scores on adaptive functioning. Previous studies using Vineland Adaptive Behaviour Scales have shown that different genetic disorders can give rise to distinctive profiles of adaptive functioning, which might also be partly age-dependent [31-33]. The relative weakness of communication in the adaptive functioning profile observed in the Vineland scores within our study is in line with the findings from direct speech and language assessments, as well as the literature on SAS so far (e.g.[34]). Variations in adaptive functioning domains with relatively strong daily-living skills based on Vineland questionnaires are commonly reported in other neurodevelopmental disorders [33,35,36]. As already shown for other genetic syndromes, the assumption that cognitive functioning is strictly related to all adaptive functioning domains does also not apply to SAS [37]. Classifying an accurate level of ID based on the required equal weighting of intellectual functioning (i.e. IQ) and level of adaptive functioning [38] is therefore a challenge and more in-depth analysis of intellectual functioning and adaptive functioning is required.

In the literature on SAS, behavioural issues have been reported in the majority of individuals, with different forms of challenging behaviour being present, including autistic traits, hyperactivity and aggression [34,39]. In our study, autistic traits were mentioned by parents or caregivers in only four individuals (none meeting requirements for a formal ASD diagnosis), and two individuals (9%) received methylphenidate because of attention problems. Broader behavioural issues in our study cohort were evaluated using CBCL/ABCL questionnaires, and we found one third of assessed individuals to have scores within the clinical range, which seems to be in line with the level of ID and/or verbal proficiency. Clinical range scores did not reflect the level of test cooperation. Growing literature on genetic syndromes from a multidisciplinary perspective (i.e., neuropsychology, psychiatry and clinical genetics) shows that particular behaviours should be interpreted in a wider context in order to understand if, and in what way, they should be regarded as specific to the phenotype [40,41]. Research in KBG syndrome, for instance, showed that social difficulties reported in patients might well be related to (the level of) ID instead of reflecting a specific ASD trait [42]. A longitudinal meta-analysis by Chow et al. [43] shows that receptive language skills in particular have a strong predictive property when it comes to challenging behaviour, and that improving (receptive) language skills can have a mitigating effect on the development of behavioural problems. Although our results do not directly support this link between problems in language and behaviour, it is possible that the relatively low levels of frustration observed in individuals in our study and a related lack of initiation contribute to the severe speech phenotype. Findings like these warrant a broad and strong dimensional approach to clinical assessment using gold standard instruments and a careful consideration of contextual factors to correctly interpret a particular behaviour as part of the SAS phenotype profile [40,41]. Research has also shown that it is necessary to interpret challenging behaviour in ID in relation to contextual variables, in order to establish an effective intervention plan [44].

Different types of heterozygous *SATB2* disruptions were found in the individuals included in our study. While there is some evidence that missense variants of *SATB2* might be associated with milder phenotypes [3], functional characterization of effects of variants in this gene has so far been limited. It is therefore unclear whether missense variants have different effects from the loss-of-function that is assumed for most other variants [3,45]. As *SATB2* encodes a transcription factor that can have pleiotropic effects on multiple different pathways and developmental processes in the brain, it is important to realize that many different factors (e.g. stochastic developmental factors) might ultimately contribute to the phenotypic presentation, even between individuals with identical pathogenic variants.

In addition to individuals with single nucleotide variants affecting *SATB2*, our cohort included two individuals (individual 4 and 20) in which a *de novo* 2q33.1 microdeletion including the *SATB2* gene was reported. Re-evaluation of the original array-CGH report of individual 20 however could not confirm the involvement of the *SATB2* gene with certainty. We therefore performed a CytoScan XON array analysis, which showed that the deletion was located just six kilobases downstream of *SATB2*. Although the breakpoints of the deletion were located outside the coding region of *SATB2*, and thus a loss-of-function effect via haploinsufficiency is unlikely for this individual, positional effects of this deletion on *SATB2* gene expression cannot be excluded.

Our study has some limitations that should be taken into account. First, due to the low prevalence of *SATB2* variants in the population, it is not possible to study a large cohort of affected individuals with the same native language in the same age range. The consequent differences in chronological ages in our cohort, as well as the varying levels of cognitive functioning, made systematic testing using comparable tests more difficult and in some cases impossible, leading to suboptimal data collection. It is also unclear if the chronological age of individuals in this study might have affected the results, as current possibilities on diagnostics, speech therapy and education are very different compared to the situation decades ago. On the other hand, while these limitations are applicable for the study data on a group level, they do not apply for the usability of the data on an individual level, e.g. for intra-individual comparisons in a longitudinal study setting.

Nonetheless, our research can serve as a base for future studies on speech, language, oral motor and cognitive functioning in SAS. Ideally, longitudinal studies should be executed in which children with a SAS diagnosis at a young age are included for early diagnostics on a speech, language and cognitive level, and for subsequent targeted interventions. In addition, for future studies we recommend the inclusion of nonverbal test batteries aimed at specific cognitive domains (i.e. attention, processing speed, executive functioning) as well as general level of intellectual functioning, combined with gold standard proxy instruments, to be able to better define cognitive performance in individuals with SAS of all levels.

In summary, with this study we provide a delineation of speech, language and oral motor skills in indviduals with SAS, combined with emerging data on neuropsychological functioning. While overlapping and highly recurrent features were seen for both the speech and language domain and the adaptive functioning profile, there was also a high variability observed, mainly in severity of features. This study can provide families, speech therapists, psychologists and other caregivers with the necessary information to guide diagnostic and treatment approaches in order to obtain the best functional outcomes in individuals with *SATB2* associated syndrome.

Acknowledgements

We thank all individuals and their families and caregivers for their contribution. Funding for this research was provided by the Netherlands Organization for Scientific Research (NWO) Gravitation Grant 24.001.006 to the Language in Interaction Consortium (L.S.B., S.E.F., and H.G.B.), the Netherlands Organization for Health Research and Development (ZonMw grant 91718310 to T.K.) and the Max Planck Society (S.E.F.). Thanks are extended to Dr. Renée Roelofs, Dr. Linde van Dongen and Dr. Lindsey Ossewaarde for their contribution to the analyses of the neuropsychological assessment data. The authors have no conflicts of interest to declare.

References

1. Deciphering Developmental Disorders S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature.* **519**(7542):223-228 (2015).
2. Zarate YA, Kaylor J, Fish J. SATB2-Associated Syndrome. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews((R)).* Seattle (WA)1993.
3. Zarate YA, Bosanko KA, Caffrey AR, et al. Mutation update for the SATB2 gene. *Hum Mutat.* **40**(8):1013-1029 (2019).
4. Britanova O, de Juan Romero C, Cheung A, et al. Satb2 is a postmitotic determinant for upper-layer neuron specification in the neocortex. *Neuron.* **57**(3):378-392 (2008).
5. Thomason A, Pankey E, Nutt B, Caffrey AR, Zarate YA. Speech, language, and feeding phenotypes of SATB2-associated syndrome. *Clin Genet.* **96**(6):485-492 (2019).
6. Castor EDC. Castor Electronic Data Capture. https://castoredc.com. Published 2019. Accessed August 28, 2019.
7. Hidecker MJ, Paneth N, Rosenbaum PL, et al. Developing and validating the Communication Function Classification System for individuals with cerebral palsy. *Dev Med Child Neurol.* **53**(8):704-710 (2011).
8. (ASHA) AS-L-HA. Augmentative and Alternative Communication. Professional Issues Web site. https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942773&section=Key_Issues Accessed 30-6, 2020.
9. Schlichting L. *Peabody Picture Vocabulary Test, Third Edition Dutch Version (PPVT-III-NL).* Amsterdam: Pearson Assessment and Information B.V.; 2005.
10. Schlichting L, Spelberg Hl. *Schlichting Test voor Taalproductie-II (Schlichting Test for Language Production).* Houten: Bohn Stafleu van Loghum; 2010.
11. Schlichting L, Spelberg Hl. *Schlichting Test voor Taalbegrip (Schlichting Test for Language Comprehension).* Houten: Bohn Stafleu van Loghum; 2010.
12. Zink I, Lambrechts D. *NNST: Nederlandstalige Nonspeech Test: aanpassing en hernormering van de Nonspeech Test van Mary Huer.* Leuven, Belgium: Acco; 2000.
13. Kort W, Schittekatte M, Compaan E. *CELF-4-NL: clinical evaluation of language fundamentals.* Amsterdam: Pearson Assessment and Information B.V; 2008.
14. Dodd B. Differential diagnosis of pediatric speech sound disorder. *Current Developmental Disorders Reports.* **1**(3):189-196 (2014).
15. Morgan AT, Liegeois F. Re-thinking diagnostic classification of the dysarthrias: a developmental perspective. *Folia Phoniatr Logop.* **62**(3):120-126 (2010).
16. Van Doornik A, Gerrits E, McLeod S, Terband H. Impact of communication partner familiarity and speech accuracy on parents' ratings of their child for the Intelligibility in Context Scale: Dutch. *Int J Speech Lang Pathol.* **20**(3):350-360 (2018).
17. van den Engel-Hoek L, Erasmus CE, van Bruggen HW, et al. Dysphagia in spinal muscular atrophy type II: more than a bulbar problem? *Neurology.* **73**(21):1787-1791 (2009).
18. van den Engel-Hoek L, Erasmus CE, Hendriks JC, et al. Oral muscles are progressively affected in Duchenne muscular dystrophy: implications for dysphagia treatment. *J Neurol.* **260**(5):1295-1303 (2013).
19. Hendriksen J, Hurks P. WPPSI-III-NL Nederlandstalige bewerking: Technische handleiding. In: Amsterdam: Pearson Assessment and Information BV; 2009.
20. Wechsler D, Hendriks MP, Ruiter S, Schittekatte M, Bos A. *WISC-V-NL: Nederlandstalige bewerking.* Pearson; 2018.
21. Wechsler D. WAIS–IV–NL Nederlandse bewerking, technische handleiding [The Dutch language version of the WAIS–IV: Administration and scoring manual]. In: Amsterdam: Pearson Assessment and Information BV; 2012.
22. Disabilities AAoIaD. Definition of intellectual disability. https://www.aaidd.org/intellectual-disability/definition Published 2020. Accessed 31-05-2020.
23. Scholte E, Van Duijn E, Dijkxhoorn Y, Noens I, Van Berckelaer-Onnes I. Vineland screener 0–6 years: manual of the Dutch adaptation. *PITS, Leiden.* 2008.
24. Sparrow SS, Balla DA, Cicchetti DV, Harrison PL. Vineland adaptive behavior scales. 1984.
25. Achenbach TM. *The Achenbach system of empirically based assessment (ASEBA): Development, findings, theory, and applications.* University of Vermont, Research Center for Children, Youth, & Families; 2009.
26. Achenbach TM, Rescorla LA. *Manual for the ASEBA preschool forms and profiles.* Vol 30: Burlington, VT: University of Vermont, Research center for children, youth …; 2000.

27. Achenbach TM, Rescorla L. *Manual for the ASEBA school-age forms & profiles: An integrated system of multi-informant assessment.* Aseba Burlington, VT:; 2001.
28. Achenbach TM, Rescorla L. Manual for the ASEBA adult forms & profiles. In: Burlington, VT: University of Vermont, Research Center for Children, Youth …; 2003.
29. Waring R, Knight R. How should children with speech sound disorders be classified? A review and critical evaluation of current classification systems. *Int J Lang Commun Disord.* **48**(1):25-40 (2013).
30. Terband H, Maassen B, Maas E. A Psycholinguistic Framework for Diagnosis and Treatment Planning of Developmental Speech Disorders. *Folia Phoniatr Logop.* **71**(5-6):216-227 (2019).
31. Vermeulen K, de Boer A, Janzing JGE, et al. Adaptive and maladaptive functioning in Kleefstra syndrome compared to other rare genetic disorders with intellectual disabilities. *Am J Med Genet A.* **173**(7):1821-1830 (2017).
32. Klaiman C, Quintin EM, Jo B, et al. Longitudinal profiles of adaptive behavior in fragile X syndrome. *Pediatrics.* **134**(2):315-324 (2014)
33. Butcher NJ, Chow EW, Costain G, Karas D, Ho A, Bassett AS. Functional outcomes of adults with 22q11.2 deletion syndrome. *Genet Med.* **14**(10):836-843 (2012).
34. Zarate YA, Smith-Hicks CL, Greene C, et al. Natural history and genotype-phenotype correlations in 72 individuals with SATB2-associated syndrome. *Am J Med Genet A.* **176**(4):925-935 (2018).
35. Visootsak J, Rosner B, Dykens E, Tartaglia N, Graham JM, Jr. Behavioral phenotype of sex chromosome aneuploidies: 48,XXYY, 48,XXXY, and 49,XXXXY. *Am J Med Genet A.* **143A**(11):1198-1203 (2017).
36. Di Nuovo SF, Buono S. Psychiatric syndromes comorbid with mental retardation: differences in cognitive and adaptive skills. *J Psychiatr Res.* **41**(9):795-800 (2007).
37. Di Nuovo S, Buono S. Behavioral phenotypes of genetic syndromes with intellectual disability: comparison of adaptive profiles. *Psychiatry Res.* **189**(3):440-445 (2011).
38. Tasse MJ, Luckasson R, Schalock RL. The Relation Between Intellectual Functioning and Adaptive Behavior in the Diagnosis of Intellectual Disability. *Intellect Dev Disabil.* **54**(6):381-390 (2016).
39. Zarate YA, Fish JL. SATB2-associated syndrome: Mechanisms, phenotype, and practical recommendations. *Am J Med Genet A.* **173**(2):327-337 (2017).
40. Waite J, Heald M, Wilde L, et al. The importance of understanding the behavioural phenotypes of genetic syndromes associated with intellectual disability. *Paediatrics and Child Health.* **24**(10):468-472 (2014).
41. Bos-Roubos AG, van Dongen L, Verhoeven WM, Egger JI. Genetic Disorders and Dual Diagnosis: Building Clinical Management on Etiology and Neurocognition. In: *Handbook of Dual Diagnosis.* Springer; 57-76 (2020).
42. van Dongen LCM, Wingbermuhle E, van der Veld WM, et al. Exploring the behavioral and cognitive phenotype of KBG syndrome. *Genes Brain Behav.* **18**(4):e12553 (2019).
43. Chow JC, Ekholm E, Coleman H. Does oral language underpin the development of later behavior problems? A longitudinal meta-analysis. *Sch Psychol Q.* **33**(3):337-349 (2018).
44. Lloyd BP, Kennedy CH. Assessment and treatment of challenging behaviour for individuals with intellectual disability: a research review. *J Appl Res Intellect Disabil.* **27**(3):187-199 (2014).
45. Bengani H, Handley M, Alvi M, et al. Clinical and molecular consequences of disease-associated de novo mutations in SATB2. *Genet Med.* **19**(8):900-908 (2017).

Supplementary Data

Table S1 is available online via doi:10.1111/gbb.12761

8

Chapter 8

The GENTOS study: Rationale and design of a prospective cohort study to investigate genetic causes of developmental language disorders

*Lot Snijders Blok, Karen van Hulst, Leenke van Haaften,
Tjitske Kleefstra, Simon E. Fisher, Han G. Brunner*

Abstract

Rationale:

The genetic architecture underlying developmental language disorders (DLD) is poorly understood. It has been established that genetic factors play an important role in the aetiology of DLD, but we still know little about the genes involved, and the proportion of children with DLD that carry a monogenic causal variant (e.g. a rare *de novo* mutation). Different studies in the literature and observations in a clinical genetics outpatient clinic setting point to a model in which there is a role for *de novo* single nucleotide DNA variants. However, so far no trio-based (i.e. screening of proband and both parents) next-generation sequencing studies of a systematically ascertained DLD cohort have been performed.

Objectives:

We determine the diagnostic yield of whole genome sequencing (WGS) in DLD. In addition to determining this diagnostic yield, we aim to gain insights into the types of variants and genetic loci that are involved.

Study design:

The study is designed as a prospective cohort study of fifty children with DLD, in whom WGS will be performed in each proband and both parents. The study population consists of children (4-18 years) with a severe form of DLD (in Dutch *TOS; taalontwikkelingsstoornis*), and a negative family history. WGS data will be analysed on an individual level to identify (possible) disease-causing variants in the probands. Data will subsequently be analysed on a group level, to determine the diagnostic yield and secondary parameters of the study. Inclusion for this study started in March 2020.

Main study parameters and endpoints:

The primary objective of this study is to define the diagnostic yield of WGS in children with a severe DLD and a negative family history. In addition to this, we aim to identify the genes in which variants are found to (possibly) cause DLD, the corresponding molecular pathways, and the type of genetic variation involved. We also aim to identify possible patient subgroups with a significantly different diagnostic yield.

Introduction

Developmental language disorders

A developmental language disorder (DLD) is a neurodevelopmental disorder that is characterized by unexpected problems with language production, comprehension and communication, despite adequate linguistic input in the absence of a clear primary cause like deafness or brain traumas. A DLD can have a major impact on communication skills and development in children, and often persists later in life[1,2]. The prevalence of developmental language disorders amongst children is not exactly known, but a study of pre-schoolers (around 5 years of age) in the USA identified impairments in ~7% of the sample[3]. Among children diagnosed with developmental language disorders, the severity and impact on daily life is extremely variable. Some children only need minor support and cope adequately in regular schools[4]. However, those with more severe forms of disorder typically require intensive treatment and special education, and they may experience problems with participation and functioning in society throughout life[5].

The terminology and criteria used to describe DLD represent an ongoing topic of debate in international diagnostic and research settings[6]. Several different names have been used in this field, including SLI (Specific Language Impairment), and SLCN (Speech, Language and Communication Needs), with results of a recent Delphi survey recommending adoption of the consensus English term DLD. In the Netherlands, since 2014 the label *TOS* (*taalontwikkelingsstoornis*) has been widely used by the various care and education organizations[4]. In this chapter, we use the English term DLD to refer to the condition that meets the accepted Dutch diagnostic criteria for the disorder, as defined by Gerrits & van Niel in 2012[7], and as used in the "Dutch guidelines for speech/language therapy in TOS" published in 2017[8].

Genetics of DLD

The genetic architecture underlying DLD is poorly understood, even though it is well established that genetic factors contribute significantly[9,10]. In children with most major neurodevelopmental disorders (e.g. intellectual disability or autism), genetic testing has become an accepted part of the diagnostic procedure[11]. In the past years, the introduction of next generation DNA sequencing techniques as a genetic diagnostic tool for intellectual disability and autism has dramatically improved the diagnostic yield, and thus also increased our knowledge of the molecular pathology involved. Enhanced testing has also highlighted the prevalence of causative *de novo* mutations (i.e. absent from parents, newly arising in the proband) in these neurodevelopmental disorders[12]. The finding of a *de novo* mutation disrupting a gene is very often the start of a research project characterizing a new neurodevelopmental disorder, and over the last years this has led to the identification of a growing number of genes involved in intellectual disability and autism, with more and more families knowing why their child develops in a way that is different from her/his peers.

But what about DLD? The large genetic cohort studies on intellectual disability and autism stand in stark contrast to the much more limited molecular genetic research so far on speech and language disorders, despite the fact that the latter are also prevalent and crucial for modern society. It has been established from familial clustering and twin studies that inherited factors play an important role in the aetiology of DLD[13], but we still know little about the specific genes involved and how they go awry. So far, *FOXP2* is one of the few well-characterized genes that have been robustly implicated in a specific speech and language disorder[14,15]. We also do not yet know what proportion of children with a DLD have a monogenic cause (e.g. a rare *de novo* mutation), and in what proportion of families the inheritance pattern is multifactorial instead of monogenic. A whole exome sequencing (WES) study in 43 unrelated probands with DLD reported rare variants in known genes associated with developmental disorders and in new candidate genes, but did not use a trio design to facilitate filtering for *de novo* variants[16]. In a whole genome sequencing (WGS) study of a small cohort of children with a specific speech disorder (childhood apraxia of speech), causative *de novo* mutations were found in three of the nine probands in which trio-sequencing was performed[17], suggesting that monogenic causes are quite common for this severe form of speech disorder. Another exome and genome sequencing study reported pathogenic or likely pathogenic variants in 11 of 34 probands with childhood apraxia of speech[18]. Another recent study using high-resolution microarrays in 58 Swedish probands with severe DLD shows an enrichment of rare clinically significant (and in some cases *de novo)* deletions and duplications, again supporting the involvement of monogenic causes[19].

In addition to these findings from the published literature, over the last couple of years several children with severe forms of DLD have been referred to the Human Genetics department at the Radboudumc for genetic testing. In some of these children for whom trio-based exome sequencing was performed in a diagnostic setting, rare *de novo* mutations were found in genes that are known to be implicated in intellectual disability and/or autism. These anecdotal observations point to a model in which there is a role for *de novo* single nucleotide DNA variants in the aetiology of DLD. Crucially, so far this possibility has not been investigated through formal trio-based next-generation sequencing of a systematically ascertained DLD cohort, and the relative contribution of such *de novo* variants remains unknown.

Aims and relevance of this study

The main goal of this study is to gain more knowledge on the use of diagnostic genetic testing by NGS in children with severe forms of DLD. It is very important to know if a child has an isolated DLD, or whether the difficulties are part of a more complex syndrome, in order to give the best support. In clinical practice, we frequently see that initially, only the speech/language problem gets attention, and it is realized (much) later that the problems are more complex. It is currently unclear in a clinical setting if and when genetic testing is indicated in children with DLD, and what the odds are of finding a genetic cause. Currently, some children are referred to clinical genetics centers already with this indication, but many other children are not, and no clear policy or guidelines exist. Possible improvements include more knowledge on the topic, more structure and clarity in referrals, and additionally clear guidelines for requesting genetic diagnostic tests. A prerequisite for such policy changes lies with systematic research, like this GENTOS study.

A second aim of this study is to gain insight into the type of genetic variants and genes involved in the pathogenesis of DLD. Many examples already exist of children with a DLD in whom a causative variant in the DNA is found in a gene already associated with autism or intellectual disability. In some cases even an identical variant is found. The degree of overlap between DLD and other developmental disorders on a molecular level is not yet known, and it is unclear if and what specific (groups of) genes are more specifically involved in the development of these speech/language disorders. More knowledge on the underlying genetic causes can have impact on how children with DLD are classified and treated. At this moment DLD is considered a separate entity, which does not facilitate the recognition of children with underlying genetic causes, or with additional medical problems.

While not the main goal of the GENTOS study, this research project might in the longer term contribute to our understanding of the molecular and neurobiological mechanisms underlying normal and abnormal speech/language development.

Materials and methods

Study design

The GENTOS study is a single-centre prospective cohort study of fifty subjects with a severe form of DLD. WGS will be performed in cases as well as both their parents. The study protocol is approved by the medical research and ethics committee Arnhem-Nijmegen (NL67516.091.19). Figure 1 shows the flowchart of the main study design.

Each trio (proband and parents) is included via an individual intake in the outpatient clinic of the Amalia's Children Hospital at the Radboudumc Nijmegen. Medical and historical data will be collected and stored in a study-specific research database. A summary of the data will be added to EPIC (electronical record system of hospital) as a research note. Blood is collected from the affected child and the unaffected parents. Data from WGS will be analysed per proband-parent trio in order to determine the origin of the identified variants in the child (paternal, maternal or *de novo*). For each patient, an individual research report is made that records variants of interest ("causative" or "possible causative" variants) for that patient, according to guidelines of the American College of Medical Genetics (ACMG) guidelines [20]. The variants of interest will be linked to the individual patient characteristics in the study database. The WGS data will also be analysed for all 50 probands combined. By this cohort analysis, recurrent mutations or recurrently involved genes can be found. Also, a group-wide analysis of diagnostic yield will be performed, including a description of the type and nature of variants found in this cohort, and an analysis to define possible subgroups with higher or lower diagnostic yields.

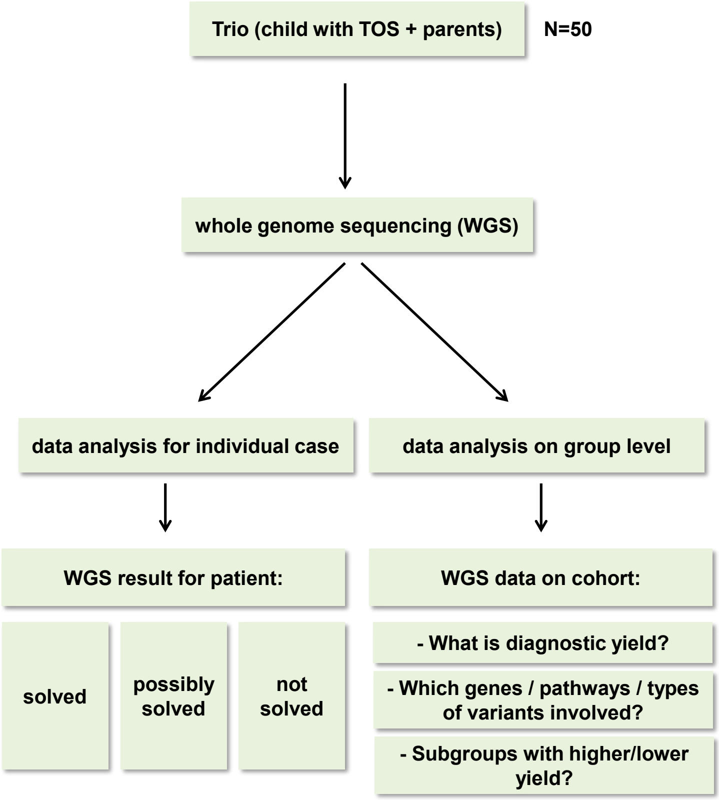

Trio (child with TOS + parents)
N=50
whole genome sequencing (WGS)
data analysis for individual case
data analysis on group level
WGS result for patient:
solved
possibly solved
not solved
WGS data on cohort:
- What is diagnostic yield?
- Which genes / pathways / types of variants involved?
- Subgroups with higher/lower yield?

Figure 1: Flowchart of study design

Participants

Individuals are eligible for participation if they have a severe TOS requiring special education or ambulatory support in a 'Cluster 2 setting' (i.e. specifically aimed at children with hearing or speech/language impairments), are between 4-18 years at the time of inclusion, have speech problems and a nonverbal IQ of at least 70. The parents and siblings of the probands should not have any speech or language disorders, learning difficulties requiring special education or autism spectrum disorders. Detailed inclusion and exclusion criteria are listed in Table 1.

Recruitment and consent

Families can show their interest in participating in the study, by submitting their e-mail address to our study website (www.radboudumc.nl/gentos). This website contains information on the study in line with the study protocol. If parents show their interest, they receive more information on the study (including patient information letter and consent forms) via post and a telephone call with the researcher is planned, in which the study details are discussed (study procedure, consent procedure, inclusion criteria, etc.). An intake appointment in the clinic is planned with the family once signed informed consent has been obtained.

Table 1: Inclusion and exclusion criteria

Inclusion criteria	
1.	Diagnosis TOS (taalontwikkelingsstoornis) given in the Netherlands
2.	Age at inclusion: 4-18y
3.	IQ > 70 on non-verbal test
4.	Indication for education at Cluster 2 school or ambulatory Cluster 2 support
5.	Speech must be affected*
6.	A negative first-degree family history for speech and language disorders, speech therapy, autism spectrum disorders and special education.
Exclusion criteria	
1.	An autism spectrum disorder diagnosis made by a child psychiatrist or GZ psychologist
2.	Next generation sequencing (including WES) has already been performed, or is being performed at the time of inclusion.
3.	One biological parent is not available for genetic testing (or both biological parents are not available)

*This will be determined by one of the speech/language therapists of the study team, based on the subdomain 'speech' in the diagnostic test report from the Audiology Center that is required for a Cluster 2 indication.

Study intake in clinic

The proband and both parents visit the hospital for an intake appointment, which takes place in the Radboudumc Amalia's Children Hospital. During this intake, the medical history of the child and a short history on the family is taken by the clinical researcher under supervision of a clinical geneticist. In addition to the clinical history, a short physical examination of the child is performed, to measure growth parameters and screen for any possible dysmorphic features. Immediately after the clinical intake, a blood withdrawal takes place at the same clinic. A certified nurse of the children's clinic from the Amalia Children's Hospital performs a venipuncture in the child and parents.

Sample collection and WGS

Blood is collected in two EDTA tubes (6,0ml per tube) and one heparin tube (5.0ml). Genomic DNA is isolated from EDTA blood from child and parents separately using the standard conditions of the Genome Diagnostics facility of the Radboudumc. Each sample will be given a DNA-number that will be linked to the study-specific identification number of the proband

and parents. The isolated DNA and remaining blood samples will be stored under standard conditions in the Tissue Culture facility of the Human Genetics Department. A fraction of the isolated DNA of the trio (proband and parents) will be sent (using the Radboudumc Clinical Utility Studies pipeline) for WGS.

Sample size

The primary outcome of this study is the diagnostic yield of WGS in children with a severe DLD. This yield will be calculated as a proportion (with corresponding confidence interval). Because of this study design, a formal power calculation is not applicable. To determine the sample size needed for our study, in order to give an informative and reliable estimation of diagnostic yield, several different factors were taken into consideration: 1) the expected diagnostic yield (~30%) based on studies in developmental speech disorders and other neurodevelopmental disorder, 2) the selection of a subgroup of children of TOS for our study that is likely enriched for monogenic causes and 3) the expected genetic heterogeneity of DLD.

All in all, we determined that the study sample size should be 50 trios. This sample size is most likely large enough to answer our primary research question: "What is the diagnostic yield of WGS in children with severe DLD and a negative family history?", with a proportion that is representative and reliable. In addition to determining the diagnostic yield, we will describe the type of variants and genes found to be (possibly) causative for DLD. Comparable study designs, sequencing strategies and sample sizes have already proven successful in multiple other cohort studies performed in the Radboudumc Human Genetics Department (see references[21] and[22] as an example), and in sequencing studies performed worldwide in paediatric patients (reviewed in reference[23]).

Primary outcomes

The primary outcome of this study is the diagnostic yield of WGS in a cohort of children with a severe DLD and a negative family history. This diagnostic yield will be determined by categorizing individual sequencing results into three categories according to ACMG guidelines[20], as further discussed in the 'Data analysis' paragraph.

Secondary outcomes

This study has three secondary parameters/endpoints: A) the types of variants found to (possibly) cause DLD (CNV versus SNV, inherited versus *de novo*, etc.); B) the types of genes in which variants are found to (possibly) cause DLD (genes already known to be involved in neurodevelopmental disorders versus genes not yet associated with neurodevelopmental disorders); and C) possible subgroups in the cohort that might have a significantly different (higher or lower diagnostic yield), as defined in Table S1.

Data analysis

The interpretation of genetic data will be carried out based on the clinical information in the individual data summary from the Castor EDC database. To aid WGS data interpretation, a predefined gene-list (referred to as gene panel), will be used consisting of genes known to play a role in the etiology of neurodevelopmental disorders. This gene list is available on the website of the Genome Diagnostics department of the Radboudumc: "Intellectual disability gene panel DG 2.14" (1158 genes). If no possible pathogenic variants are found in genes listed in this gene panel, the data (including genes not associated with disease so far) is further analyzed ('open genome analysis') using pre-defined filters used in the standard clinical exome analysis pipeline as well, including but not limited to: inheritance of variant (maternal, paternal, *de novo*), allele frequency, presence in control databases (e.g. gnomAD[24]), presence in disease databases (e.g. HGMD[25], ClinVar[26]), predicted effect on RNA/protein level (using various *in silico* prediction programs). Using this standardized data analysis pipeline, each patient will be classified into one of the following diagnostic classes: 1) No obvious pathogenic variant(s) in a known disease-causing gene are observed in the proband; no definitive genetic diagnosis is obtained 2) A possible pathogenic disease-gene variant is observed (or multiple possible pathogenic variants are observed); a possible genetic diagnosis is obtained and 3) a pathogenic disease-gene variant is observed and validated, which explains the DLD in the child; a definitive genetic diagnosis is obtained.

The diagnostic yield will be defined and described on a group level, using the data from individual WGS results. For the total group, a descriptive report will be made on the overall outcomes of the WGS analysis in all subjects, in which the variant types will be described, the inheritance patterns, characteristics of the genes involved and possible subgroups with a different (higher or lower) diagnostic yield. Figure S1 shows a simplified visual overview of data analysis procedures.

Clinical data from speech and language tests will be independently assessed by two speech therapists separately before being added to the database. Clinical data analysis will be descriptive. High quality speech and language test data from Audiology Centers, of tests that have been performed less than five years ago at the time of inclusion, will be used as reference for the clinical speech and language characterization.

Incidental findings

A concern for exome or genome wide screening technologies is the small, but non-trivial, potential for identification of incidental findings. In this light, it is noteworthy that WES has been implemented clinically in the Netherlands as a first-tier genetic test for a wide range of neurodevelopemnatl disorders since September 2012. Most, if not all, ethical, legal and societal aspects associated with the diagnostic test in this study are thus identical to those faced when WES is used in clinical practice. The risk for incidental findings with our current WGS set-up would normally be comparable with the risk in diagnostic WES studies. However, before analyzing the WGS data we will remove all possible variants found in 59 medically actionable genes, as defined by the American College of Medical Genetics and Genomics (ACMG)[27], to minimize the identification of incidental findings in this study.

The final odds of an incidental finding with the current analysis and filtering set-up is low. A study on clinical findings using WES from the Radboudumc reported a chance of 0.9% for a set of 56 'medically actionable genes'[28] that was screened[29]. In our in-house experience of approximately 1,500 diagnostic exomes using trio-based WES an incidental finding was identified in 1.4% of individuals[30]. In both studies, incidental findings included variants found in medically actionable genes. By removing all variants in these medical actionable genes from the filtering and analysis process, the actual risk for incidental findings in our study will be very low, much less than 1%. If despite all measures to reduce the chance, an incidental finding is still discovered in our study, the standard clinical procedure for incidental findings of the Radboudumc will be used. This means that an external committee, consisting of at least an ethicist, lawyer, clinical geneticist, a clinical molecular laboratory specialist and a medical expert in the disease for which the variant was identified, will review the finding, and will make a decision by balancing the possible medical benefits with the ethical principles of doing no harm and the right to (not) be informed.

Data management

Each participant receives a unique study specific identification number, which can only be traced back to personal data via an identification code list that will be kept separately from the study data. This code list is accessible only to assigned members of the study team. De-identified clinical data is stored in a electronic case report files (eCRF) in a study-specific Castor EDC database. The specific data added to the eCRF per proband is specified in table S2. A de-identified proband-specific summary will be exported from Castor and provided to the researchers that perform the WGS data analysis, and is used for the individual interpretation of WGS variants. Results of previously performed tests and reports on medical history of patients are received as electronic files or on paper. All files are digitized in password-protected digital folders. Informed consent forms are digitized but original consent forms are stored as well. All other paperwork is destroyed once the files are safely digitally stored. WGS data will be stored in FASTQ format, labelled with study-specific identification numbers. BAM files (alignment data), SNV/CNV variant annotation files (VCF-files, hcdiff-files) and variant analysis files in (Microsoft Excel) will be stored in password-protected folders.

A schematic overview of clinical data analysis and data management procedures is provided in Figure S1. All data and material will be stored for a period of 15 years. The handling of personal data complies with the Dutch Personal Data Protection Act (de Wet Bescherming Persoonsgegevens; Wbp). A qualified monitor is assigned to monitor the study, including the data management procedures.

Discussion

With this study we aim to better understand the contribution of Mendelian genetic causes of DLD. We designed this prospective study of a selected cohort of individuals diagnosed with a severe DLD in the Netherlands to systematically assess the role of rare pathogenic variants in the development of this disorder. With this, we aim to define the total diagnostic yield and the contribution of different types of variants and inheritance patterns to this yield, to find out which genes are involved and whether there are any differences observed in diagnostic yield between different subgroups.

For this study we chose to include 50 children that meet all inclusion criteria and their parents. Probands are drawn from the population of children that are living in the Netherlands, and are diagnosed with DLD. The exact prevalence of this disorder in the Netherlands is not known[4], but is likely comparable with the prevalence of 7% that is reported in a USA-based study among five-year-olds[3]. Not all children with DLD will meet the criteria for inclusion in our study. In particular, the need for didactic 'Cluster 2 support' on the one hand, and requirement of a negative family history on the other, will limit the number of children that are eligible to participate in our study. There are no recent reports available on how many Dutch children with DLD need Cluster 2 support. A report from the Dutch Inspectorate for Education from 2009 states that 9,088 children attended a Cluster 2 school in 2008 (this includes children with speech and language disorders, but also those with hearing impairments) and that 3,807 children with speech and language difficulties received ambulatory Cluster 2 support[31], which means these children go to regular schools but receive extra support from 'Cluster 2 trained' teachers and therapists. Based on the total number of children that receive Cluster 2 support, our own experiences with the target group, and close consultation with several leading experts working in the DLD field, we expect it will be feasible to recruit the planned number of 50 patients for our study.

For this study, a sample size of 50 trios (affected individuals and their unaffected parents) is defined. Formal sample size calculations are not possible for this type of study design, and no reliable data are available on the expected diagnostic yield in DLD. More data are available for other neurodevelopmental disorders. In a recent study in which trio-based WGS was performed in 309 trios with intellectual disability and/or developmental delays, a diagnostic yield of 29% was reported[32]. When WGS is used in patients with severe forms of intellectual disability, a yield as high as 62% is obtained[33]. Although diagnostic yield is largely dependent on the coverage of the sequencing platform, and on the specific inclusion criteria of cohort studies, the diagnostic yield for epilepsy is reported to be 24-36% [34], and 20-25.8% for autism spectrum disorders [35,36]. If DLD has a similar genetic architecture to other neurodevelopmental disorders such as intellectual disability, autism and epilepsy, we anticipate a diagnostic yield in the same range.

We chose to include a specifically selected group of children with DLD, which may not be representative for the population of children with a DLD as a whole. Because the size of our study is restricted by the costs of WGS, we have chosen to enrich our cohort for children that most likely have rare monogenic causes, by including those with a severe form of DLD and a negative family history. If this strategy leads to a substantial diagnostic yield, screening can be extended to cohorts with milder forms of DLD in future studies. Alternatively, if the diagnostic yield for rare monogenic causes turns out to be low, this kind of diagnostic testing might not be useful for broader populations of children with DLD.

In conclusion, our study can serve as a first study to determine the diagnostic yield of trio-based WGS in a selected cohort of children with a severe DLD in the Netherlands. Knowledge of frequencies and types of underlying monogenic causes will lead to a better understanding of the genetic architecture and pathogenic mechanisms. Moreover, such knowledge is crucial to guide referral strategies and diagnostic decisions, and to work towards better and more personalized healthcare for individuals with DLD, in the Netherlands and abroad.

Study status

Inclusion for this study started in March 2020 and is still ongoing at the time of writing.

Declaration of conflicting interests

The authors declared no potential conflicts of interest.

Acknowledgements

Funding for this study was provided by the Netherlands Organization for Scientific Research (NWO) Gravitation Grant 24.001.006 to the Language in Interaction Consortium (L.S.B., S.E.F., and H.G.B.), the Netherlands Organization for Health Research and Development (ZonMw grant 91718310 to T.K.) and the Max Planck Society (S.E.F.).

References

1. Stothard SE, Snowling MJ, Bishop DV, Chipchase BB, Kaplan CA. Language-impaired preschoolers: a follow-up into adolescence. *J Speech Lang Hear Res.* 1998;41(2):407-418.
2. Feeney R, Desha L, Ziviani J, Nicholson JM. Health-related quality-of-life of children with speech and language difficulties: a review of the literature. *Int J Speech Lang Pathol.* 2012;14(1):59-72.
3. Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res.* 1997;40(6):1245-1260.
4. Gerrits E, Beers M, Bruinsma G, Singer I. *Handboek Taalontwikkelingsstoornissen.* Coutinho; 2017.
5. Johnson CJ, Beitchman JH, Brownlie EB. Twenty-year follow-up of children with and without speech-language impairments: family, educational, occupational, and quality of life outcomes. *Am J Speech Lang Pathol.* 2010;19(1):51-65.
6. Bishop DV, Snowling MJ, Thompson PA, Greenhalgh T, consortium C. CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. *PLoS One.* 2016;11(7):e0158753.
7. Gerrits e, Niel Ev. Taalachterstand of taalontwikkelingsstoornis? *Logopedie.* 2012;84(11):6-10.
8. Foniatrie NVvLe. *Richtlijn Logopedie bij taalontwikkelingsstoornissen.* 2017.
9. Deriziotis P, Fisher SE. Neurogenomics of speech and language disorders: the road ahead. *Genome Biol.* 2013;14(4):204.
10. Graham SA, Fisher SE. Understanding Language from a Genomic Perspective. *Annual review of genetics.* 2015;49:131-160.
11. Srivastava S, Love-Nichols JA, Dies KA, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med.* 2019;21(11):2413-2421.
12. Vissers LE, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet.* 2016;17(1):9-18.
13. Bishop DV. The role of genes in the etiology of specific language impairment. *J Commun Disord.* 2002;35(4):311-328.
14. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature.* 2001;413(6855):519-523.
15. Morgan A, Fisher SE, Scheffer I, Hildebrand M. FOXP2-Related Speech and Language Disorders. In: Pagon RA, Adam MP, Ardinger HH, et al., eds. *GeneReviews(R).* Seattle (WA)1993.
16. Chen XS, Reader RH, Hoischen A, et al. Next-generation DNA sequencing identifies novel gene variants and pathways involved in specific language impairment. *Sci Rep.* 2017;7:46105.
17. Eising E, Carrion-Castillo A, Vino A, et al. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol Psychiatry.* 2018.
18. Hildebrand MS, Jackson VE, Scerri TS, et al. Severe childhood speech disorder: Gene discovery highlights transcriptional dysregulation. *Neurology.* 2020.
19. Kalnak N, Stamouli S, Peyrard-Janvid M, et al. Enrichment of rare copy number variation in children with developmental language disorder. *Clin Genet.* 2018;94(3-4):313-320.
20. Richards S, Aziz N, Bale S, et al. 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.* 2015;17(5):405-424.
21. Diets IJ, Waanders E, Ligtenberg MJ, et al. High Yield of Pathogenic Germline Mutations Causative or Likely Causative of the Cancer Phenotype in Selected Children with Cancer. *Clin Cancer Res.* 2018;24(7):1594-1603.
22. Bosch DG, Boonstra FN, de Leeuw N, et al. Novel genetic causes for cerebral visual impairment. *Eur J Hum Genet.* 2016;24(5):660-665.
23. Smith HS, Swint JM, Lalani SR, et al. Clinical Application of Genome and Exome Sequencing as a Diagnostic Tool for Pediatric Patients: a Scoping Review of the Literature. *Genet Med.* 2018.
24. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443.
25. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat.* 2003;21(6):577-581.
26. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062-D1067.

27. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med.* 2017;19(2):249-255.
28. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med.* 2013;15(7):565-574.
29. Jurgens J, Ling H, Hetrick K, et al. Assessment of incidental findings in 232 whole-exome sequences from the Baylor-Hopkins Center for Mendelian Genomics. *Genet Med.* 2015;17(10):782-788.
30. Yntema H. Experiences with the dissemination of secondary findings by diagnostic exome sequencing. In. American Society of Human Genetics, Annual Meeting 2015, Baltimore, USA.
31. Onderwijs Ivh. *Cluster 2: De kwaliteit van het onderwijs in Cluster 2.* Inspectie van het Onderwijs;2009.
32. Bowling KM, Thompson ML, Amaral MD, et al. Genomic diagnosis for children with intellectual disability and/or developmental delay. *Genome Med.* 2017;9(1):43.
33. Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature.* 2014;511(7509):344-347.
34. Lee H, Deignan JL, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA.* 2014;312(18):1880-1887.
35. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. *Genet Med.* 2016;18(7):696-704.
36. Rossi M, El-Khechen D, Black MH, Farwell Hagman KD, Tang S, Powis Z. Outcomes of Diagnostic Exome Sequencing in Patients With Diagnosed or Suspected Autism Spectrum Disorders. *Pediatric neurology.* 2017;70:34-43 e32.

Supplementary Materials

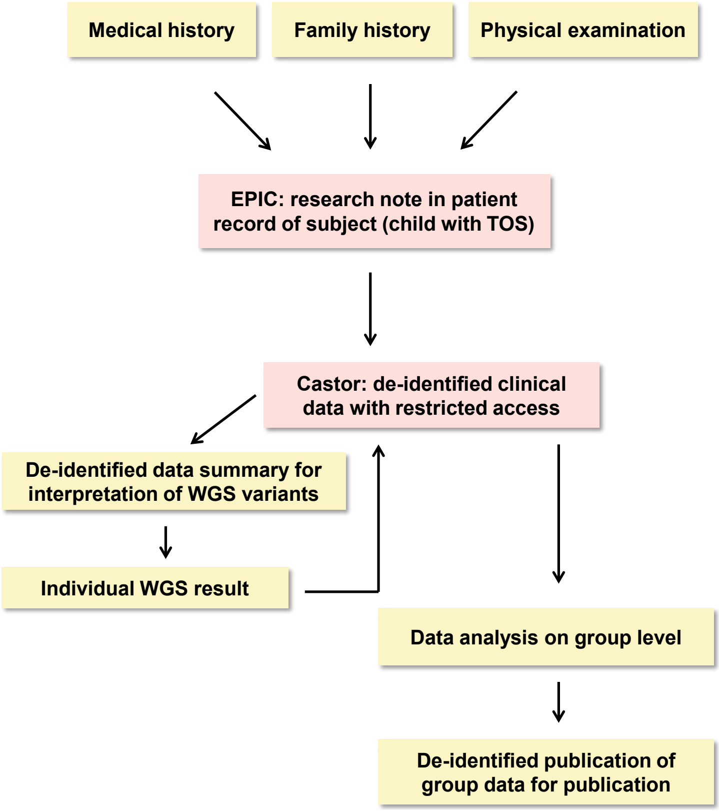

Medical history
Family history
Physical examination
EPIC: research note in patient record of subject (child with TOS)
Castor: de-identified clinical data with restricted access
De-identified data summary for interpretation of WGS variants
Individual WGS result
Data analysis on group level
De-identified publication of group data for publication

Figure S1: Schematic overview of clinical data analysis and data management procedures

Table S1: Subgroups with a possible different diagnostic yield

Category	Subgroup 1	Subgroup 2
Cluster 2 support	Attends cluster 2 school	Ambulatory support
Non-verbal IQ (last test)	< 90	≥90
Speech sound disorder reported	Yes	No
Receptive language problems	Yes	No
Receptive Language (RL) versus Expressive Language (EL)	RL>EL	RL = EL
Pragmatic language problems reported	Yes	No
Age of DLD diagnosis	< 4 years	≥ 4 years
Congenital abnormalities	Yes	No
Epilepsy/seizures in history	Yes	No
Gender	Male	Female

Table S2: Clinical data that will be stored in electronic case record files

Medical history	
Pregnancy/birth	Pregnancy abnormalities Birth abnormalities Birth weight / head circumference Congenital abnormalities Neonatal problems
Details of DLD	Age of diagnosis Details of speech problems Details of language problems Feeding history, saliva control, oral sensory issues History of therapeutic intervention Results of previously performed tests
(Neuro)development	Gross motor development/milestones Fine motor development General cognitive abilities (IQ test?) Behavioural issues Presence/absence of epilepsy Sleeping problems Presence/absence of hypotonia Other neurologic features
Other	Vision Hearing Medication Other co-morbidities
Family history (1st and 2nd degree family members)	
DLD	DLD or speech/language difficulties?
Development	Developmental delays? Cognitive impairments?
Other	Other important medical issues?
Physical examination	
Growth parameters	Height Weight Head circumference
Dysmorphic features	Description of features (in Human Phenotype Ontology (HPO) terms)

9

Chapter 9

General Discussion

Summary of results

The goal of this thesis was to study biological etiology and clinical aspects of disorders that disrupt speech and language developmental, taking advantage of the new opportunities of the next generation sequencing era. By using exome and genome sequencing data of children with major developmental speech and language disorders, we were able to identify five new genes in which pathogenic variants appeared causative: *MED13, POU3F3, CHD3, WDR5* and *FOXP4*. We characterized the full clinical spectrum of the disorders associated with disruptive mutations of these genes. Apart from speech and language deficits we found a wide range of phenotypic features. Therefore, we consider these to be Mendelian neurodevelopmental syndromes. Using both *in silico* and *in vitro* cell-based approaches, we were able to confirm pathogenicity for the identified variants. In addition to these genome-first Mendelian disorder discovery studies, we performed extensive speech/language phenotyping and neuropsychological profiling in 23 individuals with *SATB2*-associated syndrome, a known neurodevelopmental disorder involving severe impairments in communication. Lastly, we designed a novel study to systematically assess *de novo* variants in a selected cohort of children with a severe developmental language disorder.

This research demonstrated how trio-based next generation sequencing studies can be used to delineate genetic pathways underlying developmental speech and language problems, and how *de novo* variants in Mendelian disease genes can disrupt skills in these domains. We found that different underlying mechanisms and pathways can lead to similar disturbances in speech and language. Conversely, similar molecular disruptions can give rise to different phenotypes in different probands. A combination of complementary methods were used including 1) trio-based next generation sequencing in children with primary developmental speech and language disorders, 2) *in vitro* studies for assessing pathogenicity and 3) systematic phenotyping. Studying developmental speech and language disorders from this multidisciplinary perspective, including the integration of knowledge from clinical genetics, molecular biology, speech and language pathology and neuropsychology, contributes to a more complete understanding of this important class of disorders, and may be adapted in future studies.

Novel contributions to the field

We studied children with a developmental speech and/or language disorder as a starting point for characterizing new genes in which de novo mutations lead to developmental phenotypes. With this genotype-driven approach, we were able to associate five genes to a novel Mendelian developmental disorder. The disorders caused by heterozygous variants in three of these five genes are now listed in the OMIM database as Mendelian neurodevelopmental disorders: *MED13* variants cause 'intellectual developmental disorder 61' (MIM #618009), *CHD3* variants cause 'Snijders Blok-Campeau syndrome' (MIM #618205) and the disorder associated with variants in *POU3F3* has been listed as 'Snijders Blok-Fisher syndrome' (MIM #618604). For the disorders associated with heterozygous variants in the genes *FOXP4* and *WDR5*, an OMIM listing is not yet available at the time of writing.

For the genes *CHD3* and *WDR5*, our starting point was a small published cohort study (n=19) by Eising et al.[1] on children with childhood apraxia of speech, in which a *de novo* variant of unknown significance was reported in *CHD3* in one affected individual, and in *WDR5* in another individual[1]. Clinical and molecular studies on *CHD3* and *WDR5* confirmed the pathogenicity of these variants, and established new Mendelian disorders associated with severe speech and language deficits and various additional features (Chapter 4 and 5).

The study of Eising et al.[1], as well as a subsequent cohort study (n=34) by Hildebrand et al.[2], provided the first evidence to suggest a high diagnostic yield of rare *de novo* variants in children with a developmental speech disorder. To more systematically assess the contributions of pathogenic *de novo* variants to developmental language disorders, we designed a new prospective cohort study: the GENTOS study (Chapter 8).

In addition to gene discovery, we performed in-depth phenotypic characterization of individuals with heterozygous pathogenic *SATB2* variants from a speech, language, oral motor and neuropsychological perspective. Studies that combine phenotyping on a speech/language level and on a neuropsychological level are scarce. While a previous study reported childhood apraxia of speech to be present in all studied individuals with *SATB2*-associated syndrome[3] (n=61), we showed that in addition, several other speech diagnoses were found to contribute to the severe problems observed. We also described a characteristic profile in adaptive functioning assessed by Vineland screener tests for individuals with pathogenic *SATB2* variants, with relatively high scores on the ‘Daily Functioning’ domain and relatively low scores on the ‘Communication’ domain. The results can guide a personalized approach for parents and therapists.

Clinical relevance and implications of genetic studies in developmental speech and language disorders

Using trio-based next generation sequencing data to identify de novo variants

Research on the genetics of developmental speech and language disorders has often tended to focus on multifactorial inheritance[4]. Our research is distinctive in focusing largely on Mendelian causes of speech and language problems. In Chapter 2-6, we used next generation sequencing data to specifically search for Mendelian disorders involving prominent speech and language impairments. Over the last decade, many studies have already shown that next generation sequencing in parent-child trios provides a powerful tool to identify underlying molecular defects for a plethora of different types of disorders[5-7]. We applied such an approach here, with family-based sequencing data and follow-up of interesting *de novo* variants. This strategy is effective, because it does not depend on the availability of large systematically phenotyped speech and language disorder cohorts.

While our studies illustrate the possibility and relevance of finding a clinically relevant *de novo* pathogenic variant, they do not tell us which fraction of individuals with speech and language disorders carries high-penetrance variants of this nature. For developmental disorders in general, it is estimated that around 40% of individuals have a pathogenic *de novo* variant in the coding sequence[8]. We designed the GENTOS study (Chapter 8) to give an insight into the diagnostic yield for rare pathogenic variants in a selected group of children with a developmental language disorder. Preliminary literature on this topic suggests a diagnostic yield of around 30% for individuals with a developmental speech disorder such as childhood apraxia of speech[1,2]. This estimate is expected to increase, as sequencing conditions (e.g. sequencing coverage and quality of sequencing data) and knowledge on pathogenic variants improve.

Confirming pathogenicity of variants

In Chapters 2-6, we used a combination of publicly available data, *in silico* tools and *in vitro* functional assays in cell systems, in order to reliably interpret the likely effects of variants of interest. In Chapter 2, 3 and 5 we used three-dimensional modeling to derive hypotheses on pathogenic mechanisms, taking advantage of prior knowledge on protein structures. Modeling has been shown to be effective in differentiating neutral versus pathogenic missense variants in different types of genetic disorders, and is becoming increasingly available for large-scale analyses of missense variants[9,10].

In addition to *in silico* tools, functional studies can provide empirical support for etiological relevance of identified variants. In Chapter 3, 4 and 6 different *in vitro* assays were used to demonstrate pathogenicity. We showed that most of the *de novo CHD3* gene variants that we studied disrupted chromatin remodeling capacities of the CHD3 protein, and that variants in *POU3F3* and *FOXP4* interfered with transcriptional activities of the encoded transcription factors. Although proving pathogenicity can be a challenge and is often time-consuming, assigning functional evidence to variants of unknown significance is of crucial importance to their correct classification[11,12]. Functional evidence, from *in vitro* assays as well as *in vivo* studies, can elucidate molecular mechanisms and pathways involved in neurodevelopmental (disease) processes. Saturation mutagenesis approaches in combination with protein-dependent functional read-outs may enable more high-throughput tests of variants[13]. Also, models more relevant for three-dimensional organization in early brain development, such as brain organoids[14], are increasingly used for functional validation and further follow-up of variants of interest.

Follow-up of candidate genes: identification of variable phenotype expressivity

In Chapter 2-6, after the identification of a possible pathogenic candidate gene, we collected additional individuals with potential pathogenic variants in the same gene. Strategies included searches through public databases such as DECIPHER[15], ClinVar[16], HGMD[17] and denovo-db[18], using 'gene matchmaking' platforms such as GeneMatcher[19], querying large clinical and research sequencing datasets (from academic and commercial sequencing centres) and searching the medical literature for reported variants in the genes of interest.

Using this approach, we were able to capture a phenotypic spectrum associated with these rare variants, that typically appeared broader than the initial disorder of interest; both in terms of neurodevelopmental phenotypes and in terms of additional clinical features such as dysmorphisms or congenital abnormalities.

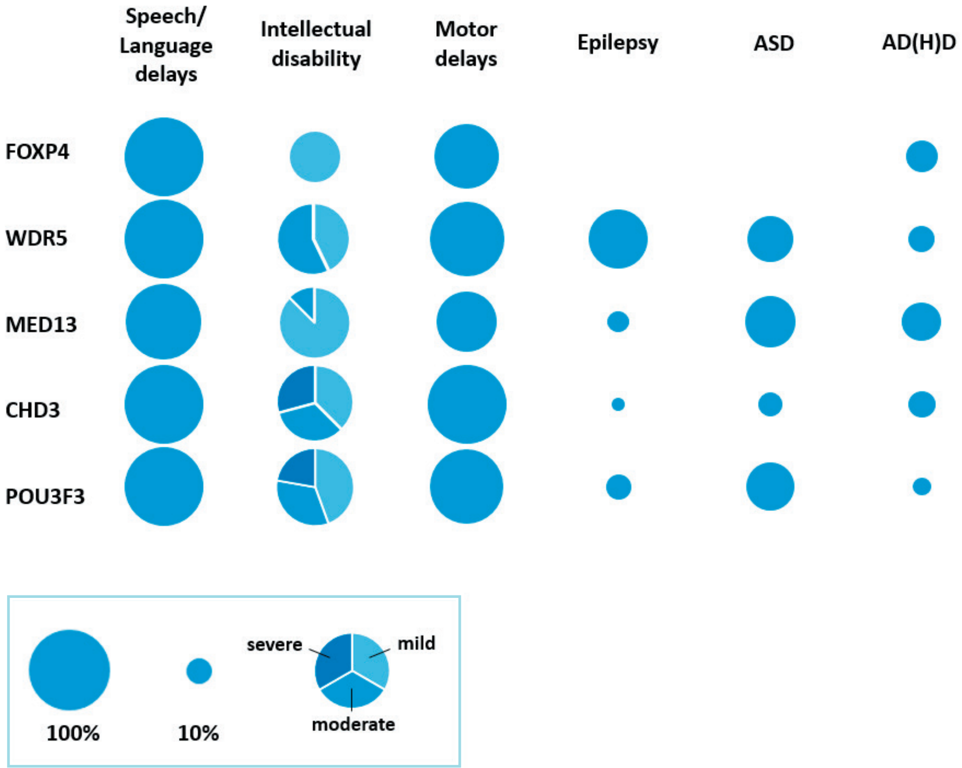


Figure 1: Frequency of neurodevelopmental phenotypes observed in five different Mendelian disorders

Frequencies of different neurodevelopmental disorder phenotypes are shown for the Mendelian disorders associated with FOXP4, WDR5, MED13, CHD3 and POU3F3, as clinically characterized in Chapter 2-6 of this thesis. The size of each circle represents the associated frequency per phenotype. For the intellectual disability phenotype, the frequency of different degrees of severity is visualized using pie charts. ASD = Autism Spectrum Disorder, AD(H)D = Attention Deficit (Hyperactivity) Disorder.

Mendelian disorders often present with more than just the neurodevelopmental phenotype. In our studies this is illustrated by the presence of gene-specific associated features in subsets of individuals with the disorders described in Chapter 2-6: e.g. optic nerve abnormalities and Duane anomaly in some individuals with a *MED13* variant (Chapter 2), characteristic prominent and cupped ears in all individuals with a loss-of-function variant in POU3F3 (Chapter 3), macrocephaly and a distinctive facial gestalt in individuals with a pathogenic variant in *CHD3* (Chapter 4), and a congenital diaphragmatic hernia and/or vertebral abnormalities in a subset of individuals with *FOXP4* variants (Chapter 6). While in clinical genetics the terms 'syndromic' and 'nonsyndromic' are often used to classify individuals with and without co-morbidities in addition to their primary phenotype (e.g. syndromic versus nonsyndromic intellectual disability)[20], our results once more highlight the unclear boundaries between these two categories, because of clinical variability. Also, some of these additional features are subtle and may initially escape recognition. Objectively assessing dysmorphisms is challenging and not always within the competencies of the attending physician.

For the results of the clinical characterization in Chapter 2-6, some skew is likely towards the more severe end of the phenotypic spectrum. NGS studies with a trio approach are known to be biased towards more severe phenotypes[21], and diagnostic testing using NGS might also be performed less frequently in individuals with phenotypes without developmental delays, compared to those with neurodevelopmental disorders. The rarity of the causative gene variants often precludes statistically significant conclusions on presence and prevalence of phenotypic features. Chapter 6 represents an example of a small discovery cohort, as *de novo* variants in *FOXP4* seem to be very rare. Although no doubt exists about the pathogenicity of the variants affecting the transcriptional repressor activity of FOXP4 as presented in this study, novel associated features remain to be identified, once more individuals with the associated disorder are studied. On the other hand, our study on *CHD3* variants in Chapter 4 involved 35 individuals with a *de novo* variant in this gene. This is a relatively large group for such a gene discovery cohort, and allowed a robust first characterization of the *CHD3*-associated phenotype. Indeed, our data on phenotypic features in this disorder have recently been confirmed in a clinical follow-up study using an independent cohort of individuals with pathogenic *CHD3* variants[22].

"Speech/language disorder-specific" genes?

The search for "speech/language disorder-specific" genes

We studied *de novo* pathogenic variants in children with developmental speech and language disorders to identify five new neurodevelopmental disorders, caused by disruption of genes that had not previously been linked to any Mendelian disorder (Chapter 2-6). As discussed in the previous paragraphs, as cohorts were investigated beyond the initial probands, the associated phenotypes were not restricted to speech and language. Although such problems turned out to be a prominent feature in all five newly characterized disorders, other neurodevelopmental issues such as intellectual disability, autism spectrum disorders, ADHD and epilepsy phenotypes were commonly found, and to variable degrees (Figure 1). As an example, in Chapter 5 we show how a single p.(Thr208Met) missense variant in *WDR5* gives rise to a severe mixed language disorder and borderline intellectual disability in one individual, and to a phenotype with moderate ID, autism spectrum disorder and absent speech in another individual.

Based on our results and those of previous cohort studies[1,2], the genetic causes of speech and language disorders clearly overlap with those of intellectual disability and autism spectrum disorders. In addition, speech/language delays diagnosed at a young age can be the first presentation of broader neurodevelopmental impairments at a later age[23]. We cannot currently explain differences in neurodevelopmental phenotypes among individuals with similar molecular defects. Such clinical variability may reflect local regulatory genetic variants[24], other effects of genetic background[25], and modulation by environmental and/or social factors. A recent publication on 22q11.2 deletion syndrome shows how polygenic scores based on common genetic variation might be used in future for better prediction of expected phenotypes in individuals with rare genetic variants[26]. It has further been posited that stochastic events or small inter-individual differences in important pathways during development might lead to large downstream differences with a major impact on the phenotype[27].

Genes associated with significant speech/language disorder phenotypes

For the newly characterized Mendelian disorders in Chapter 2-6, the speech and language problems form a prominent and important part of a broader clinical spectrum. For instance, in our *CHD3*-associated disorder cohort the overwhelming majority of individuals (25 of 27 for whom information was available) with a pathogenic variant in this gene was receiving (or had received) speech therapy (Chapter 4). The most frequent phenotypes that we observed in individuals with pathogenic variants in *CHD3* included childhood apraxia of speech, impaired intelligibility, fluency problems, and mixed or expressive language disorders. Speech and language delays or disorders were reported in all individuals with pathogenic *MED13* variants (Chapter 2), with an emphasis on expressive speech problems rather than receptive language problems. In *FOXP4*-associated disorder (Chapter 6), all six individuals with confirmed pathogenic variants in the study had speech and language delays, with a formal expressive language disorder diagnosis in two individuals. In individuals with pathogenic variants in *POU3F3* (Chapter 3) or *WDR5* (Chapter 5) severe speech and language problems are a commonly reported phenotypic feature too. And as is shown in Chapter 7, even when observed in the context of more substantial intellectual disability, as observed in *SATB2*-associated syndrome, impaired speech and language can still have a large disruptive effect on daily functioning.

All in all, pathogenic variants in a subset of neurodevelopmental disorder genes, including the genes characterized in this thesis, give rise to neurodevelopmental disorders with prominent speech and language dysfunction (Figure 2A). This means that speech and language problems and disorders are more frequently and more severely present in individuals with variants in these genes, compared to individuals with variants in other genes associated with neurodevelopmental disorders. Thus, clinical practice is better suited to a model which acknowledges the diversity of features associated with variants in a given gene, rather than a model in which variants in any single gene lead to a specific neurodevelopmental phenotype.

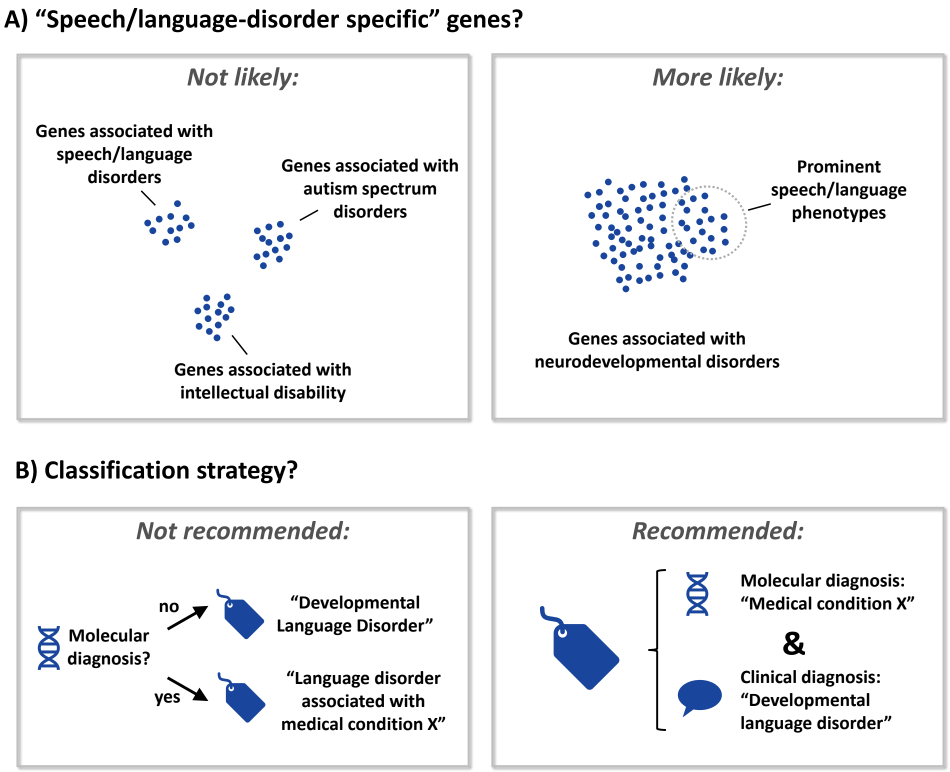

A) "Speech/language-disorder specific" genes?
Not likely:
Genes associated with speech/language disorders
Genes associated with autism spectrum disorders
Genes associated with intellectual disability
More likely:
Prominent speech/language phenotypes
Genes associated with neurodevelopmental disorders
B) Classification strategy?
Not recommended:
Molecular diagnosis?
no
"Developmental Language Disorder"
yes
"Language disorder associated with medical condition X"
Recommended:
Molecular diagnosis: "Medical condition X"
&
Clinical diagnosis: "Developmental language disorder"

Figure 2: Recommendations for classification of genes and disorders

a) The question is whether "speech/langue-disorder specific" genes exist after all. Based on the current literature and the results of the research in this thesis, a model in which separate sets of genes are associated with different phenotypes (e.g. intellectual disability, speech/language disorders or ASD) is not likely. We propose a model in which a subset of the genes in which pathogenic variants cause neurodevelopmental disorders is associated with more prominent and/or more frequent speech/ language disorder phenotypes.

b) A clinical classification strategy for individuals with speech and language disorders in which the clinical diagnosis depends on whether a causative molecular defect has been identified is not recommended. We recommend to use clinical and molecular labels or diagnoses separately and independently.

Do "speech and language-disorder specific" genes exist?

Current data do not exclude that "speech/language disorder specific" genes may exist after all, although considerations of the complexity of biological mappings between DNA and behaviour/cognition suggest that this degree of specificity is unlikely[28,29]. We and others have been unable to find genes for which Mendelian pathogenic variants specifically cause speech and language disorders, without any other associated neurodevelopmental phenotypes. In this regard, *FOXP2* remains the most striking example of genotype-phenotype associations that clearly involve a developmental speech disorder. The majority of individuals with a pathogenic *FOXP2* variant have severe speech problems (childhood apraxia of speech), but even here, the neurodevelopmental phenotypes observed in these individuals are often not limited to speech problems[30]. Our findings on speech and language are in line with recent work on monogenic causes of autism spectrum disorders, which concluded that there was presently no evidence for "autism-specific" genes [31].

To more robustly rule out the existence of "speech/language disorder-specific" genes, prospective highly selected cohort studies are needed. The GENTOS study (Chapter 8) is designed to potentially identify these genes, especially if *de novo* variants are involved. Definitive conclusions on the (non)existence of such genes, taking into account the possibility of inherited variants in combination with reduced penetrance and variable expressivity, will require much larger studies based on comprehensive data sets. An example of such a large dataset could be the UK Biobank, a long-term study in the United Kingdom which investigates the contribution of genetic and environmental factors to the development of disease [32]. This database contains clinical data and genetic data on 500,000 participants and recently released exome sequencing data on 200,000 individuals. However, this resource is mainly aimed at complex disease of middle and older ages. Currently, no suitable parameters are available in UK Biobank with respect to developmental speech and language deficits of the participants, limiting its usage for shedding light on etiology of this major class of disorders.

Consequences for classification: should speech and language disorders be considered a separate entity?

Given the overlapping genetics demonstrated here and elsewhere, one might wonder whether developmental speech and language disorders should be considered as an etiologically distinct entity, separate from other neurodevelopmental disorders. We think it is important to keep a clear distinction between a clinical and a molecular diagnosis. While different children might have an identical genetic diagnosis (even carrying the same pathogenic variant), the phenotype might still differ from one case to the next. If a child has a clinically defined developmental language disorder, and then a molecular cause is found that is predominantly associated with autism spectrum disorder or intellectual disability, the child should still receive treatment that is matched to the clinical needs (i.e. language disorder). A molecular diagnosis is important as it accounts for the developmental problems in the child, and establishes the origin of the disorder and the inheritance pattern. Sometimes, a molecular diagnosis can also be helpful for informing parents about a likely prognosis. Yet, molecular classification should not subsume clinical classification. It is not possible to classify children with clinical diagnoses based on molecular data alone. Although progress has been made in better understanding and predicting clinical outcomes in Mendelian disorders (e.g. as has recently been shown in 22q11 deletion syndrome[26]), for the foreseeable future, a pathogenic variant will not predict the exact clinical outcome for these types of disorders.

In clinical linguistics, a molecular diagnosis is sometimes used as an exclusion criterion for the label ‘developmental language disorder’ (in abbreviated form, DLD), for instance in the recommendations of the CATALISE consortium[33]. The authors argue that the term DLD should be used exclusively when there is no underlying biomedical condition present, such as e.g. Down syndrome[34]. Based on the findings of this thesis, with the current knowledge of the likely genetic underpinnings of developmental speech and language disorders, as well as the impossibility to clinically separate Mendelian from non-Mendelian disorders, we would argue strongly against making this distinction based on the presence/absence of an underlying molecular diagnosis (Figure 2B). As also illustrated by chapter 7, a genetic diagnosis can give guidance and clarity for all families involved, but the clinical diagnosis remains the cornerstone of care.

Results in the context of the neurobiology of speech and language

Networks of genes and proteins involved in shaping a “language-ready” brain

It was once believed, that the brain had highly specific regions important for language functioning, e.g. the Broca and Wernicke regions[35,36]. The current picture though is more complex than this classical neurobiological view[36,37]. While many parts of the multifaceted human-specific capacity to acquire and use complex language remain to be discovered, it is known that multiple different cortical and subcortical regions are involved, as well as the basal ganglia and cerebellum, together forming distributed circuits. Moreover,both frontal and temporal regions are involved in expressive and receptive language processes, and the initial simplistic views regarding lateralization of the crucial circuits are being replaced by a more nuanced perspective[36,38,39].

This neurobiological complexity fits with the genetic heterogeneity of speech and language deficits demonstrated here and elsewhere in the literature. A highly intricate interplay of different genes and their encoded proteins, differently expressed in time during development and in diverse brain regions, is needed to shape the developing brain in such a way that it is able process and produce speech and language - the basis of the so-called “language-ready” brain[38,40]. As a result, disruptions of an array of different types of genes and encoded proteins could potentially impact on speech and language capacities. Studying the genes and pathways involved in abnormal speech and language development might shed light on the crucial neurobiological mechanisms.

Characteristics of genes involved in Mendelian causes of speech/language disorders

Although the molecular underpinnings of developmental speech and language disorders reflect extreme genetic heterogeneity, as discussed previously, a subset of genes linked to Mendelian neurodevelopmental disorders seems to be associated with a more common and/or more prominent speech and language phenotype. Knowledge about the characteristics of this particular set of genes, and the biological processes in which the encoded proteins play a role, might lead us to molecular pathways involved in normal and abnormal speech and language development.

Nonetheless, considerable uncertainty remains with regard to the nature of these molecular pathways. Two published studies provide evidence that a set of genes co-expressed in early human brain development and involved in regulatory pathways shows enrichment in NGS screening of childhood apraxia of speech[1,2], but further studies are needed to replicate and confirm these findings.

If we look at the five genes characterized in Chapter 2-6 of this thesis, we can note that all are highly expressed during early brain development, and that the encoded proteins are known to have important functions in gene regulation processes. FOXP4 and POU3F3 are transcription factors that regulate the expression of a large variety of target genes[41,42], MED13 is involved in transcriptional regulation and transcription initiation as part of the Mediator complex[43], CHD3 is involved in chromatin remodeling[44], and WDR5 is a crucial part of different well-known protein complexes involved in chromatin-related regulatory functions[45].

So far, a large part of the literature on the molecular biology of speech and language development is focused on functions of FOXP2 and its orthologs in brain development, animal vocalizations and human speech[46]. Several of the genes and proteins that we identified have direct links to FOXP2. Strikingly, FOXP4, is not only a paralog of FOXP2 with partially overlapping expression patterns in several brain regions, but is also known to dimerize with itself or with FOXP2 in order to exert its transcriptional regulation functions[47]. As mentioned in Chapter 4, the protein CHD3 has been found to be an interactor of FOXP2 in a yeast two-hybrid assay[48]. In addition to these protein-protein interactions, we have demonstrated in Chapter 3 that an intronic element in the *FOXP2* gene can serve as a functional binding site for POU3F3. Previous work has shown that phenotypic networks often reflect underlying biological networks defined by regulation or protein-protein interactions[49,50].

While it is tempting to speculate that dysfunctions in developmental speech and language disorders might be the result of converging molecular pathways and networks, further research is needed to study this and further define the essential processes and relevant neural circuits. Although the advances and developments in genetics and molecular biology are promising, there is still a significant gap between gene/protein function and cellular networks on one side, and human brain functioning and phenotypes on the other side.

Proposal for diagnostic genetic testing using NGS in developmental speech and language disorders

Based on the results of this thesis and published literature we propose genetic testing using NGS for the following individuals:

- Individuals with a severe developmental speech disorder *(with exception of individuals with developmental speech disorders only affecting fluency of speech, e.g. stuttering)*
- Individuals with a severe developmental language disorder and additional abnormalities (e.g. dysmorphisms, growth parameters, congenital abnormalities, etc.)
- Severe developmental language disorder with a negative first-degree family history

It does not seem helpful to only focus on genes that have been associated with speech and language disorders in prior literature, due to the variable expressivity of many Mendelian neurodevelopmental disorders in combination with possibly incomplete phenotypic reports. We would therefore advocate a targeted NGS analysis of all developmental disorder genes in individuals with developmental speech and language disorders.

In view of the relevance of finding *de novo* variants, a negative family history should not be used as an exclusion criterion for genetic testing. In fact, the likelihood of finding a Mendelian cause in a child with a severe speech and/or language disorder and a completely negative family history may well exceed that of a child with a strongly positive family history. In individuals with familial aggregation of speech and language disorders, multifactorial inheritance is probably the most likely explanation, although some families with rare high-penetrance variants have also been reported (e.g. the KE family with a pathogenic variant in *FOXP2*[51]). There is currently no molecular test that can demonstrate multifactorial genetic predisposition.

Lastly, the presence of co-morbidities might point more strongly to an underlying Mendelian cause, but the absence of additional features should not be used to exclude this possibility.

Future perspectives

Clinical value of NGS in developmental speech and language disorders

We suggest not to be reluctant to use NGS in individuals who present mainly or solely with developmental speech and language disorders. This will benefit children/probands and their families, as a molecular diagnosis can provide clarity about the cause of the disorder, about the recurrence risk for family members, and can also be informative regarding presence of co-morbidities and relevant recommendations. As shown in Chapter 7, a molecular diagnosis can also guide more personalized diagnostic and treatment options for speech and language therapists. Although a small number of studies suggest otherwise[52], so far, there is no clear evidence for molecularly guided drug treatment to ameliorate language dysfunction in developmental speech and language disorders. As these disorders are the result of disrupted brain development, it is not expected that genetic therapy will become available in the future to solve all problems. However, one might speculate that for specific molecular defects, certain drugs will ultimately be found to lead to improvement on specific cognitive or speech/language related functions.

Systematic and consistent phenotyping in developmental speech and language disorders

One of the main challenges in genetics research in the speech and language field is the lack of consistency in phenotypic classification. On a clinical level, different types of classifications and labeling are used, which hamper proper comparisons. Clinical geneticists often rely on data from diverse international sources, so that when an individual from one country is reported with a developmental speech disorder, another individual from a different country with verbal dyspraxia, and a third individual has reported speech delays, these divergent terms might still indicate the same phenotype. The field of genetics of developmental speech and language disorders would benefit from comparable and systematic classification and labeling. For other Mendelian disorders, standardized phenotyping is increasingly performed using Human Phenotype Ontology (HPO) terms[53]. For developmental speech and language disorders, only a few standard terms are available in HPO, e.g. ‘delayed speech and language development’, ‘absent speech’, ‘expressive language delay’, ‘receptive speech delay’ and ‘childhood apraxia of speech’. Though these can be used in a rough stratification, they do not completely match with the proposed and commonly used diagnostic classifications, and also are insufficiently detailed to allow good distinctions between different types of speech and language disorders.

In NGS research studies in neurodevelopmental disorders, speech/language-oriented phenotype data are often incomplete or absent. While presence or absence of intellectual disability and specific somatic features are commonly reported, reliable phenotype data on behaviour as well as speech and language development and/or speech- and language-related diagnoses are often lacking. This might lead to a bias in published literature on such phenotypes of Mendelian neurodevelopmental disorders. Research studies with systematically collected speech and language data would advance the field. A number of initiatives are currently aimed at overcoming these difficulties. As an example, the GenLang consortium (http://genlang.org) brings together genetic data and researchers on normal and abnormal speech and language function from multiple different laboratories around the world, in order to set up large-scale cohort meta-analyses with enough power to detect important genetic variants. Another example of an initiative that can be used for improving systematic phenotyping in Mendelian disorders is the Dutch Computer Articulation Instrument, which is designed to systematically and automatically assess speech capacities, and has already been shown to be useful in diagnosing specific speech problems in individuals with Koolen-de Vries syndrome[54].

Consortia, data sharing and patient empowerment

There is increasing interaction of scientists studying basic biological or genetic concepts with clinicians or therapists. Such a multidisciplinary approach is crucial to bring this research field forward. For example, as developmental speech and language disorders often co-occur with (symptoms of) other neurodevelopmental disorders, it is challenging to correctly diagnose and differentiate on a phenotype level between cognitive problems, behavioural problems and speech/language problems. A multidisciplinary approach, involving speech therapists and neuropsychologists with experience with this complex group of phenotypes, would allow the selection of appropriate instruments and test settings. Therefore the establishment of expert centres with multidisciplinary teams and international embedding in EU reference networks is crucial for connecting expert clinical knowledge to fundamental research.

Data sharing efforts such as GeneMatcher[19] have greatly enhanced the identification of new Mendelian disorders, similar to other genomic collaborative efforts such as the gnomAD database[55]. Especially in the context of rare variants and rare disorders, data sharing and open science practices are of paramount importance for the progress in this research field.

Last but not least, collaboration with parents and families involved is highly recommended. Families have more experience with a disorder than most researchers will ever have, and can provide researchers and clinicians with useful suggestions and insights. Due to social media, parents from individuals with rare diseases all around the world can easily get in contact with each other and share and collect information. In addition to collecting interesting phenotype details via families (and clinicians) involved, we have collaborated with parents in our research projects to optimize study designs before execution (SATB2). For the FOXP4 study, it was the parents contacting us about an interesting FOXP4 variant found in their child, which initiated this research project. Altogether, empowering patients and families in research designs and execution will benefit all parties in performing meaningful research with both scientific and social impact.

References

1. Eising E, Carrion-Castillo A, Vino A, et al. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol Psychiatry.* 2019;24(7):1065-1078.
2. Hildebrand MS, Jackson VE, Scerri TS, et al. Severe childhood speech disorder: Gene discovery highlights transcriptional dysregulation. *Neurology.* 2020.
3. Thomason A, Pankey E, Nutt B, Caffrey AR, Zarate YA. Speech, language, and feeding phenotypes of SATB2-associated syndrome. *Clin Genet.* 2019;96(6):485-492.
4. Newbury DF, Monaco AP. Genetic advances in the study of speech and language disorders. *Neuron.* 2010;68(2):309-320.
5. Posey JE, O'Donnell-Luria AH, Chong JX, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genet Med.* 2019;21(4):798-812.
6. Vissers LE, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. *Nat Genet.* 2010;42(12):1109-1112.
7. Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nat Rev Genet.* 2012;13(8):565-575.
8. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature.* 2017;542(7642):433-438.
9. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol.* 2019;431(11):2197-2212.
10. Iqbal S, Perez-Palma E, Jespersen JB, et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc Natl Acad Sci U S A.* 2020.
11. Sunyaev SR. Inferring causality and functional significance of human coding DNA variants. *Hum Mol Genet.* 2012;21(R1):R10-17.
12. Richards S, Aziz N, Bale S, et al. 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.* 2015;17(5):405-424.
13. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. *Nat Methods.* 2015;12(3):203-206, 204 p following 206.
14. Marton RM, Pasca SP. Organoid and Assembloid Technologies for Investigating Cellular Crosstalk in Human Brain Development and Disease. *Trends Cell Biol.* 2020;30(2):133-143.
15. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet.* 2009;84(4):524-533.
16. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062-D1067.
17. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat.* 2003;21(6):577-581.
18. Turner TN, Yi Q, Krumm N, et al. denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* 2017;45(D1):D804-D811.
19. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat.* 2015;36(10):928-930.
20. Kaufman L, Ayub M, Vincent JB. The genetic basis of non-syndromic intellectual disability: a review. *J Neurodev Disord.* 2010;2(4):182-209.
21. Vissers LE, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet.* 2016;17(1):9-18.
22. Drivas TG, Li D, Nair D, et al. A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome. *Eur J Hum Genet.* 2020;28(10):1422-1431.
23. Miniscalco C, Nygren G, Hagberg B, Kadesjo B, Gillberg C. Neuropsychiatric and neurodevelopmental outcome of children at age 6 and 7 years who screened positive for language problems at 30 months. *Dev Med Child Neurol.* 2006;48(5):361-366.
24. Castel SE, Cervera A, Mohammadi P, et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet.* 2018;50(9):1327-1334.
25. Pizzo L, Jensen M, Polyak A, et al. Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genet Med.* 2019;21(4):816-825.

26. Davies RW, Fiksinski AM, Breetvelt EJ, et al. Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nat Med.* 2020.
27. Vogt G. Stochastic developmental variation, an epigenetic source of phenotypic diversity with far-reaching biological consequences. *J Biosci.* 2015;40(1):159-204.
28. Fisher SE. Tangled webs: tracing the connections between genes and cognition. *Cognition.* 2006;101(2):270-297.
29. Fisher SE. Evolution of language: Lessons from the genome. *Psychon Bull Rev.* 2017;24(1):34-40.
30. Reuter MS, Riess A, Moog U, et al. FOXP2 variants in 14 individuals with developmental speech and language disorders broaden the mutational and clinical spectrum. *J Med Genet.* 2017;54(1):64-72.
31. Myers SM, Challman TD, Bernier R, et al. Insufficient Evidence for “Autism-Specific” Genes. *Am J Hum Genet.* 2020;106(5):587-595.
32. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779.
33. Bishop DV, Snowling MJ, Thompson PA, Greenhalgh T, consortium C. CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. *PLoS One.* 2016;11(7):e0158753.
34. Bishop DVM, Snowling MJ, Thompson PA, Greenhalgh T, and the C-c. Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *J Child Psychol Psychiatry.* 2017;58(10):1068-1080.
35. Geschwind N. The organization of language and the brain. *Science.* 1970;170(3961):940-944.
36. Hagoort P. The neurobiology of language beyond single-word processing. *Science.* 2019;366(6461):55-58.
37. Tremblay P, Dick AS. Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain Lang.* 2016;162:60-71.
38. Fisher SE, Marcus GF. The eloquent ape: genes, brains and the evolution of language. *Nat Rev Genet.* 2006;7(1):9-20.
39. Dick AS, Bernal B, Tremblay P. The language connectome: new pathways, new concepts. *Neuroscientist.* 2014;20(5):453-467.
40. Arbib MA. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behav Brain Sci.* 2005;28(2):105-124; discussion 125-167.
41. Rousso DL, Pearson CA, Gaber ZB, et al. Foxp-mediated suppression of N-cadherin regulates neuroepithelial character and progenitor maintenance in the CNS. *Neuron.* 2012;74(2):314-330.
42. McEvilly RJ, de Diaz MO, Schonemann MD, Hooshmand F, Rosenfeld MG. Transcriptional regulation of cortical neuron migration by POU domain factors. *Science.* 2002;295(5559):1528-1532.
43. Poss ZC, Ebmeier CC, Taatjes DJ. The Mediator complex and transcription regulation. *Crit Rev Biochem Mol Biol.* 2013;48(6):575-608.
44. Zhang Y, LeRoy G, Seelig HP, Lane WS, Reinberg D. The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. *Cell.* 1998;95(2):279-289.
45. Guarnaccia AD, Tansey WP. Moonlighting with WDR5: A Cellular Multitasker. *J Clin Med.* 2018;7(2).
46. Graham SA, Fisher SE. Decoding the genetics of speech and language. *Curr Opin Neurobiol.* 2013;23(1):43-51.
47. Takahashi K, Liu FC, Hirokawa K, Takahashi H. Expression of Foxp4 in the developing and adult rat forebrain. *J Neurosci Res.* 2008;86(14):3106-3116.
48. Estruch SB, Graham SA, Deriziotis P, Fisher SE. The language-related transcription factor FOXP2 is post-translationally modified with small ubiquitin-like modifiers. *Sci Rep.* 2016;6:20911.
49. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. *Eur J Hum Genet.* 2006;14(5):535-542.
50. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet.* 2006;43(8):691-698.
51. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature.* 2001;413(6855):519-523.
52. Wada T, Suzuki S, Shioda N. 5-Aminolevulinic acid can ameliorate language dysfunction of patients with ATR-X syndrome. *Congenit Anom (Kyoto).* 2020;60(5):147-148.
53. Kohler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 2019;47(D1):D1018-D1027.
54. Morgan AT, Haaften LV, van Hulst K, et al. Early speech development in Koolen de Vries syndrome limited by oral praxis and hypotonia. *Eur J Hum Genet.* 2018;26(1):75-84.
55. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443.

Appendices

English summary

While most children acquire speech and language capacities relatively effortlessly, some children have speech or language problems. If these problems persist, a child might have a developmental speech or language disorder. Developmental speech and language disorders are neurobiological disorders, and children with these disorders might have problems with different aspects of speech and/or language and with different levels of severity. Developmental speech and language disorders often co-occur with other neurodevelopmental disorders, such as Attention Deficit Hyperactivity Disorder (ADHD), delayed motor development and developmental dyslexia. The presence of other neurodevelopmental disorders might complicate diagnosis and therapeutic interventions.

In **Chapter 1** we provide an overview of the current literature on developmental speech and language disorders. We discuss the clinical presentation, the commonly used terminology and diagnostic categorization, and the current knowledge on the genetics of these disorders. It is clear from family and twin studies that developmental speech and language disorders are strongly influenced by genetics. While the underlying genetic architecture might be complex and multifactorial, the exact contribution of Mendelian causes to developmental speech and language disorders is currently unclear.

The aim of this thesis was to investigate Mendelian causes of developmental speech and language disorders, in order to better understand the molecular underpinnings of these disorders and ultimately improve clinical care. We studied the impact of rare *de novo* variants from a clinical and molecular perspective, using a wide range of methods including next generation sequencing, systematic phenotyping and functional assessments with cell-based assays. Using this approach, we identified and characterized five new Mendelian disorders with speech and language impairments as a core phenotypic feature (**Chapter 2-6)**, and performed detailed speech, language and neuropsychological profiling in *SATB2*-associated syndrome, a known disorder associated with severe speech problems **(Chapter 7)**. We also designed a prospective cohort study **(Chapter 8)** to systematically investigate the role of *de novo* variants in the emergence of severe developmental language disorder (DLD).

Chapter 2 describes our research on pathogenic variants in *MED13,* a gene encoding a component of the CDK8-kinase module that can reversibly bind with the Mediator complex. We collected clinical data and variant details on 13 individuals with pathogenic variants in this gene and defined a new neurodevelopmental disorder characterized by intellectual disability (or general developmental delays) and speech/language delays or disorders. Additional features reported in a subset of individuals were autism spectrum disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD), hypotonia, optic nerve abnormalities, Duane anomaly and mild congenital heart defects. In addition to six truncating variants, we reported a clustering of non-truncating variants (missense variants and an in-frame deletion of one amino acid) in two regions of the encoded MED13 protein: an N-terminal

and C-terminal region. The four N-terminal clustering mutations affect two adjacent amino acids that are known to be involved in MED13 ubiquitination and degradation, p.Thr326 and p.Pro327. All in all, our findings add *MED13* to the group of CDK8-kinase module-associated disease genes.

In **Chapter 3** we present the results of our study on *POU3F3* (also known as *Brain-1*), a gene encoding a transcription factor that is highly important in gene regulation during early stages of brain development. We compiled a cohort of 19 individuals with heterozygous truncating or missense variants in this gene, and found that all individuals had developmental delays and/or intellectual disability, with impairments in speech and language skills. The majority of individuals had remarkable ear abnormalities: low-set, prominent and often cupped ears. Using luciferase assays, we showed that two missense variants reduced the transactivation capacity of the encoded transcription factor, while one variant displayed a gain-of-function effect. In bioluminescence resonance energy transfer (BRET) interaction assays, all the truncated POU3F3 versions showed significantly impaired dimerization capacities, whereas all missense variants had unaffected dimerization with wild-type POU3F3. All in all, we implicate disruptions of *POU3F3* in a characteristic and potentially recognizable neurodevelopmental disorder.

The identification of pathogenic variants in *CHD3* as a cause of a developmental disorder with macrocephaly and prominent speech impairment is described in **Chapter 4**. As part of a large international collaboration we collected phenotype and variant data on 35 individuals with *de novo* variants in *CHD3*, a gene encoding a protein involved in chromatin remodeling via nucleosome sliding. Individuals with pathogenic *CHD3* variants showed developmental delays or intellectual disability with varying degrees of severity, and often macrocephaly and prominent speech problems. A remarkable facial phenotype was also observed in the majority of affected individuals. Most individuals had missense variants clustering within the ATPase/helicase domain of CHD3. We performed experimental assays measuring ATPase enzyme activity and nucleosome sliding capacities in order to determine pathogenicity and better understand the molecular mechanisms underlying this disorder.

Chapter 5 shows the results of our research on WDR5, a highly conserved protein involved in a wide array of biological functions, mostly as part of protein complexes (e.g. the COMPASS complex) affecting gene regulation via post-translational modification of histones. We collected clinical data on ten unrelated individuals with six different *de novo* variants in *WDR5*. We described the phenotypic spectrum of the newly identified *WDR5*-associated neurodevelopmental disorder, which includes speech and language delays, intellectual disability, epilepsy, autism spectrum disorder, abnormal growth parameters, heart abnormalities and hearing loss. Three-dimensional protein structures indicated that all residues affected by missense variants are located at the surface of one side of the WDR5 protein, and we predicted that five out of the six found variants might disrupt interactions with RbBP5 and/or KMT2A/C. Our research on WDR5 highlights the important role of COMPASS family proteins in neurodevelopmental disorders.

While heterozygous variants in various *FOXP* genes cause developmental disorders, the phenotype associated with pathogenic variants in *FOXP4* has not been previously described. In **Chapter 6** we studied *FOXP4* by assembling a cohort of eight individuals with heterozygous and mostly *de novo* variants in this gene. We collected clinical data for six individuals: five individuals with a missense variant in the forkhead box domain of FOXP4 and one individual with a truncating variant. Overlapping features included speech and language delays, growth abnormalities, congenital diaphragmatic hernia, cervical spine abnormalities and ptosis. Luciferase assays showed loss-of-function effects for all missense variants in the forkhead box domain, while BRET assays showed intact dimerization capacities. In short, our findings show that, in addition to *FOXP1* and *FOXP2,* loss-of-function of FOXP4 is associated with an autosomal dominant developmental disorder with speech and language delays.

Besides the identification and characterization of new Mendelian disorders, in **Chapter 7** we investigated speech and language impairments and neuropsychological functioning in a known neurodevelopmental disorder: *SATB2*-associated syndrome (SAS). We performed detailed oral motor, speech and language profiling in combination with neuropsychological assessments in 23 individuals with a molecularly confirmed SAS diagnosis. We identified severe language delays in all individuals, defined underlying speech conditions for all verbal individuals and a common profile of adaptive functioning in this disorder. In addition to overlapping and recurrent phenotypes, we also observed a high variability, mainly in severity of features. This study gives more insight in speech, language and neuropsychological functioning in individuals with SAS, and provides families, therapists and other caregivers with information to guide diagnostic and treatment approaches.

While **Chapter 2-7** shows that there is no doubt that rare pathogenic variants can cause severe speech and language impairments (including DLD), we still know little about the proportion of children with a DLD that carry a monogenic causal variant. In **Chapter 8** we describe the design of a prospective cohort study, the GENTOS study, to systematically assess the diagnostic yield of whole genome sequencing in DLD. With this study we also aim to identify additional genes in which variants can cause DLD, the corresponding molecular pathways and the type of genetic variation involved. Knowledge of frequencies and types of underlying molecular causes will lead to a better understanding of the genetic architecture in DLD, but is also crucial to better guide referral strategies and diagnostic decisions. Inclusion for the GENTOS study started in March 2020 and is still ongoing at the time of writing.

In **Chapter 9** we summarize the results of our research and describe the clinical relevance and implications. We discuss how we studied *de novo* variants in children with developmental speech and language disorders, to identify five new neurodevelopmental disorders with associated phenotypes not restricted to speech and language disorders. In addition to similar molecular disruptions giving rise to different phenotypes in different probands, we showed how different underlying mechanisms and pathways can lead to similar disturbances in speech and language. We did not find any evidence for the existence of “speech and language-disorder specific” genes, but we observed that pathogenic variants in a subset of the total group of neurodevelopmental disorder genes, including the genes characterized in this thesis, give rise to neurodevelopmental disorders with prominent speech and language dysfunction. For the classification of genetic disorders of speech and language, we think that clinical practice is better suited to a model which acknowledges the diversity of features associated with variants in a given gene, rather than a model in which variants in any single gene lead to a highly specific neurodevelopmental phenotype.

For the future, we suggest not to be reluctant to use next generation sequencing in individuals who present mainly or solely with developmental speech and language disorder. A molecular diagnosis can benefit affected individuals and their families, as it can provide clarity about the cause of the disorder, recurrence risk and the incidence of possible relevant co-morbidities. To improve genetics research in the speech and language field, systematic and consistent phenotyping is of paramount importance and remains one of the big challenges. Furthermore, research consortia, data sharing and patient empowerment are highly recommended and will benefit all parties in performing meaningful research with large scientific and social impact.

Nederlandse samenvatting

De meeste kinderen verwerven zonder al te veel moeite spraak- en taalvaardigheden, maar voor sommige kinderen geldt dit niet: zij hebben problemen met de spraak- en taalontwikkeling. Als deze problemen aanhouden, kan een kind een spraak- of taalontwikkelingsstoornis hebben. Spraak- en taalontwikkelingsstoornissen zijn neurobiologische ontwikkelingsstoornissen. Kinderen met deze aandoeningen kunnen problemen hebben met verschillende aspecten van de spraak- en taalontwikkeling, en ook de ernst van de problemen kan zeer variabel zijn. Spraak- en taalontwikkelingsstoornissen komen vaak samen voor met andere ontwikkelingsstoornissen zoals bijvoorbeeld ADHD, een vertraagde motorische ontwikkeling en dyslexie. De aanwezigheid van andere ontwikkelingsstoornissen kan het stellen van een juiste diagnose of het toepassen van de juiste behandeling soms bemoeilijken.

In **Hoofdstuk 1** geven we een overzicht van de huidige literatuur over spraak- en taalontwikkelingsstoornissen. We beschrijven de klinische presentatie, de veelgebruikte terminologie en diagnostische categorisatie, en de huidige kennis over de genetische achtergrond van deze aandoeningen. Vanuit familiestudies en tweelingstudies is al bekend dat het ontstaan van spraak- en taalontwikkelingsstoornissen sterk wordt beïnvloed door genetische factoren. De onderliggende genetische basis is echter vaak complex en multifactorieel, en de precieze bijdrage van Mendeliaans overervende DNA-varianten aan het ontstaan van spraak- en taalontwikkelingsstoornissen is tot op heden onbekend.

Het doel van dit proefschrift was om Mendeliaans overervende oorzaken van spraak- en taalontwikkelingsstoornissen te bestuderen, om de moleculaire achtergrond van deze aandoeningen beter te begrijpen om zo uiteindelijk de zorg te kunnen verbeteren. We bestudeerden hiervoor de impact van zeldzame *de novo* varianten vanuit een klinisch en moleculair perspectief, en maakten hierbij gebruik van veel verschillende methoden zoals *next generation sequencing*, systematisch fenotyperen en functionele laboratoriumexperimenten. Door deze aanpak hebben we vijf nieuwe Mendeliaans overervende aandoeningen ontdekt en gekarakteriseerd waarbij spraak- en taalproblemen een belangrijk symptoom zijn (**Hoofdstuk 2-6**), en hebben we het spraak-, taal en neuropsychologisch profiel van SATB2-geassocieerd syndroom (een bekende aandoening waarbij ernstige spraakproblemen op de voorgrond staan) uitgebreid in kaart gebracht (**Hoofdstuk 7**). We hebben ook een prospectieve cohort studie opgezet (**Hoofdstuk 8**) waarmee op een systematische manier de bijdrage van *de novo* varianten aan het ontstaan van taalontwikkelingsstoornissen (TOS) kan worden onderzocht.

Hoofdstuk 2 beschrijft ons onderzoek naar pathogene varianten in *MED13*, een gen dat codeert voor een component van de CDK8-kinase module, dat reversibel kan binden met het zogenoemde *Mediator complex*. We hebben klinische gegevens en gegevens over de gevonden varianten verzameld van 13 individuen met pathogene varianten in dit gen, en zo hebben we een nieuwe ontwikkelingsstoornis beschreven, die wordt gekarakteriseerd door een verstandelijke beperking (of algehele ontwikkelingsachterstand) en spraak- en taalstoornissen. Bijkomende kenmerken die werden gezien in een deel van de personen met varianten in *MED13* waren autismespectrumstoornis, ADHD, hypotonie, afwijkingen aan de oogzenuw, Duane anomalie en milde aangeboren hartafwijkingen. Naast zes truncerende varianten vonden we ook een clustering van niet-truncerende varianten (missense varianten en een in-frame deletie van één aminozuur) in twee verschillende regio's van het MED13 eiwit: een N-terminale en een C-terminale regio. De vier mutaties in de N-terminale regio beïnvloeden twee naast elkaar gelegen aminozuren waarvan bekend is dat zij een rol spelen bij de ubiquitinatie en degradatie van MED13; p.Thr326 en p.Pro327. Al met al kan door ons onderzoek *MED13* worden toegevoegd aan de groep van 'CDK8-kinase module-associated disease genes'.

In **Hoofdstuk 3** presenteren we de resultaten van onze studie over *POU3F3* (ook bekend als *Brain-1*), een gen dat codeert voor een transcriptiefactor die essentieel is voor een goede genregulatie tijdens de vroege hersenontwikkeling. We vonden 19 personen met heterozygote truncerende varianten of missense varianten in dit gen, en stelden vast dat al deze personen een ontwikkelingsachterstand en/of verstandelijke beperking hadden met hierbij ook problemen met de spraak- en taalvaardigheden. De meerderheid van deze personen had opvallende oorafwijkingen: laagstaande, prominente oren en vaak zogenoemde *cupped ears*. Met luciferase experimenten konden we aantonen dat twee missense varianten de transactivatie capaciteit van de transcriptiefactor verlaagden, terwijl een andere variant juist een toegenomen functie liet zien. Met *bioluminescence resonance energie transfer* (BRET) interactie experimenten zagen we dat alle getrunceerde versies van POU3F3 een significant verminderde dimerisatie lieten zien, terwijl alle missense varianten wel een normale dimerisatie met POU3F3 vertoonden. Alles bijeengenomen toonden wij aan dat mutaties in *POU3F3* een karakteristieke en mogelijk ook herkenbare ontwikkelingsstoornis kunnen veroorzaken.

De identificatie van pathogene varianten in *CHD3* als oorzaak voor een ontwikkelingsstoornis met macrocefalie en opvallende spraakproblemen is beschreven in **Hoofdstuk 4**. Als onderdeel van een grote internationale samenwerking verzamelden we gegevens over fenotypes en varianten van personen met een *de novo* variant in *CHD3*. *CHD3* is een gen dat codeert voor een eiwit dat betrokken is bij *chromatin remodeling* via *nucleosome sliding*. Personen met een pathogene variant in *CHD3* hadden een ontwikkelingsachterstand of verstandelijke beperking met verschillende gradaties van ernst, en vaak ook macrocefalie en duidelijke spraakproblemen. Bij de meerderheid werden ook opvallende uiterlijke kenmerken gezien. De meeste personen in onze studie hadden missense varianten, clusterend in het ATPase/helicase domein van CHD3. Met functionele experimenten hebben wij de de ATPase enzymactiviteit en de *nucleosome sliding* capaciteit van verschillende varianten bepaald om zo de pathogeniciteit te kunnen beoordelen, en zo de moleculaire mechanismen van deze aandoening beter te kunnen begrijpen.

Hoofdstuk 5 laat de resultaten zien van ons onderzoek naar WDR5, een sterk geconserveerd eiwit betrokken bij zeer veel verschillende biologische processen, meestal als onderdeel van een eiwitcomplex (bijvoorbeeld het COMPASS complex) dat genregulatie beïnvloedt via prosttranslationele modificatie van histonen. We verzamelden klinische gegevens van tien personen met zes verschillende *de novo* varianten in *WDR5*. We beschreven het spectrum van fenotypes geassocieerd met deze nieuwe ontwikkelingsstoornis veroorzaakt door *WDR5*-varianten, waarbij o.a. een achterlopende spraak- en taalontwikkeling, verstandelijke beperking, epilepsie, autismespectrumstoornissen, abnormale groeiparameters, hartafwi-jkingen en gehoorverlies werden gezien. Door de varianten te visualiseren in driedimensionale eiwitstructuren zagen we dat alle betrokken aminozuren (aminozuren beïnvloed door missense varianten) aan het oppervlak van één kant van het WDR5 eiwit lagen, en voorspelden we dat vijf van de zes gevonden varianten de interactie met RbBP5 en/of KMT2A/C verstoren. Ons onderzoek over WDR5 onderstreept de belangrijke rol van eiwitten uit de 'COMPASS complex'-familie in ontwikkelingsstoornissen.

Hoewel heterozygote varianten in verschillende *FOXP* genen ontwikkelingsstoornissen kunnen veroorzaken, was het fenotype veroorzaakt door pathogene varianten in *FOXP4* tot nu toe nog niet beschreven. In **Hoofdstuk 6** bestudeerden we *FOXP4* door een cohort van acht personen met een heterozygote en meestal *de novo* variant in dit gen samen te stellen. We verzamelden klinische gegevens over zes personen: vijf personen met een missense variant in het forkhead box domein van FOXP4 en een persoon met een truncerende variant. Overlappende kenmerken waren spraak- en taalproblemen, afwijkende groeiparameters, een aangeboren breuk in het middenrif (hernia diafragmatica), afwijkende nekwervels en een hangend ooglid (ptosis). Luciferase experimenten lieten een *loss-of-function* effect zien voor alle missense varianten in het *forkhead box* domein, terwijl BRET experimenten voor deze varianten een normale dimerisatie capaciteit liet zien. Al met al toonden wij aan dat dat, naast *FOXP1* en *FOXP2*, een verlies van functie van *FOXP4* ook geassocieerd is met een autosomaal dominante ontwikkelingsstoornis met taal- en spraakproblemen.

Naast het identificeren en karakteriseren van nieuwe Mendeliaans overervende aandoeningen, onderzochten we in **Hoofdstuk 7** de spraak- en taalstoornissen en het neuropsychologisch functioneren in een bekende ontwikkelingsstoornis: SATB2-geassocieerd syndroom (SAS). We hebben hiervoor gedetailleerd logopedisch onderzoek verricht van zowel de mondmotoriek, spraak- en taalvaardigheden, in combinatie met neuropsychologische testen in 23 personen met een moleculair bevestigde diagnose SAS. We stelden vast dat alle personen ernstige taalachterstanden hadden, konden onderliggende spraakproblemen definiëren voor alle verbale personen, en zagen een overeenkomend profiel voor het adaptief functioneren bij deze aandoening. In aanvulling op de overlappende en terugkerende fenotypes was er ook een sterke variabiliteit, met name in ernst, van de kenmerken van deze aandoening. Deze studie geeft meer inzicht in het functioneren op spraak-, taal en neurologisch gebied in personen met SAS en geeft families, behandelaars en andere zorgverleners informatie die van belang kan zijn voor goede diagnostiek, behandeling en begeleiding.

Hoewel **Hoofdstuk 2-7** laten zien dat er geen twijfel is dat zeldzame pathogene varianten ernstige spraak- en taalstoornissen (inclusief TOS) kunnen veroorzaken, is het nog niet duidelijk hoeveel kinderen met TOS zo'n monogene oorzakelijke DNA-variant hebben. In **Hoofdstuk 8** beschrijven wij de opzet van een prospectieve cohort studie, de GENTOS studie, om systematisch de 'diagnostische opbrengst' van genoomsequencing bij TOS te onderzoeken. Met deze studie hopen we nog meer genen te vinden waarin varianten TOS kunnen veroorzaken, evenals de bijbehorende moleculaire *pathways* en het type varianten dat hierbij betrokken is. Meer kennis over frequentie van genetische aandoeningen en de soorten genetische varianten is belangrijk om de genetische achtergrond van TOS te begrijpen, maar is ook cruciaal voor het bepalen van de juiste verwijsstrategieën en diagnostische beslissingen. Inclusie voor de GENTOS studie is gestart in maart 2020 en is nog steeds bezig op het moment van schrijven.

In **Hoofdstuk 9** vatten we de resultaten van ons onderzoek samen en beschrijven we de klinische relevantie en implicaties. We beschrijven hoe we *de novo* varianten in kinderen met spraak- en taalontwikkelingsstoornissen bestudeerden, en zo vijf nieuwe ontwikkelingsstoornissen hebben geïdentificeerd waarbij het geassocieerde fenotype breder was dan alleen spraak- en taalstoornissen. We lieten zien hoe vergelijkbare moleculaire defecten verschillende fenotypes kunnen geven bij verschillende personen, en hoe verschillende onderliggende moleculaire mechanismen vergelijkbare stoornissen in de spraak- en taalontwikkeling kunnen veroorzaken. We hebben geen bewijs gevonden voor het bestaan van 'spraak- en taalstoornis-specifieke genen', maar we zagen wel dat pathogene varianten in een deel van de totale groep van genen geassocieerd met ontwikkelingsstoornissen, inclusief de genen gekarakteriseerd in dit proefschrift, ontwikkelingsstoornissen kunnen veroorzaken met prominente spraak- en taalproblemen. Voor de classificatie van genetische stoornissen met spraak- en taalproblematiek, denken wij dat een model dat de diversiteit van kenmerken geassocieerd met varianten in een specifiek gen erkent beter aansluit bij de klinische praktijk dan een model waarbij varianten in een specifiek gen altijd leiden tot een heel specifiek fenotype.

Voor de toekomst is het wat ons betreft belangrijk om niet terughoudend te zijn met het gebruik van *next generation sequencing* in personen waarbij een spraak- of taalontwikkelingsstoornis op de voorgrond staat. Een genetische diagnose kan voor aangedane personen en hun familieleden belangrijk zijn, omdat het duidelijkheid kan geven over de oorzaak van de problemen, de herhalingskans en over het optreden van mogelijk relevante co-morbiditeit. Om wetenschappelijk onderzoek over genetica van spraak- en taalstoornissen te verbeteren, is systematisch en consistente fenotypering cruciaal, dit blijft één van de grootste uitdagingen. Daarnaast zijn research consortia, het delen van data en *patient empowerment* belangrijk, en kan dit alle partijen ten goede komen bij het verrichten van betekenisvolle studies met grote wetenschappelijke en sociale impact.

Dankwoord

Daar is het moment dan eindelijk, mijn boekje is af! En dit boekje was absoluut nooit tot stand gekomen zonder de steun en bijdragen van heel veel verschillende mensen op heel veel verschillende momenten. Ik voel me vereerd met zoveel mooie, leuke en getalenteerde mensen te hebben gewerkt (of nog steeds te werken), maar ook om buiten het werk zoveel fijne mensen om mij heen te hebben…dus bij deze alvast voor iedereen die daar ook maar op enige wijze aan heeft bijgedragen: mijn dank is groot!

Allereerst natuurlijk heel veel dank voor mijn promotieteam: Han, Simon en Tjitske. Han, wat heb ik veel van jou geleerd, en wat was het fijn om jou als promotor te hebben. Als ik iets moeilijk vond, gefrustreerd was over een samenwerking of situatie, of even niet wist hoe verder kon ik altijd bij jou terecht en nam je me altijd serieus. Jouw brede kennis over de genetica-geschiedenis maar ook visie op de toekomst van ons vakgebied is een fantastische bron van inspiratie voor elke geneticus-in-spé. Je nuchtere e-mails met toepasselijke citaten of slechte grappen waren een zeer welkome verademing in de soms rare wereld van wetenschappelijke tijdschriften, ethische commissies en bureaucratische procedures. Ik vergeet nooit meer hoe ik onderweg naar de ESHG in Göteborg een mail kreeg van jou, als reactie op de mededeling dat onze studie-opzet wéér niet was goedgekeurd door de ethische commissie, met daarin: ‘als ik je zie krijg je een ijsje’. Precies wat ik nodig had ☺ (niet dat ijsje, wel die e-mail).

Simon, it was an honour to be able to work in your lab and to be a member of the very famous Protein Group. You, as well as all the other Language & Genetics people, have created a great atmosphere and a very safe space for me to develop my wetlab skills and research skills in general. I admire your knowledge on a wide range of genetics-related topics, varying from basic molecular biology, cell-based assays and animal studies to complex trait analyses and evolutionary genetics, in combination with the fact that you are always open for questions and discussions. Over the last four-and-a-half years, you have patiently corrected all my English mistakes and every time I am writing an English text now, I think about your text balloons with comments like: “the word *data* is plural”. I also want to specifically thank you and Han for the endless support and trust in me in the most difficult months of my PhD in May and June 2017.

Tjitske, al sinds ik als co-assistent bij de Genetica rondliep werken we samen aan onderzoeksprojecten. Jij hebt mij, altijd geduldig, de eerste kneepjes van onderzoek doen geleerd, terwijl ik zelf soms nog geen idee had waar ik eigenlijk mee bezig was. Nadat het DDX3X project een beetje uit de hand liep qua grootte, wist ik zeker dat ik verder wilde met onderzoek, en het was ontzettend fijn om dit onder jouw begeleiding te kunnen doen. Ik vind het bewonderenswaardig hoe je altijd voor mij klaar stond en klaar staat, ook als jouw thuissituatie het eigenlijk niet altijd toe laat, of als ik ’s avonds om 22.00 een artikel probeer te submitten en toch nog ineens iets van je nodig heb. Ik hoop dat we nog vele jaren mogen samenwerken en de 'Friese vibes' op de afdeling hoog kunnen houden!

I want to sincerely thank all the wetlab members in the Max Planck Institute, for helping me out and introducing me to the fantastic world of molecular genetics. I have to admit I was a bit nervous, at my first morning, when I arrived in the lab around 8.30. But that changed immediately after I met all my new colleagues, and discovered the common lab working ours started around 9.30-10.00, what a blessing! Pela, I could not have wished for a better supervisor in the first year of my PhD. Although I did not even know what a plasmid was when I started, you patiently helped me and your knowledge and experience, alternated with a great sense of humour (and many comments on the Dutch food and healthcare system), made it a pleasure to work with you. I'm still sad that you left me, but I'm very glad you're living the good life now in the UK and you still support me via Whatsapp messages now and then!

Sara, such a pity our time together in the lab was too short, but the 'Lanterneta Paranoica' labels at my bench have always reminded me of our good times (and PLA struggles). I'm looking forward to visiting you (and my best friend Ona!) in Badalona, I'm sure that will happen again someday. Elliot, thank you for all your help and positivity, especially with setting-up BRET experiments. Ary, I would simply not have survived in the lab without you. You are the corner stone of the protein group, always close by to help out, when I couldn't find something, had no clue what I was doing or get the best tips & tricks to do everything faster/smarter/better (or to provide me with Taralli or self-made cake). Your creativity with cooking skills/birthday surprises/experiment solutions is unbelievable. And outside the lab we had so many good times too, with the 'Dutch songs playlist' at your wedding in Italy as one of the highlights. Laura, thank you for being a great desk-neighbor and a Pancake Day co-organizer. We should still start filming for our 'do's and don'ts series'; great lab life hacks (by Laura) and how to screw up your co-IP or western blot (by Lot). Joery, of course you deserve a very special thanks too. As a Master student, and later as a PhD colleague, you have taught me so much about lab skills and experimental designs (and about ants, not to forget). After Pela left, I appointed you as my 'lab supervisor' and you took that job seriously. You are a great colleague and friend, and I'm sure you will have a great career ahead of you, with lots of 'kwark' and peanut butter! Maggie, I am so happy that you decided to join our group too. Although I still cannot handle the idea of 'Essence of Chicken' or 'BBQ-fried dough', it was great to have you around and to have so many funny discussions on science or life in general. Roos, although our time in the lab together was short, I am grateful for all the cells you split for me (and the good conversations of course), and I'm sure we will see each other in future, either on a bike or maybe as a new colleague? I also want to thank many other MPI colleagues: Fabian, Midas, Paolo, Elpida, Kai, Moritz, Janine, Ine, Jelle, Cleo, Karthik, Jasper, Jurgen, Else, Ellen, Martina and ………. (please fill in you own name if I accidentally forgot you), for all the help, for great collaborations, fun Christmas parties, Friday pizza events, random ice cream moments and so much more!

Het voordeel aan twee werkplekken is dat je ook dubbel zoveel leuke collega's kunt hebben. De sectie Genetica in het Radboud is een warm bad waarin ik me ontzettend thuis voel, en dat komt vooral door al die leuke mensen die daar rondlopen. Ik wil alle onderzoekers van de Klinische Genetica bedanken: Linde, Elke, Chantal, Janet, Joost, Jeroen, Dmitrijs, Bianca, Sandra, Jolijn, maar ook iedereen die al lang weer weg is en natuurlijk de nieuwe lichting die nu de onderzoekskamer bemant. Linde, ik hoop dat onze speciale samenwerking (L&L voor al uw DNA en neuropsychologie-vragen) nog lang in stand blijft! Janet, dank voor al je betrokkenheid en ondersteuning bij Biobank-, Castor- of ethische commissie-vragen. Elke, wat een heerlijk persoon ben jij, jammer dat onze kamer zo plotseling werd opgeheven, ik mis de stapels Autodrop en verse dadelvoorraad! Jolijn, jij hebt als student fantastisch werk verricht voor de WDR5 studie, hopelijk mogen we je in de nabije toekomst als arts-assistent of arts-onderzoeker verwelkomen.

En ook al heb ik alle AIOS/ANIOS voor meer dan vier jaar in de steek gelaten, ik wil jullie toch ook bedanken voor alle support. Lex, Ozlem, Illja, Anneke, Femke, Inci, Milou, Jeroen, Erika, Nynke en Thatjana (voor mij hoor je er nog steeds bij hoor!), wat hebben we een fijne groep samen, hopelijk snel weer meer borrels en 'buitenschoolse' activiteiten! Bregje, heel veel dank voor je steun en flexibiliteit als opleider, waardoor ik later terug kon komen als AIOS dan oorspronkelijk gepland. Mede door jou ben ik als student ooit helemaal enthousiast geworden over Genetica, en daar heb ik nog geen moment spijt van gehad. Ik vind het fantastisch dat ik nu onder jouw hoede (en die van Yvonne!) mijn opleiding tot klinisch geneticus mag afmaken. Corrie, ik hoop dat we snel weer eens kunnen lunchen samen, om alle roddels van de afdeling te kunnen doornemen ☺. En natuurlijk voor alle stafleden, PAs/genetisch consulenten, ons fantastische (staf)secretariaat en de onmisbare maatschappelijk werkers, en iedereen die ik nu vergeet: dank voor jullie mental support, inhoudelijke bijdragen, gezellige lunches en voor de altijd fijne sfeer op route 836. Door al dat thuiswerken besef ik des te meer hoe gezegend ik ben met zulke fijne collega's. Laten we snel weer de dansvloer onveilig maken met z'n allen, 'klinische genetica-style'!

Een speciaal bedankje ook voor alle Klinisch Genetiski-deelnemers van de afgelopen jaren. Wat begon als een spontaan plan in de Aesculaaf, is geëindigd in een jaarlijkse traditie die (na een onderbreking in 2021) hopelijk nog héél vaak plaats gaat vinden. Ik kan niet wachten op de volgende Genetiski editie met foute skipakken, heuptasjes en haarbanden, een goede après ski-ski ratio, op slippers naar het 'galadiner' en natuurlijk vele andere legendarische momenten en anekdotes waar we weer een jaar op kunnen teren, maar waar is Joost?

Naast de collega's van de sectie Klinische Genetica zijn er nog heel veel andere geniale Genetica collega's die hebben bijgedragen aan het tot stand komen van mijn proefschrift.

Rolph en Nicole, wat fijn dat ik jullie altijd kan bellen als ik even twijfel over een variant of weer een query wil doen voor een nieuw kandidaat-gen. De afdeling Genetica is nergens zonder jullie! Christian, wat fijn dat ik altijd bij je kan binnenlopen (of eigenlijk loop jij altijd bij ons binnen) om te sparren over van alles en nog wat, variërend van nieuwe ideeën over *de novo varianten* of *reduced penetrance* tot inspiratie voor vakantieplannen of leuke borrels. Lisenka, dank voor al je hulp en kennis over o.a. CMO aanvragen, het is altijd leuk om met jou naar Hinxton of andere congressen te gaan, en ik hoop dat we in de toekomst meer kunnen samenwerken! Alex, jouw kennis over Genetica-land en sequencing-technieken is fantastisch, maar de vele borrels bij jou met over-the-top borrelplanken of versgemaakte pizza's en (te veel) wijn zijn minstens zo fantastisch ☺.

Ik weet nog goed hoe ik na mijn afstuderen twijfelde of ik wel bij de afdeling Genetica wilde gaan werken, want 'daar werken vast alleen maar saaie mensen', dacht ik. Maar na het eerste dagje uit met een afterparty in de Sjors en Sjimmie was ik he-le-maal om. Ik denk met een grote glimlach terug aan alle borrels, uitjes en congressen die mijn PhD tijd hebben verrijkt: het Airbnb huis op een vakantiepark in Orlando (met lichtgevende koelkast en de partybus naar een vakantiehuis-feest), de nachtelijke fietstocht door Göteborg na een prachtige ESHG-feestavond, de 'Nijmegen houses' met bijbehorende 'house parties', de kerstdiners met afterparty in de Malle Babbe, het jaarlijkse dagje uit (maar zo leuk als het dagje uit 2015 wordt het helaas nooit meer, toch Martijn?), 'spontane Genetica avonden' tijdens de Vierdaagse en natuurlijk de vrijdagmiddagborrels in de Aesculaaf. Ik kan me niet meer voorstellen dat ik ooit dacht dat Genetica-mensen saai waren. A big thanks for all current or former Human Genetics colleagues making this all possible: Marloes, Martijn, Margot, Laurens, Stefan, Petra, Manon, Jakob, Judith, Ideke, Roos, Ralph, Susanne, Silvia, Laura, Ralph, and all the others I forget now: whether it's in a serious Theme Discussion, at the dancing floor, in the Aesculaaf, or another random place: It's always fun with you guys around!

Margot, ik ben ontzettend blij dat wij elkaar gevonden hebben in onze beginjaren op de afdeling. Wat hebben we veel meegemaakt samen, daar kunnen we wel een boek over schrijven (hoewel ik nu wel even klaar ben met boeken schrijven). Ik weet nog goed hoe we samen onze eerste artikelen gingen submitten (eindeloze lijsten met co-auteurs invullen) en ons eerste praatje op een congres gingen geven ("haploinsufFICiency"). Ik hoop dat ik tijdens mijn promotie net zo relaxed kan zijn als jij tijdens de jouwe, en vind het fantastisch dat jij als paranimf naast me zult staan. Ik weet zeker dat we elkaar nog heel vaak gaan zien, niet alleen in Genetica-land maar ook daarbuiten!

Karen en Leenke, wat fijn om met jullie samen te kunnen werken, ik kan me geen beter logopedistenteam wensen. Of het nou over wetenschap, promotiefrustraties of niet-werk gerelateerde onderwerpen gaat, het is altijd leuk met jullie! Ik hoop dat we ook de komende jaren nog samen kunnen werken aan research of in de kliniek, en ik weet jullie te vinden als ik weer eens niet snap wat precies het verschil is tussen verschillende fasen of processen van de spraakmotoriek. Max, wat goed dat jij ons uit de brand kwam helpen toen we hard op zoek waren naar een psycholoog voor ons project. We gaan elkaar vast nog zien als jij je vervolgstudie uit gaat voeren!

Aardbananen, lievelingsmensen, Sjoeke, Afke, Lotver, AJ en Jildou: we kennen elkaar al meer dan twintig jaar, en nooit is het saai! Elk etentje, weekendje of reisje met jullie is fantastisch en o zo grappig, een heerlijke afwisseling van het soms serieuze leven. Ook al lopen onze levens soms aardig uiteen en wonen we niet bij elkaar om de hoek, ik weet zeker dat we elkaar altijd zullen blijven vinden, zullen blijven trakteren op random gadgets met aardbeien en bananen, en ooit gaat het lukken met die lustrumreis naar IJsland hoor, ik kijk er naar uit!

Ireen, Yoen, Helma, Dieke, Kars-Jan, Floor, Bob en vele andere oud MFVers en Formosa-vrouwen, wat fijn om jullie nog steeds om mij heen te hebben als vrienden. Thee-avondjes, lunches en borrels zijn altijd een goede afwisseling van werkgerelateerde zaken. Floor, al vanaf het moment dat we samen meekeken bij Formosa hadden we elkaar gevonden. Ook al zijn we zo verschillend, we weten elkaar altijd weer te vinden voor wijze raad, fietsdates of kopjes thee. Ireen, Laura, Marleen, ook al wonen we allemaal niet meer bij elkaar in de buurt, het is altijd leuk om met jullie te zijn! Leonie, ik waardeer je niet-aflatende enthousiasme en positiviteit over zo ongeveer alles, en hoop dat we nog vele fietsuitjes of andere spontane activiteiten kunnen ondernemen. Heleen, Maaike, Leonie: ik heb de Tajine dates gemist in Corona-tijden, ik zal er snel weer een organiseren! Maaike, wat fantastisch om zoveel verschillende dingen met jou te kunnen delen: van hockeyteamgenoten tot huisgenoten, van samen reisleider zijn nu wekelijks samen op de tennisbaan. Ik weet zeker dat je een fantastische paranimf gaat zijn, en kom je ooit nog bij ons werken? ☺

De laatste vijf maanden van mijn PhD contract bestonden met name uit thuiswerken tijdens de ‘Corona-lockdown’, niet helemaal hoe ik het me had voorgesteld. Maar gelukkig waren daar de online pubquizzes, veel dank Martijn, Remi, Alex, Simon, Maaike, Martin en Marloes voor het wekelijkse enthousiasme, en voor het toevoegen van alle kennis over acteurs en films die bij mij (nog steeds, sorry!) ontbreekt. En gelukkig was fietsen nooit verboden tijdens Corona-tijden, en is dat altijd een prachtige afleiding van alles waar je maar afgeleid van wil worden, dus veel dank ook alle KEKkies en andere fietsmaatjes voor de mooie avonturen op de racefiets. Of het nou in de Ooijpolder is, in Girona, de Heuvelrug of Italië, op de fiets is het altijd genieten en zijn ‘PhD problemen’ ver te zoeken.

I want to thank all clinical geneticists and other collaborators that have contributed to all the studies in this thesis. Working in rare disease genetics can only be a success if there is collaboration and trust in each other, and I am very grateful the worldwide Human Genetics community is a great example of this.

Daarnaast kan onderzoek in zeldzame ziekten niet plaatsvinden zonder families die bereid zijn hun eigen gegevens of de gegevens van hun kind te willen delen ‘voor de wetenschap’, bereid zijn om extra naar het Radboudumc te komen voor een studie, die ons op het spoor zetten van een nieuw gen of een nieuwe bevinding bij een bestaand gen… Daarom bij deze voor alle MED13, CHD3, POU3F3, WDR5, FOXP4 en SATB2 families die hebben meegewerkt, en ook de families die nog steeds deelnemen aan de GENTOS studie: Mijn dank is groot, zonder jullie bijdrage had dit proefschrift nooit geschreven kunnen worden!

En als laatste wil ik mijn familie bedanken. Merel en Arjen, Wender en Janneke, ook al wonen we niet super dicht bij elkaar in de buurt, ik ben blij dat ik altijd bij jullie terecht kan voor fietsuitjes, logeeruitjes, om even te blijven eten of gewoon lekker sport kijken op de bank! Ik denk overigens dat Maren, Elin en Tessel misschien wel voor de meeste afleiding hebben gezorgd tijdens de laatste maanden van mijn PhD in het voorjaar van 2020 waarin de wereld vrij klein werd en ik de hoeveelheid schrijfwerk voor mijn proefschrift nog groot was. De wekelijkse speel- en oppasdagjes in Bunnik waren een mooie afwisseling van het serieuze schrijfwerk, en ik ben trots dat ik tante mag zijn van zulke lieve, stoere en schattige nichtjes! Pap en mam, ook voor jullie heel veel dank. Thuiskomen in Sneek voelt altijd nog als thuiskomen, ook al is het huis niet meer hetzelfde. Wat ik ook doe, waar ik ook heen ga, ik weet dat jullie me altijd steunen en trots op me zijn. Mam, ook al weet je soms even niet meer het verschil tussen een ethische commissie of een manuscriptcommissie, je bleef altijd geïnteresseerd informeren naar hoe het met het onderzoek (en vooral met mij) ging. En pap, ik waardeer enorm hoe je altijd probeert te begrijpen wat ik precies heb gepubliceerd (en hoe je mijn artikelen van Twitter plukt als ik er zelf niet over begin!). Maar mijn allergrootste fan is misschien wel mijn oma: Oma Joke. Oma, als er iemand de afgelopen jaren vaak vroeg hoe het met mijn proefschrift ging dan was jij het wel! Je hebt altijd gezegd dat je dit bijzondere moment nog hoopt mee te mogen maken. Lieve oma, deze is voor jou!

Curriculum vitae

Lot Snijders Blok was born on the 24th of April 1987 in Sneek, the Netherlands. In 2005 she graduated with distinction ('cum laude') from the RSG Magister Alvinus in Sneek, after which she started studying Medicine at the Radboud University in Nijmegen. She obtained her Bachelor's degree in 2009 and her Master of Science in Medicine in December 2012. In the last year of her studies, Lot was a senior intern in Neurology at the Rijnstate Hospital in Arnhem, and she did a six-month research internship in the laboratory of John Christodoulou in the Children's Hospital at Westmead, Sydney, Australia. The subject of this internship was 'Flow Cytometry as a diagnostic tool in respiratory chain disorders'. During her Medicine degree Lot was active as a board member of the Medical Faculty Association Nijmegen (MFVN) and the Student organization for Education and Study (SOOS), and she was a member of the Education Committee Medicine.

After graduating as a medical doctor, Lot went back to Australia to work for three months as a research assistant in the laboratory of John Christodoulou. Back in the Netherlands, she started working as a junior resident (ANIOS) Clinical Genetics in the University Medical Center in Utrecht. In December 2013, Lot started as a resident (AIOS) Clinical Genetics in the Department of Human Genetics of the Radboud University Medical Center in Nijmegen, the Netherlands.

In February 2016, Lot paused her training in Clinical Genetics and was appointed as a PhD student at the Department of Human Genetics (Radboud University Medical Center) and Language & Genetics (Max Planck Institute for Psycholinguistics) in Nijmegen, the Netherlands. Her research project was part of the 'Language and Interaction' consortium, and for this project Lot was supervised by Prof. Dr. Han Brunner, Prof. Dr. Tjitske Kleefstra and Prof. Dr. Simon Fisher. As a PhD student Lot cycled back and forth between the two involved research institutes, to study both the clinical characteristics of developmental speech and language disorders as well as the underlying molecular mechanisms. In addition to her research activities, Lot was involved in teaching basic courses in Human Genetics as part of the Bachelor's programs Medicine and Biomedical Sciences.

Over the last years, Lot has presented the results of her research studies at many different national and international conferences, as well as meetings with speech/language therapists and family days for children with developmental disorders. She has been a semi-finalist twice for the American Society of Human Genetics (ASHG) Charles J. Epstein Trainee Award for Excellence in Human Genetics Research, and was awarded with the European Society of Human Genetics (ESHG) Young Investigator Award in 2019.

In July 2020 Lot returned to the clinic, to restart her residency program in Clinical Genetics. She expects to complete her training in 2022, and to remain involved in several research projects and teaching activities during the last years of her training. In her spare time, Lot prefers to be outside on a road bike, mountain bike, hiking boots or a snowboard.

Data management

Ethics

This thesis includes studies with human subjects. All studies were conducted in accordance with the principles of the Declaration of Helsinki (64th WMA General Assembly, Fortaleza, Brazil, October 2013). Written consent for collecting these data was obtained from the participants and/or from their parents or legal representatives. The studies in Chapter 7 and 8 were subject to the 'Medical Research Involving Human Subjects Act' (WMO), and were approved by a Medical Research Ethics committee (CMO Arnhem-Nijmegen) under study number NL64562.091.18 (Chapter 7) and NL67516.091.19 (Chapter 8).

Funding

Funding for this PhD project was provided by the Netherlands Organization for Scientific Research (NOW) Gravitation Grant 24.001.006 to the Language and Interaction Consortium.

FAIR principles

Findable, Accessible

The raw and processed clinical data and accompanying files of Chapter 2-6 in this thesis are stored in a folder on the server of the Human Genetics Department at the Radboud University Medical Center (H:\KG Algemeen\Onderzoek Zeldzaam\Lot). For the study on SATB2 (Chapter 7) and the GENTOS study (Chapter 8), all data were captured in study-specific Castor EDC databases, and for all source data and additional data files protected digital folders were used (H:\KG Algemeen\SPELA-SAS studie and H:\KG Algemeen\GENTOS studie).

All laboratory experimental data are stored in a folder on the server of the Language & Genetics Department at the Max Planck Institute for Psycholinguistics (P:\workspaces\lg-protein-group\archive_deposit\Lot). Cell lines, DNA constructs and primers used are stored in freezers of the Language & Genetics laboratory in the Max Planck Institute for Psycholinguistics. The location of all samples can be found in the Sample Storage folder on the Language & Genetics laboratory server of the Max Planck Institute for Psycholinguistics (P:\shared_spaces\wetlab\6. Laboratory Records).

The digital folders described above are only accessible by members of the research groups involved. All paper files have been digitized, original consent forms have been preserved and are stored in a locked file cabinet at the Clinical Genetics section of the Human Genetics Department. Raw data are available upon request, and if in line with the study-specific consent procedure, can be requested via lot.snijdersblok@radboudumc.nl or secretariaatstafklinischegenetica@radboudumc.nl.

Interpretable, Reusable

All raw data are stored in their original form. A description of the experimental setup can be found in the methods section of each study in this thesis. All laboratory experiments were documented using the Electronic Lab Notebook (ELN) or Labfolder systems of the Max Planck Institute for Psycholinguistics. Where applicable, additional documentation to make data interpretable is available in the folders with the datasets.

All data will be stored for at least 15 years, and can therefore also be reused in this time. There is no embargo on the accessibility of the data.

Privacy

The privacy of participants is protected by the use of unique individual subject codes. The key files that link these subject codes to identifiable information are stored in separate (password-protected) folders, only accessible by a small selection of involved researchers.

PhD Portfolio

Name PhD student	Lot Snijders Blok
Departments	Department of Human Genetics, Radboud University Medical Center Language & Genetics Department, Max Planck Institute for Psycholinguistics Nijmegen, the Netherlands
Graduate school	Donders Graduate School for Cognitive Neuroscience
Promotors	Prof. dr. H.G. Brunner Prof. dr. S.E. Fisher Prof. dr. T. Kleefstra

Training Activities

A) Courses and workshops	**Year(s)**	**ECTS**
Language in Interaction Summer School	2016	2.5
Introduction to Data Analysis	2016	1.0
Basic Course on Regulations and Organization for Clinical Investigators (BROK)	2017	1.5
The Art of Presenting Science	2017	1.5
Management for PhD students	2017	2.0
Molecular and Cellular Neurobiology (SOW-DGCN46)	2017	6.0
Scientific Integrity Course Radboud University Medical Center	2017	1.0
Writing Retreat Language Interaction consortium (1st edition)	2018	1.5
Education in a Nutshell	2018	1.0
‘Onderzoeker in de klas’	2019	1.5
Writing Retreat Language Interaction consortium (2nd edition)	2019	1.5
B) Seminars & Lectures		
Theme Discussions and Seminars Human Genetics Department Radboudumc	2016-2020	3.0
Colloquia and Seminars Max Planck Institute for Psycholinguistics	2016-2020	3.0
C) (Inter)national Symposia and Congresses		
Oral presentations		
Genetics Retreat – NVHG graduate meeting – Kerkrade	2017	0.75
Annual Meeting of the American Society of Human Genetics (ASHG) - Orlando	2017	1.25
Auris Information Market – Rotterdam	2017	0.25
Joint Meeting Clinical Genetics UK/NL - Utrecht	2018	0.75
Genomics of Rare Disease - Hinxton	2018	1.0
51st European Society of Human Genetics (ESHG) Conference – Milan	2018	1.25
Donders Discussions (Organizer/Moderator of session ‘from DNA to disease’)	2018	1.5
Genomics of Rare Disease - Hinxton	2019	1.0
52nd European Society of Human Genetics (ESHG) Conference - Gothenburg	2019	1.25
NVK ‘Inherited and Congenital Disease’ young researcher day - Maastricht	2019	0.5

Poster presentations		
Taalstaal Conference – Nieuwegein	2020	0.25
European Human Genetics Conference – Virtual meeting	2020	1.0

D) Other

Peer review of scientific publications	2017-2020	0.5
Oral presentation award, third prize – NVHG Graduate Meeting	2017	
Semifinalist Charles J. Epstein Trainee Award – 67th ASHG Annual Meeting	2017	
Young Investigator Award – European Human Genetics Conference (ESHG)	2019	

Teaching Activities

E) Lecturing

Teacher in different courses (Q1 & Q2) of the Bachelors Medicine & Biomedical Sciences	2016-2020	4.0
Workshop NGS Bioinformatics Analysis and Clinical Curation – UGM Yogyakarta	2019	1.5

F) Supervision

Supervision of research internships of three Master students (Medicine/Biology)	2017-2019	3.0
	Total ECTS	**46.75**

List of publications

Snijders Blok, L., N. Corsten-Janssen, D. R. FitzPatrick, C. Romano, M. Fichera, G. A. Vitello, M. H. Willemsen, J. Schoots, R. Pfundt, C. M. van Ravenswaaij-Arts, L. Hoefsloot and T. Kleefstra (2014). **"Definition of 5q11.2 microdeletion syndrome reveals overlap with CHARGE syndrome and 22q11 deletion syndrome phenotypes."** Am J Med Genet A 164A(11): 2843-2848.

Snijders Blok, L., E. Madsen, J. Juusola, C. Gilissen, D. Baralle, M. R. Reijnders, H. Venselaar, C. Helsmoortel, M. T. Cho, A. Hoischen, L. E. Vissers, T. S. Koemans, W. Wissink-Lindhout, E. E. Eichler, C. Romano, H. Van Esch, C. Stumpel, M. Vreeburg, E. Smeets, K. Oberndorff, B. W. van Bon, M. Shaw, J. Gecz, E. Haan, M. Bienek, C. Jensen, B. L. Loeys, A. Van Dijck, A. M. Innes, H. Racher, S. Vermeer, N. Di Donato, A. Rump, K. Tatton-Brown, M. J. Parker, A. Henderson, S. A. Lynch, A. Fryer, A. Ross, P. Vasudevan, U. Kini, R. Newbury-Ecob, K. Chandler, A. Male, D. D. D. Study, S. Dijkstra, J. Schieving, J. Giltay, K. L. van Gassen, J. Schuurs-Hoeijmakers, P. L. Tan, I. Pediaditakis, S. A. Haas, K. Retterer, P. Reed, K. G. Monaghan, E. Haverfield, M. Natowicz, A. Myers, M. C. Kruer, Q. Stein, K. A. Strauss, K. W. Brigatti, K. Keating, B. K. Burton, K. H. Kim, J. Charrow, J. Norman, A. Foster-Barber, A. D. Kline, A. Kimball, E. Zackai, M. Harr, J. Fox, J. McLaughlin, K. Lindstrom, K. M. Haude, K. van Roozendaal, H. Brunner, W. K. Chung, R. F. Kooy, R. Pfundt, V. Kalscheuer, S. G. Mehta, N. Katsanis and T. Kleefstra (2015). **"Mutations in DDX3X Are a Common Cause of Unexplained Intellectual Disability with Gender-Specific Effects on Wnt Signaling."** Am J Hum Genet 97(2): 343-352.

Mulhern, M. S., C. Stumpel, N. Stong, H. G. Brunner, L. Bier, N. Lippa, J. Riviello, R. P. W. Rouhl, M. Kempers, R. Pfundt, A. P. A. Stegmann, M. K. Kukolich, A. Telegrafi, A. Lehman, C. study, E. Lopez-Rangel, N. Houcinat, M. Barth, N. den Hollander, M. J. V. Hoffer, S. Weckhuysen, E.-R. E. S. M. A. E. w. g. Euro, J. Roovers, T. Djemie, D. Barca, B. Ceulemans, D. Craiu, J. R. Lemke, C. Korff, H. C. Mefford, C. T. Meyers, Z. Siegler, S. M. Hiatt, G. M. Cooper, E. M. Bebin, **L. Snijders Blok**, H. E. Veenstra-Knol, E. H. Baugh, E. H. Brilstra, C. M. L. Volker-Touw, E. van Binsbergen, A. Revah-Politi, E. Pereira, D. McBrian, M. Pacault, B. Isidor, C. Le Caignec, B. Gilbert-Dussardier, F. Bilan, E. L. Heinzen, D. B. Goldstein, S. J. C. Stevens and T. T. Sands (2018). **"NBEA: Developmental disease gene with early generalized epilepsy phenotypes."** Ann Neurol 84(5): 788-795.

Snijders Blok, L., S. M. Hiatt, K. M. Bowling, J. W. Prokop, K. L. Engel, J. N. Cochran, E. M. Bebin, E. K. Bijlsma, C. A. L. Ruivenkamp, P. Terhal, M. E. H. Simon, R. Smith, J. A. Hurst, D. D. D. study, H. McLaughlin, R. Person, A. Crunk, M. F. Wangler, H. Streff, J. D. Symonds, S. M. Zuberi, K. S. Elliott, V. R. Sanders, A. Masunga, R. J. Hopkin, H. A. Dubbs, X. R. Ortiz-Gonzalez, R. Pfundt, H. G. Brunner, S. E. Fisher, T. Kleefstra and G. M. Cooper (2018). **"De novo mutations in MED13, a component of the Mediator complex, are associated with a novel neurodevelopmental disorder."** Hum Genet 137(5): 375-388.

Snijders Blok, L., J. Rousseau, J. Twist, S. Ehresmann, M. Takaku, H. Venselaar, L. H. Rodan, C. B. Nowak, J. Douglas, K. J. Swoboda, M. A. Steeves, I. Sahai, C. Stumpel, A. P. A. Stegmann, P. Wheeler, M. Willing, E. Fiala, A. Kochhar, W. T. Gibson, A. S. A. Cohen, R. Agbahovbe, A. M. Innes, P. Y. B. Au, J. Rankin, I. J. Anderson, S. A. Skinner, R. J. Louie, H. E. Warren, A. Afenjar, B. Keren, C. Nava, J. Buratti, A. Isapof, D. Rodriguez, R. Lewandowski, J. Propst, T. van Essen, M. Choi, S. Lee, J. H. Chae, S. Price, R. E. Schnur, G. Douglas, I. M. Wentzensen, C. Zweier, A. Reis, M. G. Bialer, C. Moore, M. Koopmans, E. H. Brilstra, G. R. Monroe, K. L. I. van Gassen, E. van Binsbergen, R. Newbury-Ecob, L. Bownass, I. Bader, J. A. Mayr, S. B. Wortmann, K. J. Jakielski, E. A. Strand, K. Kloth, T. Bierhals, D. D. D. study, J. D. Roberts, R. M. Petrovich, S. Machida, H. Kurumizaka, S. Lelieveld, R. Pfundt, S. Jansen, P. Deriziotis, L. Faivre, J. Thevenon, M. Assoum, L. Shriberg, T. Kleefstra, H. G. Brunner, P. A. Wade, S. E. Fisher and P. M. Campeau (2018). **"CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language."** Nat Commun **9**(1): 4619.

Snijders Blok, L., T. Kleefstra, H. Venselaar, S. Maas, H. Y. Kroes, A. M. A. Lachmeijer, K. L. I. van Gassen, H. V. Firth, S. Tomkins, S. Bodek, D. D. D. Study, K. Ounap, M. H. Wojcik, C. Cunniff, K. Bergstrom, Z. Powis, S. Tang, D. N. Shinde, C. Au, A. D. Iglesias, K. Izumi, J. Leonard, A. Abou Tayoun, S. W. Baker, M. Tartaglia, M. Niceta, M. L. Dentici, N. Okamoto, N. Miyake, N. Matsumoto, A. Vitobello, L. Faivre, C. Philippe, C. Gilissen, L. Wiel, R. Pfundt, P. Deriziotis, H. G. Brunner and S. E. Fisher (2019). **"De Novo Variants Disturbing the Transactivation Capacity of POU3F3 Cause a Characteristic Neurodevelopmental Disorder."** Am J Hum Genet 105(2): 403-412.

Zarate, Y. A., K. A. Bosanko, A. R. Caffrey, J. A. Bernstein, D. M. Martin, M. S. Williams, E. M. Berry-Kravis, P. R. Mark, M. A. Manning, V. Bhambhani, M. Vargas, A. H. Seeley, J. I. Estrada-Veras, M. F. van Dooren, M. Schwab, A. Vanderver, D. Melis, A. Alsadah, L. Sadler, H. Van Esch, B. Callewaert, A. Oostra, J. Maclean, M. L. Dentici, V. Orlando, M. Lipson, S. P. Sparagana, T. J. Maarup, S. I. Alsters, A. Brautbar, E. Kovitch, S. Naidu, M. Lees, D. M. Smith, L. Turner, V. Raggio, L. Spangenberg, S. Garcia-Minaur, E. R. Roeder, R. O. Littlejohn, D. Grange, J. Pfotenhauer, M. C. Jones, M. Balasubramanian, A. Martinez-Monseny, **L. Snijders Blok**, R. Gavrilova and J. L. Fish (2019). **"Mutation update for the SATB2 gene."** Hum Mutat 40(8): 1013-1029.

Connaughton, D. M., R. Dai, D. J. Owen, J. Marquez, N. Mann, A. L. Graham-Paquin, M. Nakayama, E. Coyaud, E. M. N. Laurent, J. R. St-Germain, **L. Snijders Blok**, A. Vino, V. Klambt, K. Deutsch, C. W. Wu, C. M. Kolvenbach, F. Kause, I. Ottlewski, R. Schneider, T. M. Kitzler, A. J. Majmundar, F. Buerger, A. C. Onuchic-Whitford, M. Youying, A. Kolb, D. Salmanullah, E. Chen, A. T. van der Ven, J. Rao, H. Ityel, S. Seltzsam, J. M. Rieke, J. Chen, A. Vivante, D. Y. Hwang, S. Kohl, G. C. Dworschak, T. Hermle, M. Alders, T. Bartolomaeus, S. B. Bauer, M. A. Baum, E. H. Brilstra, T. D. Challman, J. Zyskind, C. E. Costin, K. M. Dipple, F. A. Duijkers, M. Ferguson, D. R. Fitzpatrick, R. Fick, I. A. Glass, P. J. Hulick, A. D. Kline, I. Krey, S. Kumar, W.

Lu, E. J. Marco, I. M. Wentzensen, H. C. Mefford, K. Platzer, I. S. Povolotskaya, J. M. Savatt, N. V. Shcherbakova, P. Senguttuvan, A. E. Squire, D. R. Stein, I. Thiffault, V. Y. Voinova, M. J. G. Somers, M. A. Ferguson, A. Z. Traum, G. H. Daouk, A. Daga, N. M. Rodig, P. A. Terhal, E. van Binsbergen, L. A. Eid, V. Tasic, H. M. Rasouly, T. Y. Lim, D. F. Ahram, A. G. Gharavi, H. M. Reutter, H. L. Rehm, D. G. MacArthur, M. Lek, K. M. Laricchia, R. P. Lifton, H. Xu, S. M. Mane, S. Sanna-Cherchi, A. D. Sharrocks, B. Raught, S. E. Fisher, M. Bouchard, M. K. Khokha, S. Shril and F. Hildebrandt (2020). **"Mutations of the Transcriptional Corepressor ZMYM2 Cause Syndromic Urinary Tract Malformations."** Am J Hum Genet 107(4): 727-742.

Drivas, T. G., D. Li, D. Nair, J. T. Alaimo, M. Alders, J. Altmuller, T. S. Barakat, E. M. Bebin, N. L. Bertsch, P. R. Blackburn, A. Blesson, A. M. Bouman, K. Brockmann, P. Brunelle, M. Burmeister, G. M. Cooper, J. Denecke, A. Dieux-Coeslier, H. Dubbs, A. Ferrer, D. Gal, L. E. Bartik, L. B. Gunderson, L. Hasadsri, M. Jain, C. Karimov, B. Keena, E. W. Klee, K. Kloth, B. Lace, M. Macchiaiolo, J. L. Marcadier, J. M. Milunsky, M. P. Napier, X. R. Ortiz-Gonzalez, P. N. Pichurin, J. Pinner, Z. Powis, C. Prasad, F. C. Radio, K. J. Rasmussen, D. L. Renaud, E. T. Rush, C. Saunders, D. Selcen, A. R. Seman, D. N. Shinde, E. D. Smith, T. Smol, **L. Snijders Blok**, J. M. Stoler, S. Tang, M. Tartaglia, M. L. Thompson, J. M. van de Kamp, J. Wang, D. Weise, K. Weiss, R. Woitschach, B. Wollnik, H. Yan, E. H. Zackai, G. Zampino, P. Campeau and E. Bhoj (2020). **"A second cohort of CHD3 patients expands the molecular mechanisms known to cause Snijders Blok-Campeau syndrome."** Eur J Hum Genet 28(10): 1422-1431.

Johnson-Kerner, B., **L. Snijders Blok**, L. Suit, J. Thomas, T. Kleefstra and E. H. Sherr (2020). **DDX3X-Related Neurodevelopmental Disorder.** GeneReviews((R)). M. P. Adam, H. H. Ardinger, R. A. Pagon et al. Seattle (WA).

Lennox, A. L., M. L. Hoye, R. Jiang, B. L. Johnson-Kerner, L. A. Suit, S. Venkataramanan, C. J. Sheehan, F. C. Alsina, B. Fregeau, K. A. Aldinger, C. Moey, I. Lobach, A. Afenjar, D. Babovic-Vuksanovic, S. Bezieau, P. R. Blackburn, J. Bunt, L. Burglen, P. M. Campeau, P. Charles, B. H. Y. Chung, B. Cogne, C. Curry, M. D. D'Agostino, N. Di Donato, L. Faivre, D. Heron, A. M. Innes, B. Isidor, B. Keren, A. Kimball, E. W. Klee, P. Kuentz, S. Kury, D. Martin-Coignard, G. Mirzaa, C. Mignot, N. Miyake, N. Matsumoto, A. Fujita, C. Nava, M. Nizon, D. Rodriguez, **L. Snijders Blok**, C. Thauvin-Robinet, J. Thevenon, M. Vincent, A. Ziegler, W. Dobyns, L. J. Richards, A. J. Barkovich, S. N. Floor, D. L. Silver and E. H. Sherr (2020). **"Pathogenic DDX3X Mutations Impair RNA Metabolism and Neurogenesis during Fetal Cortical Development."** Neuron 106(3): 404-420 e408.

Snijders Blok, L., A. Vino, J. den Hoed, H. R. Underhill, D. Monteil, H. Li, F. J. Reynoso Santos, W. K. Chung, M. D. Amaral, R. E. Schnur, T. Santiago-Sim, Y. Si, H. G. Brunner, T. Kleefstra and S. E. Fisher (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(3): 534-542.

den Hoed, J., E. de Boer, N. Voisin, A. J. M. Dingemans, N. Guex, L. Wiel, C. Nellaker, S. M. Amudhavalli, S. Banka, F. S. Bena, B. Ben-Zeev, V. R. Bonagura, A. L. Bruel, T. Brunet, H. G. Brunner, H. B. Chew, J. Chrast, L. Cimbalistiene, H. Coon, D. D. D. Study, E. C. Delot, F. Demurger, A. S. Denomme-Pichon, C. Depienne, D. Donnai, D. A. Dyment, O. Elpeleg, L. Faivre, C. Gilissen, L. Granger, B. Haber, Y. Hachiya, Y. H. Abedi, J. Hanebeck, J. Y. Hehir-Kwa, B. Horist, T. Itai, A. Jackson, R. Jewell, K. L. Jones, S. Joss, H. Kashii, M. Kato, A. A. Kattentidt-Mouravieva, F. Kok, U. Kotzaeridou, V. Krishnamurthy, V. Kucinskas, A. Kuechler, A. Lavillaureix, P. Liu, L. Manwaring, N. Matsumoto, B. Mazel, K. McWalter, V. Meiner, M. A. Mikati, S. Miyatake, T. Mizuguchi, L. H. Moey, S. Mohammed, H. Mor-Shaked, H. Mountford, R. Newbury-Ecob, S. Odent, L. Orec, M. Osmond, T. B. Palculict, M. Parker, A. K. Petersen, R. Pfundt, E. Preiksaitiene, K. Radtke, E. Ranza, J. A. Rosenfeld, T. Santiago-Sim, C. Schwager, M. Sinnema, **L. Snijders Blok**, R. C. Spillmann, A. P. A. Stegmann, I. Thiffault, L. Tran, A. Vaknin-Dembinsky, J. H. Vedovato-Dos-Santos, S. A. Schrier Vergano, E. Vilain, A. Vitobello, M. Wagner, A. Waheeb, M. Willing, B. Zuccarelli, U. Kini, D. F. Newbury, T. Kleefstra, A. Reymond, S. E. Fisher and L. Vissers (2021). **"Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction."** Am J Hum Genet 108(2): 346-356.

Snijders Blok, L., Y. M. Goosen, L. van Haaften, K. van Hulst, S. E. Fisher, H. G. Brunner, J. I. M. Egger and T. Kleefstra (2021). **"Speech-language profiles in the context of cognitive and adaptive functioning in SATB2-associated syndrome."** Genes Brain Behav: e12761.

Donders Graduate School for Cognitive Neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit:
http://www.ru.nl/donders/graduate-school/phd/