INFORMATICS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Integrated Sequence Analysis Pipeline Provides One-Stop Solution for Identifying Disease-Causing Mutations

Hao Hu,[1]* Thomas F. Wienker,[1] Luciana Musante,[1] Vera M. Kalscheuer,[1] Kimia Kahrizi,[2] Hossein Najmabadi,[2] and H. Hilger Ropers[1]

[1]*Max-Planck Institute for Molecular Genetics, Berlin, Germany ;* [2]*Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran*

Communicated by Graham R. Taylor

Received 13 January 2014; accepted revised manuscript 28 August 2014.
Published online 13 September 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22695

ABSTRACT: Next-generation sequencing has greatly accelerated the search for disease-causing defects, but even for experts the data analysis can be a major challenge. To facilitate the data processing in a clinical setting, we have developed a novel medical resequencing analysis pipeline (MERAP). MERAP assesses the quality of sequencing, and has optimized capacity for calling variants, including single-nucleotide variants, insertions and deletions, copy-number variation, and other structural variants. MERAP identifies polymorphic and known causal variants by filtering against public domain databases, and flags nonsynonymous and splice-site changes. MERAP uses a logistic model to estimate the causal likelihood of a given missense variant. MERAP considers the relevant information such as phenotype and interaction with known disease-causing genes. MERAP compares favorably with GATK, one of the widely used tools, because of its higher sensitivity for detecting indels, its easy installation, and its economical use of computational resources. Upon testing more than 1,200 individuals with mutations in known and novel disease genes, MERAP proved highly reliable, as illustrated here for five families with disease-causing variants. We believe that the clinical implementation of MERAP will expedite the diagnostic process of many disease-causing defects.

Hum Mutat 35:1427–1435, 2014.

KEY WORDS: next-generation sequencing; intellectual disability; indels; logistic model

Introduction

Since the introduction of next-generation sequencing (NGS) techniques, there have been many articles on single gene defects underlying known or novel genetic diseases [Bamshad et al., 2011; Najmabadi et al., 2011; Rauch et al., 2012; Yu et al., 2013]. Such gene defects are not confined to familial cases; de novo and inherited mutations explain a significant portion of common and rare diseases, particularly in outbred Western populations [Vissers et al., 2010]. Given the growing awareness that the total number of monogenic disorders is much larger than previously assumed and the rapidly increasing number of known gene defects [Boycott et al., 2013], there is a growing need to implement NGS techniques in the clinic as diagnostic tools [Hu et al., 2009; Mamanova et al., 2010; Hu et al., 2011; Bainbridge et al., 2013]. While diagnostic tests combining targeted exon enrichment and NGS to rule out mutations in several dozen to more than 1,000 disease genes are gaining ground, the clinical implementation of whole-exome or whole-genome sequencing still lags behind. This is a major problem for the molecular diagnosis of known diseases and for the identification of hitherto unknown ones, as large-scale medical genome sequencing and central storage and comparison of the clinical and sequence information are crucially important for identifying disease-causing mutations in a sea of functionally neutral sequence variants. Whole-exome or genome sequencing may generate unsolicited genetic information, which is considered as an issue even though the same is true for a wide variety of established diagnostic procedures. More serious problems hampering the introduction of medical genome sequencing relate to the performance of available NGS techniques, the complexity and low concordance of variant-calling pipelines, and the identification and prioritization of potentially disease-causing sequence variants [O'Rawe et al., 2013].

Here, we describe a novel, easy-to-install and to-use medical resequencing analysis pipeline (MERAP; https://sourceforge.net/projects/merap) that compares favorably with several of the existing ones in various respects and is designed as a one-stop solution for clinical applications and for research laboratories with limited bioinformatics support. It consists of eight sequentially operating software modules that interact with several databases to generate a prioritized shortlist of plausible disease-causing mutations, as shown in Figure 1.

Medical Resequencing Analysis Pipeline

MERAP is a comprehensive solution for handling the NGS data, evaluating the performance, and for detecting, filtering, annotating, and prioritizing sequence variants. The pipeline is compatible with a variety of enrichment platforms used in combination with Illumina sequencers, and it requires only minor adjustment for use in combination with other NGS systems.

MERAP uses SOAP2 as the default mapping tool for aligning the raw sequencing reads to the human reference genome, but is also compatible with other tools after minor adjustment [Li et al., 2009b]. Its alignment report provides elementary information such as total output (Gb), percentage of bases reaching the Q20 quality threshold

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Hao Hu, Max-Planck Institute for Molecular Genetics, Berlin 14195, Germany. E-mail: hu@molgen.mpg.de

Contract grant sponsors: Max Planck Society; European Commission Framework Program 7 (FP7) project GENCODYS (grant number 241995).

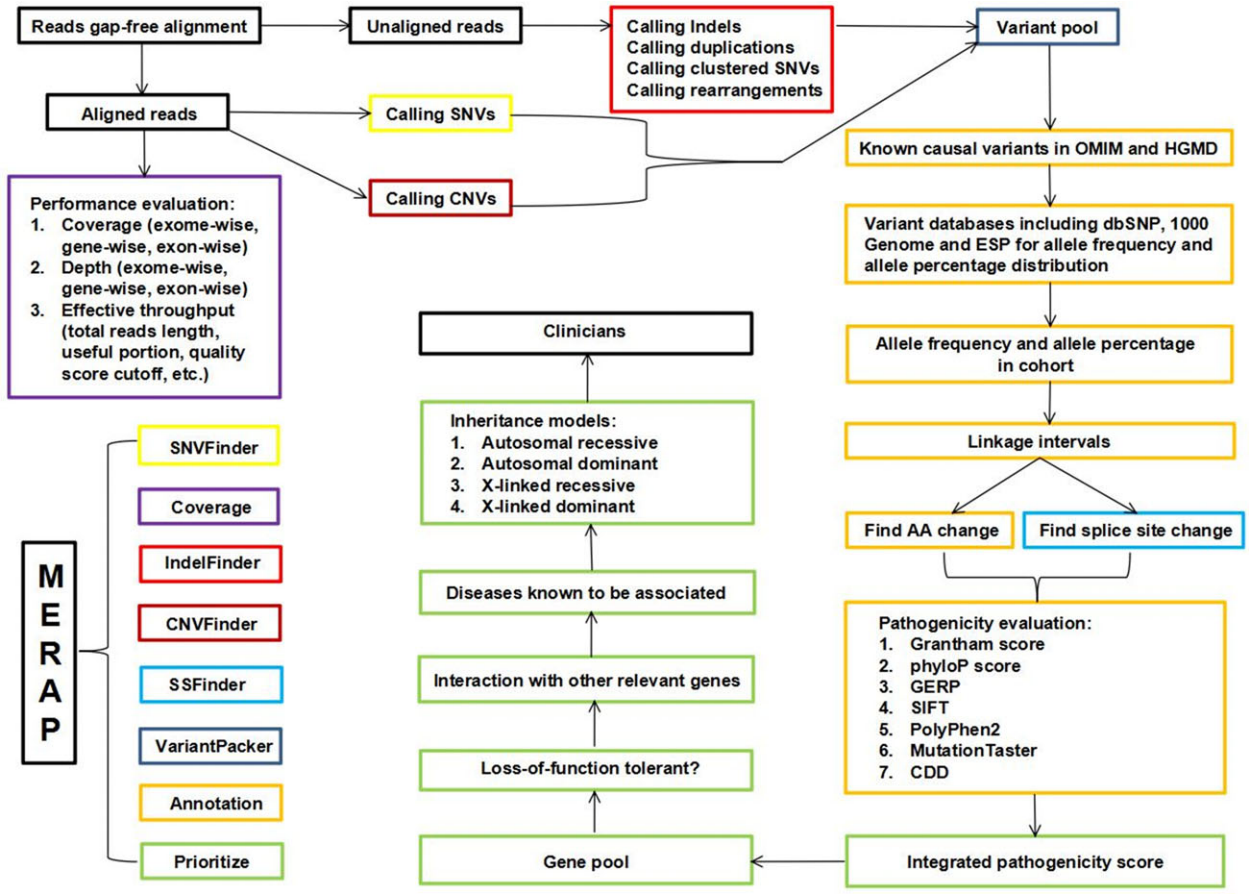

Figure 1. Flowchart and components of MERAP. MERAP is composed of eight software modules, namely, SNVFinder for calling SNVs and generating the base-wise sequencing depth, IndelFinder for calling indels and other structural variants, CNVFinder for calling CNVs, SSFinder for defining cryptic splice sites, Coverage for evaluating sequencing performance by coverage information, VariantPacker for packing variants from a cohort into a general variant list for subsequent annotation and prioritization, Annotation for filtering the variants and annotating by gene models, and Prioritize for prioritizing candidate mutations by considering their pathogenicity and other relevant information. Boxes connected by arrows show the direction of the analytical procedure. The color of each box is matched with the software responsible for each specific step.

[Cock et al., 2010], percentage of aligned reads, and so on. The report also lists the coverage and sequencing depth for each gene, its exons, coding sequences, and its transcript. In order to evaluate the evenness of the sequencing depth, a correlation between coverage and sequencing depth is provided for the entire target region (Supp. Table S1, also see the Supporting Information for algorithm details of coverage).

Functionality

Calling sequence variants involves three different software modules, starting with SNVFinder which focuses on single-nucleotide variants (SNVs). Characteristically, heterozygous SNVs are called if 30%–70% of the nonredundant reads carry identical sequence variants, whereas the other reads correspond to the wild-type sequence. The quality score of each SNV reflects the sequencer-generated Phred-like score (i.e., the sequencing chip image quality) as well as the positional mappability (the probability of a read being aligned to the cognate region) [Ewing and Green, 1998]. The corrected quality score is rescaled to the customary range of 0–40. Artificial SNVs due to nearby indels are predicted and removed. SNVFinder can automatically accommodate itself to different read length (allowed in the same batch), quality format (both Illumina format and Sanger format are allowed), and allows for both redundant and nonredundant reads (see Supporting Information for algorithm details of SNVFinder.pl; see Supp. Figs. S1 and S2).

Copy-number variant (CNV) calling is based on the detection of exons with a significantly elevated or reduced number of overlapping sequence reads. Normalization is performed to compensate for different exon lengths, GC-content, and possible gaps due to the enrichment of target sequences, using preferably more than 20 samples in a batch. After log2 transformation, a value of zero corresponds to the normal diploid state, with significantly higher or lower values for duplications and deletions, respectively (see Supp. Fig. S3). In practice, even after stringent normalization, the variable sequencing depth renders this method unsuitable for detecting indels that involve only a single exon. Since most indels detected correspond to known CNVs, MERAP filters out known and repeatedly occurring CNVs and retains only indels that encompass two or more adjacent exons. The price for this is minimal, as most of the smaller CNVs can be identified by another MERAP module (see below). Further details concerning the CNVFinder algorithm are provided in the Supporting Information.

IndelFinder, another subroutine of MERAP, detects indels and other complex genomic rearrangements. As shown in Figure 2, the

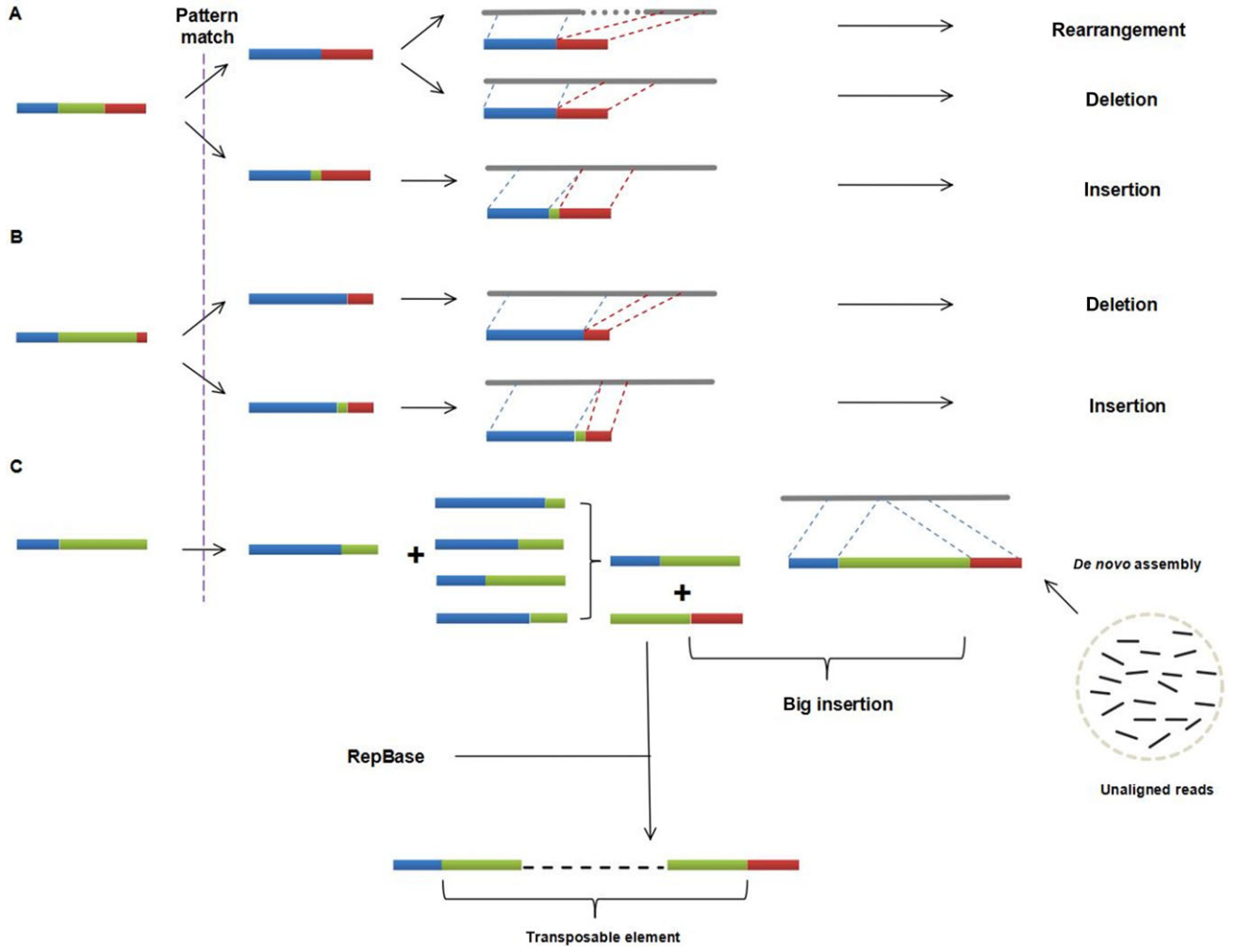

Figure 2. Algorithm of IndelFinder. **A**, **B**, and **C** show three examples of successful split-read mapping. **A**: Both substrings (blue and red) are uniquely mapped to the reference genome; after pattern match and extension, the breakpoints are defined, which are used to define the events of insertion, deletion, and rearrangement. **B**: Only one substring (blue) is uniquely mapped to the reference genome; the other substring is shortened (red) and mapped to the 10-kb genomic vicinity of the mapped substring (blue). After pattern matching and extension, the breakpoints are established and the molecular details of the rearrangement are inferred. **C**: Only one substring (blue) is uniquely mapped to the reference genome; after pattern match and extension, the breakpoint is defined; reads with the same breakpoint from the same direction are collected to find the one with longest unidentified sequence. If the same breakpoint is identified from the other direction, and if the de novo assembled sequence aligned to the reference genome with a large insertion shows the same breakpoints, a large insertion is thus defined. The sequences flanking large insertions are aligned to RepBase to trace their origins.

underlying algorithm combines a split-read strategy with a de novo assembly approach. Two (e.g., 36 bp) substrings are extracted from the termini of each unaligned read and mapped to human genome. If both substrings can be uniquely mapped, they will be used as anchors for subsequent analyses. If only one substring can be unambiguously mapped, the other substring will be shortened to 10 bp and used to search in silico for a complementary sequence within of a radius of 10 kb around the mapped substring. Then, the length of the two anchors is increased and that of the insert separating them is reduced by pattern match and extension through comparison with the reference genome. Subsequently, the positions of the two anchors and the size of the insert are used to infer the presence of insertions, deletions, and other rearrangements including complex substitutions.

To identify insertions that cannot be accommodated on a single sequence read, IndelFinder starts out from reads harboring only one anchor to recover flanking sequences from both sides of the insertion. With these sequences as probes, de novo assembly is then performed on the pool of unaligned reads, resulting in a series of contigs that are then aligned to the human reference genome by BLAT (http://genome.ucsc.edu/FAQ/FAQblat.html). Medium to large size insertions are also mapped to the repeat sequence database (RepBase, http://www.girinst.org/repbase/) to characterize their origin (see Supporting Information for algorithm details of IndelFinder). In practice, deletions as large as 25,859 bp have been detected in this way; full-length sequences could be inferred for insertions as large as 400 bp; and complex rearrangements such as tandem duplications ranging in size from 40 to 1,178 bp have been identified.

Typically, MERAP detects 0.82 variant per 1 kb transcript regions, among which the percentages of SNV, small Indel, other variants such as large insertion, tandem duplication, and CNV, are 85.4, 12.2, and 2.4, respectively. All variants called by MERAP are presented in a format complying with the nomenclature for the description of sequence variants (Human Genome Variation Society [HGVS], http://www.hgvs.org/mutnomen/). Figure 3 shows nine examples of variants identified by MERAP. The specificity of

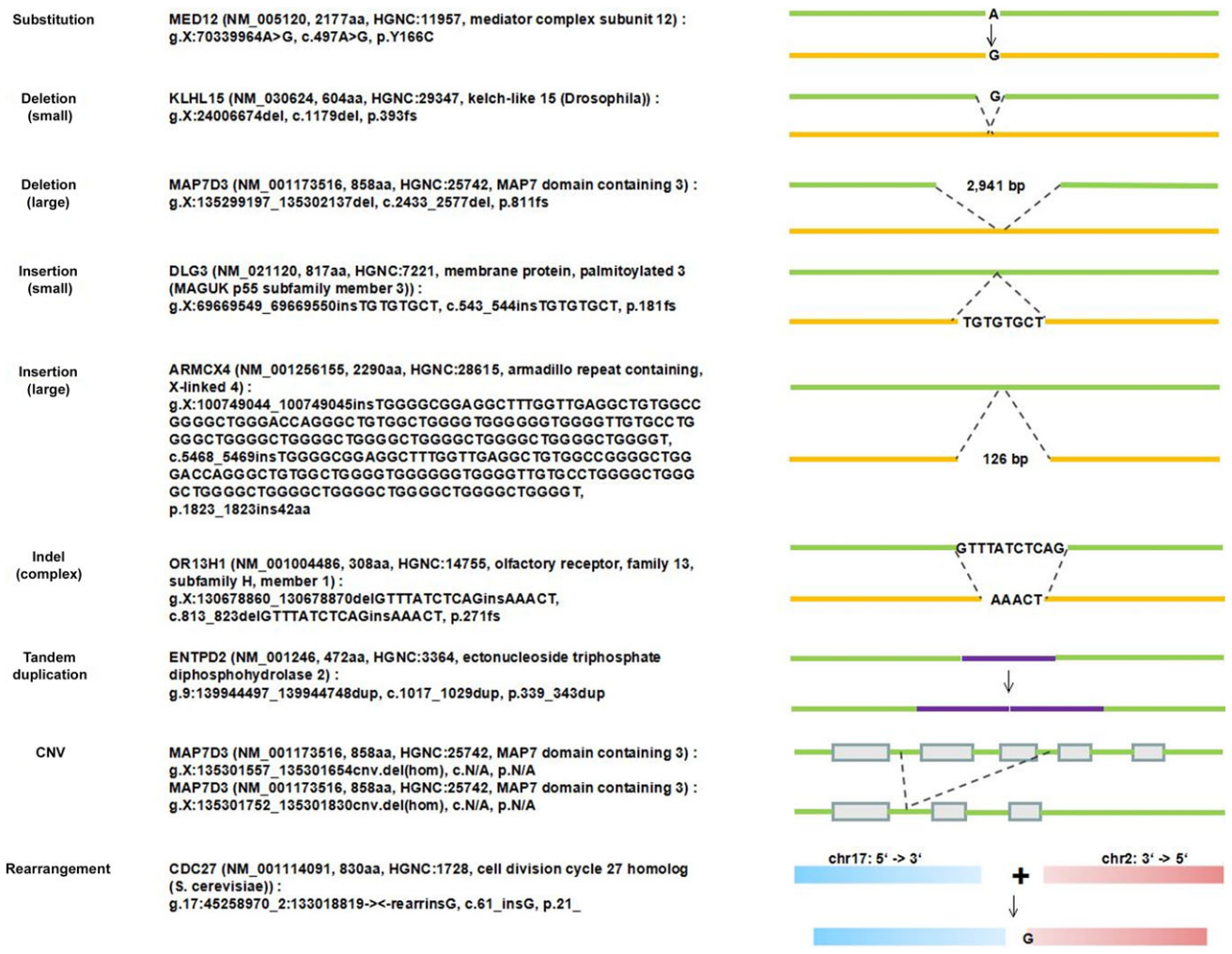

Figure 3. Examples of variants identified by MERAP. The identified variant types include base substitutions, small deletions, large deletions, small insertions, large insertions, complex indels (with both insertion and deletion), tandem duplications, CNVs, and other rearrangement.

MERAP variant calling has been evaluated by conventional Sanger sequencing. In practice, >99.9% of the variants called by MERAP could be confirmed, and false-positive results were only obtained for variants supported by less than five reads. We evaluated the sensitivity of MERAP variant calling by comparison with known variants from Affymetrix genotyping array and found the concordance to be as high as 99.5%. The high reliability of MERAP variant calling is also reflected by the consistent appearance of highly recurrent (>10% in cohort) variants in public databases including dbSNP, 1000 Genome, and ESP 6500 exomes [Sherry et al., 2001; Altshuler et al., 2012; Fu et al., 2013]. All variants identified in a cohort are packed by VariantPacker into a general variant list for the subsequent annotation and prioritization.

All variants identified by MERAP are first filtered through comparison with more than 126,000 disease-associated variants extracted from Human Gene Mutation Database (HGMD) and Online Mendelian Inheritance in Man (OMIM) to identify known disease-causing mutations. In order to filter out neutral variants, MERAP uses up to one million entries from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), 1000 Genome (http://www.1000genomes.org/), and NHLBI Exome Sequencing Project (ESP, http://evs.gs.washington.edu/EVS/) as databases. MERAP does not only consider matches in the databases, but also their frequency in the population and in the cohort studied. For example, homozygosity for mutations causing recessive disorders should not occur in healthy controls, and the frequency of heterozygotes should not exceed one or few percent; and mutations causing severe dominant disorders should not be observed at all. More than 90% of the variants can be ruled out in this way.

For translating DNA variants into amino acid changes and for assessing the functional relevance of nonsynonymous or splice-site changes, MERAP uses RefSeq genes (http://www.ncbi.nlm.nih.gov/refseq/) as reference, because of their nonredundancy and consistency. Nonsynonymous changes are described in terms of gene ID, base change, protein change, genomic coordinate, transcript coordinate, protein coordinate, protein length, affiliated with gene description from the Human Gene Nomenclature Committee (HGNC, http://www.genenames.org/) (Fig. 3; Table 1). MERAP identifies changes destroying conventional splice sites (2 bp flanking sequences of exons) or introducing novel splice sites (see Supporting Information for algorithm details of SSFinder).

To assess the pathogenicity of missense mutations, MERAP generates a single score that is based on Logit modeling and integrates the results of seven different algorithms, including the Grantham score (codon replacement conservation based on chemical dissimilarity) [Grantham, 1974], phyloP (base replacement

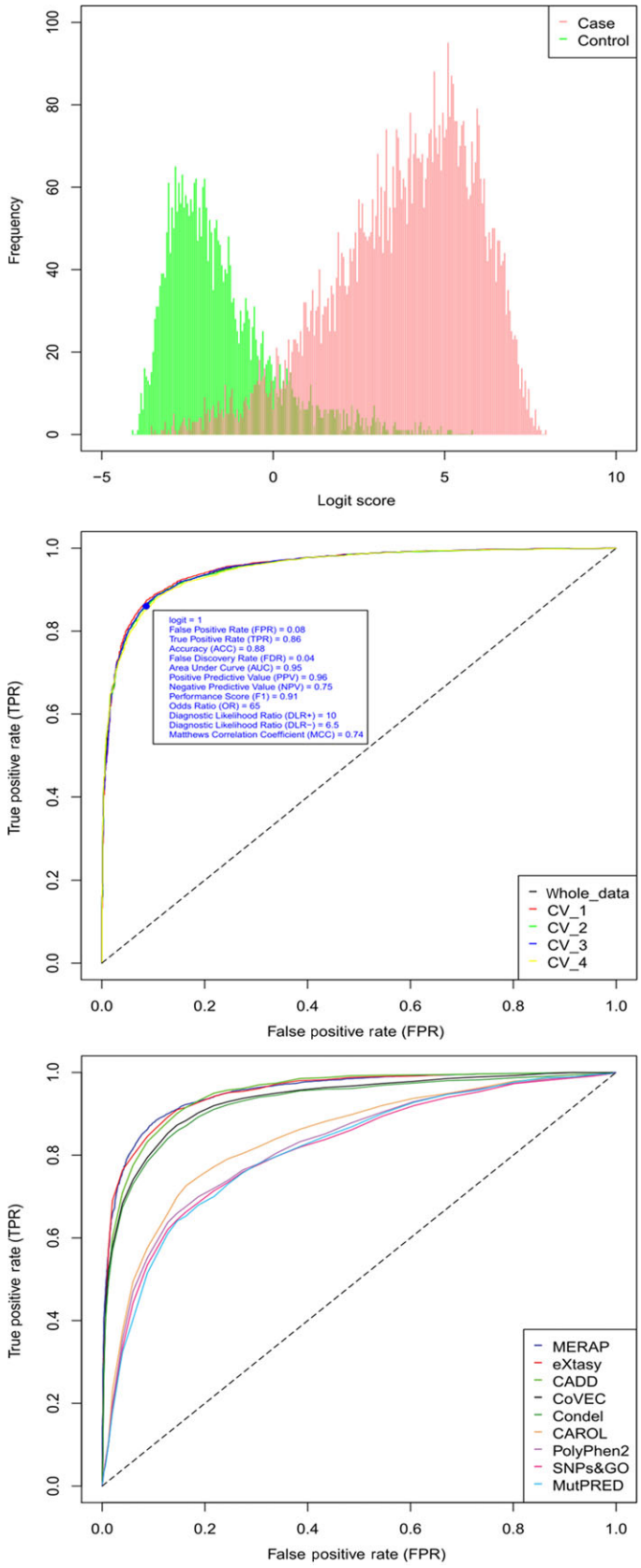

Figure 4. Logit modeling and ROC curves. **A**: Shows the Logit scores distribution for case set and control set. **B**: Shows the ROC curves for the whole data and four test data (see Supporting Information for details). **C**: Shows the ROC curves of MERAP and the other eight software, and the AUCs of MERAP, eXtasy, CADD, CoVEC, Condel, CAROL, PolyPhen2, SNPs&GO, and MutPRED are 0.954, 0.952, 0.949, 0.933, 0.929, 0.867, 0.841, 0.834, and 0.831, respectively.

Table 1. The Specification of MERAP Result

Detail	Example	Note
Family and patient ID	M999_1274	Family ID M999, patient ID 1274
Gene name, RefSeq ID, protein length, HGNC ID, gene description, variant coordinate in terms of genome, cDNA, and protein	HPD(NM_002150,393aa,HGNC:5147,4-hydroxyphenylpyruvate dioxygenase): g.12:122277904G>C,c.1005C>G,p.I335M	Multiple isoforms of RefSeq genes are shown together separated by "|"
Number of nonredundant reads supporting the variant	76	
Allele percentage of the variant	0.98	
Phred-like quality score (0–40)	40	
Subjects in the cohort, incidence of the variant, homozygote frequency, heterozygote frequency	371|3|1|2	In a cohort of 371 subjects, the variant is observed three times with once as homozygote and twice as heterozygote
Definition of linkage interval, LOD score, length of interval, the variant location in the interval	Homozygous|3.1|—*—	In a homozygous interval with LOD score 3.1 and length 5.5 Mb, the variant is located at the center of the interval. "-" stands for length unit 0.5 Mb and "*" stands for the location of variant and also a length unit 0.5 Mb
HGMD match of the variant	DM;HPD;Tyrosinaemia 3;Hum Genet:v.106,p.654,y.2000	There is a match of the variant in HGMD, with the classification of DM (disease causing); the host gene name is HPD; the associated phenotype is Tyrosinaemia 3; it is reported in Hum Genet (volumn:106, page:654, year:2000)
OMIM match of the variant	TYROSINEMIA,TYPE III	
dbSNP match of the variant	rs137852868	
1000 Genome match of the variant	HOM_REF:HOM_VAR:HET = 1,090:0:2; AF = 0.0009; AMR = 0.01; ASN = 0; AFR = 0; EUR = 0	The incidences of the variant in homozygous wild type, homozygous variant, and heterozygous variant are 1,090, 0, and 2, respectively. The allele frequency of the variant in population is 0.0009, with ethnic-specific frequencies of 0.01, 0, 0, and 0 in American population, Asian population, African population, and European population, respectively
ESP match of the variant	N/A	
Grantham score for AA change	10	
phyloP score for base conservation	2,547	
GERP score for base conservation	3.25	
SIFT prediction and score	Damaging (0.000000)	
PolyPhen2 prediction and score	Probably damaging (0.978)	
MutationTaster prediction and score	Disease causing (0.999999)	
CDD match and score	cl14632:Glo_EDI_BRP_like superfamily (0.498727735368957)|	Multiple matches are shown together separated by "|"
Integrated pathogenicity score by Logit modeling	4,172	Pass the cutoff 3.57 where false discovery rate is <0.01
If a loss-of-function tolerant gene	N	N means negative, P means positive
Interaction partner of the gene	HPD<->IKBKG	The gene interacts physically with IKBKG
Known diseases caused by the gene	Hawkinsinuria;Tyrosinaemia 3	
Proposed inheritance model for the disease	Recessive	

There are 24 fields shown in MERAP results, including the sample ID, the variant information, the reads number supporting the variant, the allele percentage of the variant, the quality of the call, the variant frequency in cohort, the variant position in terms of the linkage intervals, the HGMD match, the OMIM match, the dbSNP match, the 1000 Genome match, the ESP6500 match, the Grantham score, the phyloP score, the GERP score, the SIFT score, the PolyPhen2 score, the MutationTaster score, the CDD score, the Logit score, the LOF score, the interaction partner, the known disease, and the inheritance model.

conservation) [Pollard et al., 2010], GERP (base replacement conservation) [Davydov et al., 2010], SIFT (amino acid residue conservation) [Ng and Henikoff, 2003], PolyPhen2 (various physical and comparative parameters) [Adzhubei et al., 2010], MutationTaster (integrated evaluation of base and amino acid conservation) [Schwarz et al., 2010], and the Conserved Domains Database (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml; details of the scoring algorithm are provided in the Supporting Information). With empirical false discovery rate cutoffs, the integrated Logit scores allow for dichotomized pathogenicity predictions even if SIFT, PolyPhen2, and MutationTaster predictions do not coincide, as is often the case. To calibrate this integrated score, we collected 7,703 disease-causing (DM) missense mutations from HGMD, all with heterozygote frequencies of <1% in the 1000 Genome or ESP6500 databases, as well as 3,520 neutral missense variants that occur in homozygous form in the 1000 Genome and ESP6500 databases, with allele frequencies exceeding 10%. The seven aforementioned algorithms generate the numerical predictions for the case and control sets, which are then converted into ranks. A Logit model was fitted to optimize the discrimination between case and control sets. Its receiver operating characteristic (ROC) curve shows high performance of the model (the area under the curve [AUC] = 0.95) and its robustness by resampling without replacement (Fig. 4). Compared with other pathogenicity scores combining multiple predictors, advantages of the MERAP Logit score model include the availability of stringent training sets and the use of ranks instead of numerical values [Calabrese et al., 2009; Li et al., 2009a; Adzhubei et al., 2010; Gonzalez-Perez and Lopez-Bigas, 2011; Lopes et al., 2012; Frousios et al., 2013; Sifrim et al., 2013; Kircher et al., 2014]. As shown by the ROC curves, MERAP performs as well as eXtasy and CADD, outperforms CoVEC and Condel, and is significantly superior to CAROL, PolyPhen2, SNPs&GO, and MutPRED (for further information about the annotation algorithm and details of Logit modeling and ROC evaluation, see Supporting Information).

Even after filtering and annotation, we are typically left with several to dozens of candidate mutations for each case. When cohorts of patients with the same ethnicity are studied, information on the proportion of homozygotes and heterozygotes carrying a particular variant is also useful for detecting population-specific

polymorphisms (see Table 1). MERAP tracks candidate genes reported to harbor homozygous loss-of-function (LOF) variants in healthy individuals, which applies to >1% of the human genes. If more than three independent truncating variants are observed in >10 of the exomes listed in the 1000 Genome and ESP6500 databases, the relevant gene is flagged as LOF tolerant.

To facilitate the choice between few remaining candidate genes, MERAP also provides a list of ~3,000 known disease genes extracted from OMIM (http://www.ncbi.nlm.nih.gov/omim), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), and HGMD (http://www.hgmd.org/), as well as ~8,000 associated disorders and their symptoms. For novel candidate genes without prior link to disease, MERAP offers information on their interaction with known disease genes, based on data from Biogrid (http://thebiogrid.org/) and IntAct (http://www.ebi.ac.uk/intact/). The rationale is that genes implicated in clinically similar disorders tend to cluster in gene or protein interaction networks. Further details about the prioritization algorithm are provided in Supporting Information. A practical example illustrating the entire MERAP workflow and outcome is given in Table 1.

Performance

MERAP is faster than other sequence analysis pipelines, and its computational requirements are modest. For example, on a server with 48 CPUs sharing 248Gb RAM, MERAP managed to process 221 WES samples (>60× mean coverage) and to complete the analysis within 4 days (92 hr). For smaller targets, the turnover was even much higher (see Supp. Table S1). For example, it took MERAP only 17 hr to screen 195 samples for mutations in 520 genes implicated in severe related to recessive childhood disorders birth defects [Bell et al., 2011], and no more than 11 hr to analyze 60 individuals for mutations in a total of 1,222 disease genes, including most genes that have been linked to intellectual disability so far. For WES with >60× mean coverage, MERAP necessitates only 60 Gb RAM. MERAP has been designed for laboratories with limited infrastructure and resources for DNA sequence analysis. With its optimized default parameters and automatic links to all relevant databases, it can be downloaded and installed within 1 day (https://sourceforge.net/projects/merap), with minimal manual intervention.

To compare its performance with that of the widely used Genome Analysis Tool Kit (GATK) [DePristo et al., 2011], we have analyzed WES results from 22 unrelated individuals with both pipelines, focusing on SNVs and indels. For MERAP and GATK, average processing times were 5 and 18 hr, respectively. 98% of the SNVs called by MERAP were also detected by GATK, whereas 75% of the indels identified by MERAP and 88% of GATK-identified indels were shared (see Supp. Fig. S4). It is of note that that one-third of the indels exclusively identified by MERAP correspond to known dbSNP entries, suggesting that at least some of the indels unique to MERAP in this study are real. For further details concerning the relative performance of MERAP and GATK, see the Supporting Information. The comparison with another SNV caller, VarScan, was conducted based upon the same dataset [Koboldt et al., 2009]. About 98.5% of SNVs identified by MERAP were also called by VarScan, whereas 99% of SNVs called by VarScan were identified by MERAP. The comparison with another indel caller, Pindel, was implemented based on the same dataset [Ye et al., 2009]. More than 95% of the indels called by Pindel could be found by MERAP but only 66% of the indels called by MERAP can be identified by Pindel. Up to 39% of the indels uniquely called by MERAP were variants listed in dbSNP. MERAP was also compared with CNVnator for identifying unique CNVs from the aforementioned dataset [Abyzov et al., 2011]. MERAP called on average 12 multiple-exon-spanning CNVs for each sample, 90% of which could also be identified by CNVnator. Seven CNVs on average were exclusively identified by CNVnator, most of which were deletions, but when we randomly picked three such CNVs and did quantitative PCR, none of them could be confirmed.

Discussion

Since 2008, we have used NGS techniques to process and analyze more than 1,200 human DNA samples (Supp. Table S3). In the course of this work, which enabled us to identify many dozen novel disease genes, we have developed several novel sequence alignment and analysis tools. Improved versions of these tools have now been linked to form MERAP, an integrated sequence analysis pipeline that provides a one-stop solution for identifying disease-causing mutations, for example, in a clinical setting. MERAP was instrumental in the reanalysis of numerous previously unsolved cases, and in the vast majority of cases that we had considered as solved, it confirmed our previous conclusions (Supp. Table S3). Moreover, it turned out to be a major asset for the investigation of more than 800 novel families with autosomal-recessive ID, autosomal-dominant ID, and X-linked ID [Hu et al., 2011; Kahrizi et al., 2011; Najmabadi et al., 2011; Pak et al., 2011; Rafiq et al., 2011; Ropers et al., 2011; Schraders et al., 2011; Strobl-Wildemann et al., 2011; Huang et al., 2012; Zanni et al., 2012; Bainbridge et al., 2013; Dreha-Kulaczewski et al., 2013; Hirata et al., 2013; Lesca et al., 2013; Puttmann et al., 2013; Belet et al., 2014; de Brouwer et al., 2014; Larti et al., 2014; Masurel-Paulet et al., 2014; Philips et al., 2014], as illustrated below for five of these that were found to carry defects in known or novel plausible candidate genes (see Supporting Information; Supp. Figus. S5–S9; Supp. Table S4).

MERAP can be retrieved from https://sourceforge.net/projects/merap. It includes eight software programs, the relevant databases as well as manuals. NGS data mentioned in this paper can be retrieved from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) under the accession numbers SRA036250. The aforementioned list of 1,222 genes for severe recessive childhood disorders and intellectual disability is available upon request. Most inherited disorders are genetically heterogeneous, and for some, the heterogeneity is overwhelming. For example, more than 700 different genes have been implicated in severe forms of intellectual disability (ID) (extracted from the latest version of HGMD [version 4, 2013] and OMIM [February, 2014]), and there is reason to believe that the total number of ID genes will run into the thousands, most of which have not been identified yet [Ropers, 2010]. This strongly argues for introducing medical genome sequencing as standard diagnostic procedure, which in turn necessitates sequence analysis tools that are adapted to the needs of diagnostic laboratories with limited bioinformatics experience and infrastructure. MERAP has been developed to meet the rapidly growing demand for fast, easy to install, and user-friendly pipelines enabling the search for clinically relevant mutations in whole exomes or whole genomes, but also in defined segments of the human genome that can be isolated by targeted enrichment. Given its superior sensitivity and specificity, it should be equally useful for large, more experienced groups.

While in the current version of MERAP, settings have been optimized for its present task, future versions will enable users to fine-tune them at will to meet specific demands. Also, there will

be room for implementing additional algorithms for pathogenicity prediction and for other adjustments by individual users. Finally, future versions of MERAP will be adapted to handle much larger data sets, for example, through further improvement of the processing efficiency and by implementing data compression.

Acknowledgment

We are very grateful to Dr. Peter N. Robinson from Universitätsklinikum Charité, Berlin for his concrete contribution to this work.

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974–984.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Bainbridge MN, Hu H, Muzny DM, Musante L, Lupski JR, Graham BH, Chen W, Gripp KW, Jenny K, Wienker TF, Yang Y, Sutton VR, Gibbs RA, Ropers HH. 2013. De novo truncating mutations in ASXL3 are associated with a novel clinical phenotype with similarities to Bohring-Opitz syndrome. Genome Med 5:11.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745–755.

Belet S, Fieremans N, Yuan X, Van Esch H, Verbeeck J, Ye Z, Cheng L, Brodsky BR, Hu H, Kalscheuer VM, Brodsky RA, Froyen G. 2014. Early frameshift mutation in PIGA identified in a large XLID family without neonatal lethality. Hum Mutat 35:350–355.

Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, et al. 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3:65ra4.

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14:681–691.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025.

de Brouwer AP, Nabuurs SB, Verhaart IE, Oudakker AR, Hordijk R, Yntema HG, Hordijk-Hos JM, Voesenek K, de Vries BB, van Essen T, Chen W, Hu H, et al. 2014. A 3-base pair deletion, c.9711_9713del, in DMD results in intellectual disability without muscular dystrophy. Eur J Hum Genet 22:480–485.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

Dreha-Kulaczewski S, Kalscheuer V, Tzschach A, Hu H, Helms G, Brockmann K, Weddige A, Dechent P, Schluter G, Kratzner R, Ropers HH, Gartner J, Zirn B. 2013. A novel SLC6A8 mutation in a large family with X-linked intellectual disability: clinical and proton magnetic resonance spectroscopy data of both hemizygous males and heterozygous females. JIMD Rep 13:91–99.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194.

Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. 2013. Predicting the functional consequences of non-synonymous DNA sequence variants—evaluation of bioinformatics tools and development of a consensus strategy. Genomics 102:223–238.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ; NHLBI Exome Sequencing Project, Akey JM. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Gonzalez-Perez A, Lopez-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88:440–449.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Hirata H, Nanda I, van Riesen A, McMichael G, Hu H, Hambrock M, Papon MA, Fischer U, Marouillat S, Ding C, Alirol S, Bienek M, et al. 2013. ZC4H2 mutations are associated with arthrogryposis multiplex congenita and intellectual disability through impairment of central and peripheral synaptic plasticity. Am J Hum Genet 92:681–695.

Hu H, Eggers K, Chen W, Garshasbi M, Motazacker MM, Wrogemann K, Kahrizi K, Tzschach A, Hosseini M, Bahman I, Hucho T, Muhlenhoff M, et al. 2011. ST3GAL3 mutations impair the development of higher cognitive functions. Am J Hum Genet 89:407–414.

Hu H, Wrogemann K, Kalscheuer V, Tzschach A, Richard H, Haas SA, Menzel C, Bienek M, Froyen G, Raynaud M, Bokhoven HV, Chelly J, Ropers H, Chen W. 2009. Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing. Hugo J 3:41–49.

Huang L, Jolly LA, Willis-Owen S, Gardner A, Kumar R, Douglas E, Shoubridge C, Wieczorek D, Tzschach A, Cohen M, Hackett A, Field M, et al. 2012. A noncoding, regulatory mutation implicates HCFC1 in nonsyndromic intellectual disability. Am J Hum Genet 91:694–702.

Kahrizi K, Hu CH, Garshasbi M, Abedini SS, Ghadami S, Kariminejad R, Ullmann R, Chen W, Ropers HH, Kuss AW, Najmabadi H, Tzschach A. 2011. Next generation sequencing in a family with autosomal recessive Kahrizi syndrome (OMIM 612713) reveals a homozygous frameshift mutation in SRD5A3. Eur J Hum Genet 19:115–117.

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.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–2285.

Larti F, Kahrizi K, Musante L, Hu H, Papari E, Fattahi Z, Bazazzadegan N, Liu Z, Banan M, Garshasbi M, Wienker TF, Ropers HH, Galjart N, Najmabadi H. 2014. A defect in the CLIP1 gene (CLIP-170) can cause autosomal recessive intellectual disability. Eur J Hum Genet.[Epub ahead of print]

Lesca G, Moizard MP, Bussy G, Boggio D, Hu H, Haas SA, Ropers HH, Kalscheuer VM, Des Portes V, Labalme A, Sanlaville D, Edery P, Raynaud M, Lespinasse J. 2013. Clinical and neurocognitive characterization of a family with a novel MED12 gene frameshift mutation. Am J Med Genet A 161A:3063–3071.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009a. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009b. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967.

Lopes MC, Joyce C, Ritchie GR, John SL, Cunningham F, Asimit J, Zeggini E. 2012. A combined functional annotation score for non-synonymous variants. Hum Hered 73:47–51.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118.

Masurel-Paulet A, Kalscheuer VM, Lebrun N, Hu H, Levy F, Thauvin-Robinet C, Darmency-Stamboul V, El Chehadeh S, Thevenon J, Chancenotte S, Ruffier-Bourdet M, Bonnet M, et al. 2014. Expanding the clinical phenotype of patients with a ZDHHC9 mutation. Am J Med Genet A 164A:789–795.

Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, Hosseini M, Behjati F, Haas S, Jamali P, Zecha A, Mohseni M, et al. 2011. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 478:57–63.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon GJ. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5:28.

Pak C, Garshasbi M, Kahrizi K, Gross C, Apponi LH, Noto JJ, Kelly SM, Leung SW, Tzschach A, Behjati F, Abedini SS, Mohseni M, et al. 2011. Mutation of the conserved polyadenosine RNA binding protein, ZC3H14/dNab2, impairs neural function in Drosophila and humans. Proc Natl Acad Sci USA 108:12390–12395.

Philips AK, Siren A, Avela K, Somer M, Peippo M, Ahvenainen M, Doagu F, Arvio M, Kaariainen H, Van Esch H, Froyen G, Haas SA, Hu H, Kalscheuer VM, Jarvela I. 2014. X-exome sequencing in Finnish families with intellectual disability—four novel mutations and two novel syndromic phenotypes. Orphanet J Rare Dis 9:49.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121.

Puttmann L, Stehr H, Garshasbi M, Hu H, Kahrizi K, Lipkowitz B, Jamali P, Tzschach A, Najmabadi H, Ropers HH, Musante L, Kuss AW. 2013. A novel ALDH5A1 mutation is associated with succinic semialdehyde dehydrogenase deficiency and severe intellectual disability in an Iranian family. Am J Med Genet A 161A:1915–1922.

Rafiq MA, Kuss AW, Puettmann L, Noor A, Ramiah A, Ali G, Hu H, Kerio NA, Xiang Y, Garshasbi M, Khan MA, Ishak GE, et al. 2011. Mutations in the alpha 1,2-mannosidase gene, MAN1B1, cause autosomal-recessive intellectual disability. Am J Hum Genet 89:176–182.

Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, et al. 2012. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380:1674–1682.

Ropers F, Derivery E, Hu H, Garshasbi M, Karbasiyan M, Herold M, Nurnberg G, Ullmann R, Gautreau A, Sperling K, Varon R, Rajab A. 2011. Identification of a novel candidate gene for non-syndromic autosomal recessive intellectual disability: the WASH complex member SWIP. Hum Mol Genet 20:2585–3590.

Ropers HH. 2010. Genetics of early onset cognitive impairment. Annu Rev Genomics Hum Genet 11:161–187.

Schraders M, Haas SA, Weegerink NJ, Oostrik J, Hu H, Hoefsloot LH, Kannan S, Huygen PL, Pennings RJ, Admiraal RJ, Kalscheuer VM, Kunst HP, Kremer H. 2011. Next-generation sequencing identifies mutations of SMPX, which encodes the small muscle protein, X-linked, as a cause of progressive hearing impairment. Am J Hum Genet 88:628–634.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7:575–576.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y. 2013. eXtasy: variant prioritization by genomic data fusion. Nat Methods 10:1083–1084.

Strobl-Wildemann G, Kalscheuer VM, Hu H, Wrogemann K, Ropers HH, Tzschach A. 2011. Novel GDI1 mutation in a large family with non-syndromic X-linked intellectual disability. Am J Med Genet A 155A: 3067–3070.

Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA. 2010. A de novo paradigm for mental retardation. Nat Genet 42:1109–1112.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871.

Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, Schmitz-Abe K, Harmin DA, Adli M, Malik AN, D'Gama AM, Lim ET, et al. 2013. Using whole-exome sequencing to identify inherited causes of autism. Neuron 77:259–273.

Zanni G, Cali T, Kalscheuer VM, Ottolini D, Barresi S, Lebrun N, Montecchi-Palazzi L, Hu H, Chelly J, Bertini E, Brini M, Carafoli E. 2012. Mutation of plasma membrane Ca2 +ATPase isoform 3 in a family with X-linked congenital cerebellar ataxia impairs Ca2 +homeostasis. Proc Natl Acad Sci USA 109: 14514–14519.