A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease

Damian Smedley,[1,2,15] Max Schubach,[3,15] Julius O.B. Jacobsen,[4,15] Sebastian Köhler,[3] Tomasz Zemojtel,[3,5] Malte Spielmann,[3,6] Marten Jäger,[3,7] Harry Hochheiser,[8] Nicole L. Washington,[9] Julie A. McMurry,[10] Melissa A. Haendel,[10] Christopher J. Mungall,[9] Suzanna E. Lewis,[9] Tudor Groza,[11,12] Giorgio Valentini,[13] and Peter N. Robinson[3,6,7,14,16,*]

The interpretation of non-coding variants still constitutes a major challenge in the application of whole-genome sequencing in Mendelian disease, especially for single-nucleotide and other small non-coding variants. Here we present Genomiser, an analysis framework that is able not only to score the relevance of variation in the non-coding genome, but also to associate regulatory variants to specific Mendelian diseases. Genomiser scores variants through either existing methods such as CADD or a bespoke machine learning method and combines these with allele frequency, regulatory sequences, chromosomal topological domains, and phenotypic relevance to discover variants associated to specific Mendelian disorders. Overall, Genomiser is able to identify causal regulatory variants as the top candidate in 77% of simulated whole genomes, allowing effective detection and discovery of regulatory variants in Mendelian disease.

Introduction

Medical genetics is being transformed by next-generation sequencing (NGS) technologies that enable the simultaneous investigation of all relevant disease genes, all protein-coding genes, and even the entire genome.[1,2] Whole-genome sequencing (WGS) can detect a broader range of genetic variation than other sequencing approaches, including not only single-nucleotide variants (SNVs) and insertion or deletions (indels), but also structural variants such as copy-number variants (CNVs) and translocations. Pilot studies have shown that WGS can reveal disease-causing variants missed by other genetic tests.[3] In addition to interrogating more of the non-coding genome, WGS also has better coverage, even in exome regions.[3] Therefore, WGS is best poised to investigate the relevance of nucleotide substitutions and other small non-coding variants (NCVs) in Mendelian disease, but substantial obstacles remain.

We hypothesize that the rarity of reported Mendelian regulatory mutations is related to a long-standing observational bias toward coding sequences in human genetic diagnostic and research projects. Genome-wide association studies (GWASs) have identified more than 10,000 strong associations ($p < 10^{-5}$) between diseases or traits and SNVs, most of which are located in non-coding sequences;[4] however, in Mendelian disease, mutations in non-coding regions represent a tiny minority of all those published to date. In fact, of the more than 100,000 Mendelian-disease-causing variants in ClinVar,[5] the vast majority affect coding sequences or conserved splice sites. Accordingly, a large number of bioinformatics tools have been developed to predict the pathogenicity of sequence variants in these traditional categories.[6] The "regulatory code" that determines whether and how a given genetic variant affects the function of a regulatory element remains poorly understood for most classes of regulatory variation. Thus, given our lack of understanding and tooling, it is not surprising that so far very few disease-causing NCVs less than 25 nucleotides have been identified as causal in Mendelian disease. To address this, we therefore sought to develop an effective approach to detect regulatory variants causative of Mendelian disease.

Recently, several machine learning (ML) methods have been developed to evaluate arbitrary genomic SNVs with respect to their potential to cause disease or affect genetic regulation.[7–11] None of the methods were designed specifically for the detection of NCVs associated with Mendelian

[1]Queen Mary University of London, London E1 4NS, UK; [2]Genomics England Ltd., London EC1M 6BQ, UK; [3]Institute for Medical and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; [4]Skarnes Faculty Group, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK; [5]Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland; [6]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany; [7]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; [8]Department of Biomedical Informatics and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15206, USA; [9]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [10]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, USA; [11]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia; [12]St Vincent's Clinical School, Faculty of Medicine University of New South Wales, Darlinghurst, NSW 2010, Australia; [13]Anacleto Lab Department of Computer Science, University of Milan, Via Comelico, 20135 Milan, Italy; [14]Institute for Bioinformatics, Department of Mathematics and Computer Science, Freie Universität Berlin, Takustrasse, 14195 Berlin, Germany
[15]These authors contributed equally to this work
[16]Present address: The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA
*Correspondence: peter.robinson@jax.org
http://dx.doi.org/10.1016/j.ajhg.2016.07.005.

disease. To address this, we introduce Genomiser, a complete framework for the prioritization of NCVs and discovery of SNVs causative of specific Mendelian diseases. It has been designed primarily for use in two contexts: clinical diagnosis and novel disease gene identification. Our approach combines two major components: (1) a machine learning method for scoring NCVs and (2) an integrative algorithm for ranking NCVs in whole-genome sequence data. The ML method scores each position of the non-coding genome based on predicted pathogenicity in Mendelian diseases. The integrative algorithm factors in multiple inputs: (1) phenotypes, (2) variants in coding regions, (3) variants in non-coding regions, and (4) existing published gene-phenotype associations. We show by cross-validation studies that the ML method outperforms previous, more general-purpose, pathogenicity scoring schemes in the particular task of identifying Mendelian disease-associated variants. Simulations performed for more than 10,000 case subjects were able to recover the correct regulatory variant in first place in 77% of diagnostic genomes.

Material and Methods

Observed Probably Non-deleterious Variant Sites

We identified single-nucleotide sites in the human genome at which the human genome reference sequence differs from the inferred ancestral primate genome based on the Ensembl Enredo-Pecan-Ortheus (EPO) whole-genome alignments of six primate species[12,13] (Ensembl Compara release e71). A file containing the inferred ancestral sequences was downloaded from the 1000 Genomes Project website. We selected all positions of high-confidence alignments that differed from the human reference sequence (hg19). Low-confidence calls are defined in the file as those where the ancestral state is supported by one sequence only. This file was compared with the human (*Homo sapiens*) genome sequence (hg19) via an in-house Java program that cataloged the differences found according to location with respect to genomic annotations.

We excluded nucleotide positions associated with variants present in the most recent 1000 Genomes Project[14] data at a frequency of higher than 5% (meaning that the derived allele in the human genome is present at a frequency of less than 95% so that it is less certain that the allele has been exposed to many generations of natural selection). The file ALL.wgs.phase3_shapeit2_mvncall_integrated_v5a.20130502.sites.vcf was downloaded from the 1000 Genomes Project FTP site on May 30, 2015, and the AF (allele frequency) field was used as the threshold.

All variants were annotated with Jannovar[15] v.0.14 using NCBI Reference Sequence Database[16] (annotation release 105) and only variants of non-coding variant effect are used as final non-deleterious variant sites (negative positions). Table S1 shows the distribution of variants extracted in this way and the variant categories selected for analysis are marked. This yielded a total of 14,755,199 sites; because deleterious variants are depleted by natural selection in fixed or nearly fixed derived alleles, we infer that variation in these sites is unlikely to be associated with Mendelian disease, and we therefore chose to use this set of genomic sites as negative examples for training.

Biocuration of Non-coding Mendelian Disease-Associated Mutations

Comprehensive literature review was performed to identify non-coding variants that are convincingly associated with Mendelian disease. We included only those variations and publications judged to provide plausible evidence of pathogenicity. First, the phenotypic abnormalities of the individual carrying the variant were assessed and a variant was included only if the disease association was regarded as plausible on the basis of evidence such as familial cosegregation or experimental validation, using techniques such as luciferase reporter assays, electrophoretic mobility assay, or telomerase activity assay. In some cases pathogenicity was assigned based on curator judgment or computational predictions; for instance, mutations in RNA genes that affected RNA secondary structure elements such as stem loops were included. To identify articles for biocuration, a number of review articles were consulted on non-coding mutations,[17–23] including 5′ and 3′ untranslated region (UTR) mutations,[24–29] enhancer mutations,[30–32] promoter mutations,[33,34] and mutations affecting microRNA (miRNA) genes or miRNA recognition sites in mRNAs.[35–38] Additionally, locus-specific databases were consulted for selected genes.[39–41] We did not include variants that represent susceptibility loci for common, complex disease (i.e., "GWAS hits" were excluded). Likewise, somatic variants associated with cancer were not included. A total of 453 unique non-coding Mendelian disease-associated variants were identified (Table S6). Mutations were manually mapped to GRCh37, if necessary. Each variant was cataloged according to its sequence variant type (Table 1). The disease associated with the variant was mapped to an OMIM disease identifier, and the affected gene was encoded with an NCBI Entrez Gene identifier.

Genomic Attributes Used for Machine Learning

Every position in the genome was annotated with 26 numeric features. Conservation scores PhastCons and PhyloP[42] for 9 primates, 32 mammals, and 45 vertebrates multi-species alignments were derived from UCSC.[43] GERP++ element scores and the corresponding p values were downloaded from the GERP[44] website on June 6, 2015. CpG and G/C content as well as the observed to expected CpG ratio were downloaded directly from the UCSC table browser[45] on June 6, 2015. The GC content in the human genome (hg19) in a range of ±75 nt for every position was computed (Ns are not counted). Transcription and regulation annotations were downloaded from UCSC.[43] We used the maximum ENCODE H3K27 acetylation level along with the maximum ENCODE H3K4 methylation level and the maximum ENCODE H3K4 trimethylation level. DNase hypersensitive scores were derived from the UCSC ENCODE Regulation DNase Clusters track V3 along with the number of overlapping transcription factor binding sites conserved in the human/mouse/rat alignment. In addition, permissive and robust enhancers were taken from the FANTOM5 project.[46] Population-based features were computed by counting the number of rare (≤0.5% AF) and common (>0.5% AF) 1000 Genomes[14] (release 5a of 05/02/2013) variants in a window of ±500 and using the ratio of rare variants (≤0.5%) and common variants (>0.5%) (zero if common variants are zero). Finally, overlapping Database of Genomic Variants[47] (DGV), dbVar,[48] and ISCA[49] (study IDs nstd37, nstd45, nstd75) CNVs for every position in the human genome were counted for each position. All attributes are listed in Table S2.

Table 1. Mendelian Regulatory Mutations

Category	Example	Count
Enhancer	triphalangeal thumb, type I (*SHH* [MIM: 174500])	42
Promoter	hemophilia B (*F9* [MIM: 306900])	142
5′ UTR		153
Transcription (core promoter)	acute intermittent porphyria (*HMBS* [MIM: 176000])	52/153
uORF	Marie Unna hereditary hypotrichosis (*HR* [MIM: 146550])	37/153
Secondary structure	hyperferritemia cataract syndrome (*FTL* [MIM: 600886])	31/153
Kozak sequence	beta thalassemia (*HBB* [MIM: 613985])	2/153
Unclassified	thrombocytopenia 2 (*ANKRD26* [MIM: 188000])	31/153
3′ UTR		43
Polyadenylation	permanent neonatal diabetes (*INS* [MIM: 606176])	14/43
miRNA binding	autosomal-dominant spastic paraplegia 31 (*REEP1* [MIM: 610250])	5/43
Other	autosomal-dominant myopia 21 (*ZNF644* [MIM: 614167])	24/43
Large non-coding RNA gene	microcephalic osteodysplastic primordial dwarfism, type 1 (*RNU4ATAC* [MIM: 210710])	65
MicroRNA gene	autosomal-dominant deafness 50 (*MIR96* [MIM: 613074])	5
Imprinting control region	Beckwith-Wiedemann syndrome (*H19* [MIM: 130650])	3
Total		453
Total single-nucleotide variants		406

A total of 453 unique, non-coding, regulatory mutations were identified by manual biocuration (Table S6). The pathomechanism of a subset of the 5′ and 3′ UTR mutations was indicated in the original publications and is shown here. 406 of the 453 mutations were single-nucleotide variants that were used for machine learning. One example of a disease caused by each pathomechanistic category is shown together with the affected gene and the OMIM number of the disease.

Regulatory Mendelian Mutation Score

The regulatory Mendelian mutation (ReMM) framework uses ML techniques to train a classifier to predict the potential of an arbitrary position in the non-coding genome to cause a Mendelian disease if mutated. The hand-curated set of Mendelian mutations was used as a positive training set, and non-coding nucleotides that have diverged in humans as compared with the inferred ancestral primate genome sequence were used as a negative training set.

Our experimental setting is characterized by a high imbalance between the available positive and negative training data: there were only 453 regulatory Mendelian mutations, of which 406 were single-nucleotide variants suitable for training, compared with 14,755,199 negative examples. Thus, approximately 36,000 negative examples are available for every positive one. In such extremely unbalanced conditions, classical computational and machine learning methods tend to perform poorly. This is because they learn overwhelmingly from negative examples, which leads to a sensitivity and precision close to zero on new (test) data.[50]

In order to train the ReMM model, we first divided the majority class (probably non-deleterious variant sites) randomly into n = 100 partitions and then we added all the minority instances (non-coding Mendelian mutations) to every partition. We chose 100 partitions because no substantial performance improvements were observed when more partitions were utilized (data not shown). Moreover, in each partition we synthetically oversampled the minority positive class, using the synthetic minority over-sampling technique[51] (SMOTE) with a number of nearest neighbors $k = 5$. With the SMOTE approach we generated synthetic instances two times the cardinality of the positive class. We then randomly undersampled the majority negative class to obtain a three times larger set of negative examples. The resulting dataset was used to train a random forest (RF) classifier[52] (forest size 10; larger forests do not significantly improve the performances; data not shown) that outputs a probability to estimate whether a given position in non-coding genome can cause a Mendelian disease if mutated. The overall process of over- and undersampling and the training of the RF was repeated for all the n partitions. Finally, the probabilities estimated by each RF were averaged and the resulting "consensus" probability of the hyperensemble represents the final ReMM score. Our method was implemented in Java using Weka.[53]

One ReMM score was generated for each position of the non-coding genome, with 0.0 being the least and 1.0 being the highest prediction of deleteriousness. In order to predict the pathogenicity of deletions, the maximum ReMM score of any nucleotide affected by the deletion is used (note that our tests include deletions of up to 24 nt only). For insertions, the maximum ReMM score of the two nucleotides that surround the insertion are used.

Model Testing and Validation

Model performance was tested with a "cytogenetic band-aware" 10-fold cross validation: to ensure that mutations of the same location, gene, or disease do not occur in the training and test set, we partitioned the mutations into the chromosomal bands. Bands with at least one positive mutation were assigned to one of the ten folds so that each fold contains around 40 positives. The remaining bands were randomly assigned to the different folds and negative variants were added to the partition of their associated band. For each round of the cross-validation, the nine folds corresponding to the training set underwent a subdivision

in n = 100 partitions and were over- and undersampled according to the procedures described above in the model training section. The trained ensemble was then tested on the remaining held-out unchanged fold not used for training. In this way, across the ten rounds of the cross-validation procedure, we tested all the genomic positions available in our data (more than 14,000,000 genomic positions). For all other positions in the human genome, we built a global model using the complete negative and positive positions and annotated the remaining 2,845,135,389 unambiguous (i.e., not "N") positions of the human reference genome (release hg19).

The ReMM score was compared to the non-coding variant scores CADD[7] v.1.3, GWAVA[8] v.1.0, FATHMM-MKL,[54] Eigen,[11] and DeepSEA.[10] CADD and Eigen scores are extracted from the provided precomputed genome-wide file. For Eigen all variants on allosomes were removed, because Eigen is available only on autosomes. Position scores of GWAVA, FATHMM-MKL (commit d4af576240fb872179805fb113e892597248441d), and DeepSEA were computed using the source code provided by the authors.

Regulatory Filtering and Genomiser Application

Genomiser was implemented by extending the existing Exomiser codebase.[55,56] To allow Genomiser to run in a reasonable time frame (~4–10 min) and with a minimal memory (~4–10 GB), we had to reimplement Exomiser to be able to stream variants from a VCF file and run the various filtering and prioritization steps in a user-configured manner rather than the predefined filtering followed by prioritization steps of Exomiser. We also introduced the ability to filter genes by their phenotypic similarity so that as a first step genes associated with diseases that have little or no similarity to the observed phenotypes can be removed along with their associated variants. Note in this step, distal (>20 kb from a gene) variants that reside in predicted enhancers from the FANTOM5 consortium[46] (downloaded on 8/7/15) or Ensembl regulatory feature build[57] (downloaded Ensembl regulatory features dataset from Ensembl Biomart on 8/7/2015) are associated with the most phenotypically similar gene in the topologically associated domain (TAD)[58] containing the enhancer. TADs are defined using Hi-C to identify higher-order chromatin interactions in the three-dimension organization of the genomes and they organize the genome into chromosome neighborhoods within which most enhancer-promoter contact occurs. We replaced the existing behavior of removing all non-coding variants with a new configurable filter that can remove any combination of variant types or none at all in the case of Genomiser. Pathogenicity scoring was extended to use the REMM scores for all non-coding variants. In the case of non-coding insertions, the maximum of the REMM score for the two bases either side of insertion site is used. For non-coding deletions, the highest REMM score for the deleted positions is taken. Finally, we introduced a regulatory feature filter where all variants that lie more than 20 kb from a gene are removed unless they lie in one of the predicted enhancers. The binaries and data for Genomiser are available as part of Exomiser and are free for academic use from Exomiser website. The version used in all results presented here is 7.2.0. The Genomiser_README file describes how to download, install, and run the application to perform Genomiser analysis.

Performance Evaluation

Benchmarking experiments for Genomiser were performed using 10,419 simulated rare disease genomes based on the 453 regulatory Mendelian mutations and 1,092 whole-genomes VCF files from the 1000 Genomes Project[14] (05/02/2013 release). For autosomal-dominant diseases, one heterozygous mutation was added, and for autosomal-recessive diseases, either one homozygous mutation or two heterozygous mutations were added to the 1000 Genomes VCF file. For these experiments, the phenotypic (HPO) annotations for the corresponding disease in OMIM were taken on 8/7/2015 from the annotation files of the HPO team. To measure the ability of Genomiser to detect known disease-gene associations, we repeated the analysis with incomplete (maximum of three HPO annotations), noisy (two random HPO terms added), and imprecise (two of the original HPO annotations replaced by the more general parent terms in the ontology) annotations.

These simulated genomes were run through the default settings of Genomiser. In the first step, genes and associated variants are removed where there is little similarity between observed phenotypes and direct or inferred knowledge from disease and model organism databases. Note in this step, distal (>20 kb from a gene) variants that reside in predicted enhancers from FANTOM5 and Ensembl are associated with the most phenotypically similar gene in the topological domain containing the enhancer rather than simply taking the closest gene. Distal variants that do not reside in a predicted enhancer are removed, followed by the exclusion of any that are common (>1% minor allele frequency [MAF]) in the 1000 Genomes Project, NHLBI Exome Sequencing Project (ESP), and Exome Aggregation Consortium (ExAC) datasets. Finally, the remaining variants are prioritized by a composite score of the minor allele frequency, phenotypic similarity, and pathogenicity (using the ReMM score for non-coding and the existing hiPHIVE method for coding and splice sequences). To assess our performance, we measured how often the seeded regulatory Mendelian variant was ranked first among the full set of the variants of the simulated Mendelian disease genomes.

Results

In this work, we developed a complete framework for the prioritization of non-coding variants in Mendelian disease by combining a bespoke pathogenicity score with phenotype-based measures of gene candidacy. We first developed a pathogenicity score to assess Mendelian non-coding variation. Next we developed methods to integrate the pathogenicity score, candidate regulatory regions, and the phenotypic relevance of the associated genes. Here we describe the development process and present benchmarking of the entire framework.

The Regulatory Mendelian Mutation Score

We reasoned that ML techniques for building a scoring model of non-coding Mendelian variants would perform better with a highly reliable training set, consisting of mutations that had been validated by experimentation or co-segregation studies, or for which other convincing evidence of pathogenicity was available. However, to date such a catalog of validated non-coding variants associated with Mendelian disease has not existed. Therefore, we performed detailed and comprehensive biocuration to identify experimentally or otherwise validated non-coding variants (<25 nucleotides) associated with Mendelian disease

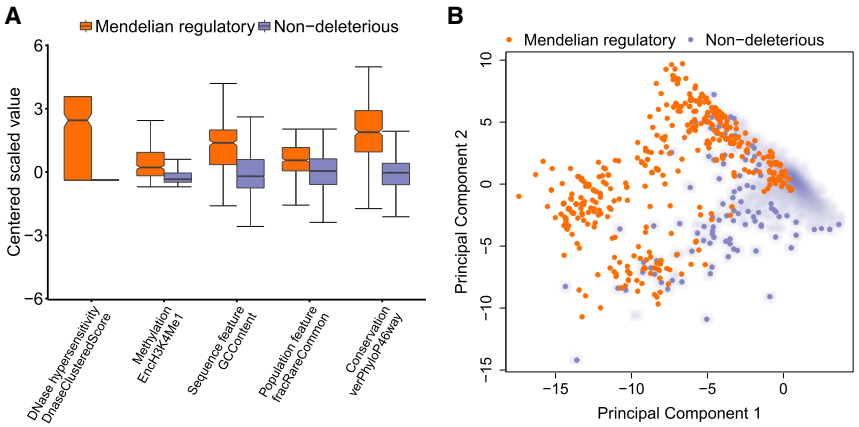

Figure 1. Genomic Attributes of Regulatory Mendelian Mutations

(A) Centered mean and scaled genomic attributes of Mendelian non-coding mutations as compared with the derived non-deleterious positions. Five highly informative attributes of different attribute groups are shown. The information content of single attributes was computed with a univariate logistic regression model (Table S3).

(B) Principal-component analysis plot showing the first two principle components, which make up 32% of the total variability.

from the medical literature. The resulting 453 mutations were located in 5′ and 3′ UTRs, promoters, enhancers, large RNA genes, microRNA genes, and imprinting control regions (ICR) (Table 1, Table S6).

To assess whether the regulatory Mendelian mutations differ from non-deleterious variants, we compared the regulatory mutations based upon a set of genomic characteristics (attributes) representing typical indicators of variant functionality such as GC-content, conservation, histone modifications, DNase I accessibility, and overlap with enhancers and transcription factor binding sites. Additionally, Mendelian candidacy measures were incorporated, such as the ratio of rare to common variation around the position (Table S2). Information content of attributes associated with each variant is computed using an univariate logistic regression model,[59] and results are shown in Table S3. The negative variant set is derived from positions that differ from the inferred primate ancestral genome with an allele frequency > 95% (Table S1). The Mendelian regulatory mutations displayed a number of substantial differences as compared to the neutral variants (Figure 1A). Principal-component analysis (PCA) was performed on the two variant classes with all 26 features. The first two components show a certain separation between our Mendelian regulatory mutations and the negative variants (Figure 1B).

This analysis suggested that the genomic attributes characterizing the nucleotide positions affected by the Mendelian regulatory mutations differ sufficiently from those of non-deleterious variants and could therefore be used to construct a classifier using machine learning techniques. Our experimental setting is characterized by an extreme imbalance between the available positive and negative data (406 Mendelian disease-associated SNVs and 14,755,199 negative data points). As detailed in the Material and Methods section, we developed a hyper-ensemble (ensemble of ensembles) approach in which multiple RFs[52] are used as base learners, together with a combination of over- and undersampling techniques to compensate for the unbalanced sizes of positive and negative training data. In total, an ensemble of 100 RFs are trained in this way to promote balanced and comprehensive coverage of the training data. The probabilities of the 100 RFs are then averaged to compute the Regulatory Mendelian Mutation (ReMM) score (Figure 2A).

We tested the performance of the ReMM score using a 10-fold "cytogenetic band-aware" cross-validation scheme. This scheme was designed to minimize bias due to distinct variants associated with the same disease gene being used for both training and testing. Cytogenetic bands containing at least one disease-associated variant were thereby assigned to one of ten folds for cross validation, with each fold containing a total of approximately 40 disease-associated variants. The remaining bands covering the rest of the genome were randomly assigned to one of the folds. Because our ML involves assessing individual genomic positions, our training set excluded the 47 indels and dinucleotide block mutations. However, in the subsequent analysis of phenotype-driven prioritization (see below) and software implementation, we did include deletions as well as insertions. We evaluated the area under the receiver operating characteristics (ROC) curve and the area under the precision-recall (PR) curve to compare our ReMM score against five other leading scoring methods: CADD,[7] GWAVA,[8] FATHMM,[54] DeepSEA,[10] and Eigen[11] (Figures 2B and 2C). PR and ROC curves show that in the context of the prioritization of the Mendelian mutations, the ReMM score substantially outperforms other state-of-the-art scoring methods. It is worth noting that in the context of extremely unbalanced data, the area under the PR curve is more informative than the area under the ROC,[60] but even the small differences between the ROC curves are in most cases statistically significant according to the Delong test for the comparison of the areas under dependent ROC curves (Table S4, Supplemental Note, Figures S1, S2, and S3).

Phenotype-Driven Prioritization of Non-coding Variants

We and others have previously shown that phenotypic information can effectively boost the prioritization of disease-associated genes.[1,61,62] Regulatory mutations can lead to identical or similar phenotypic abnormalities as

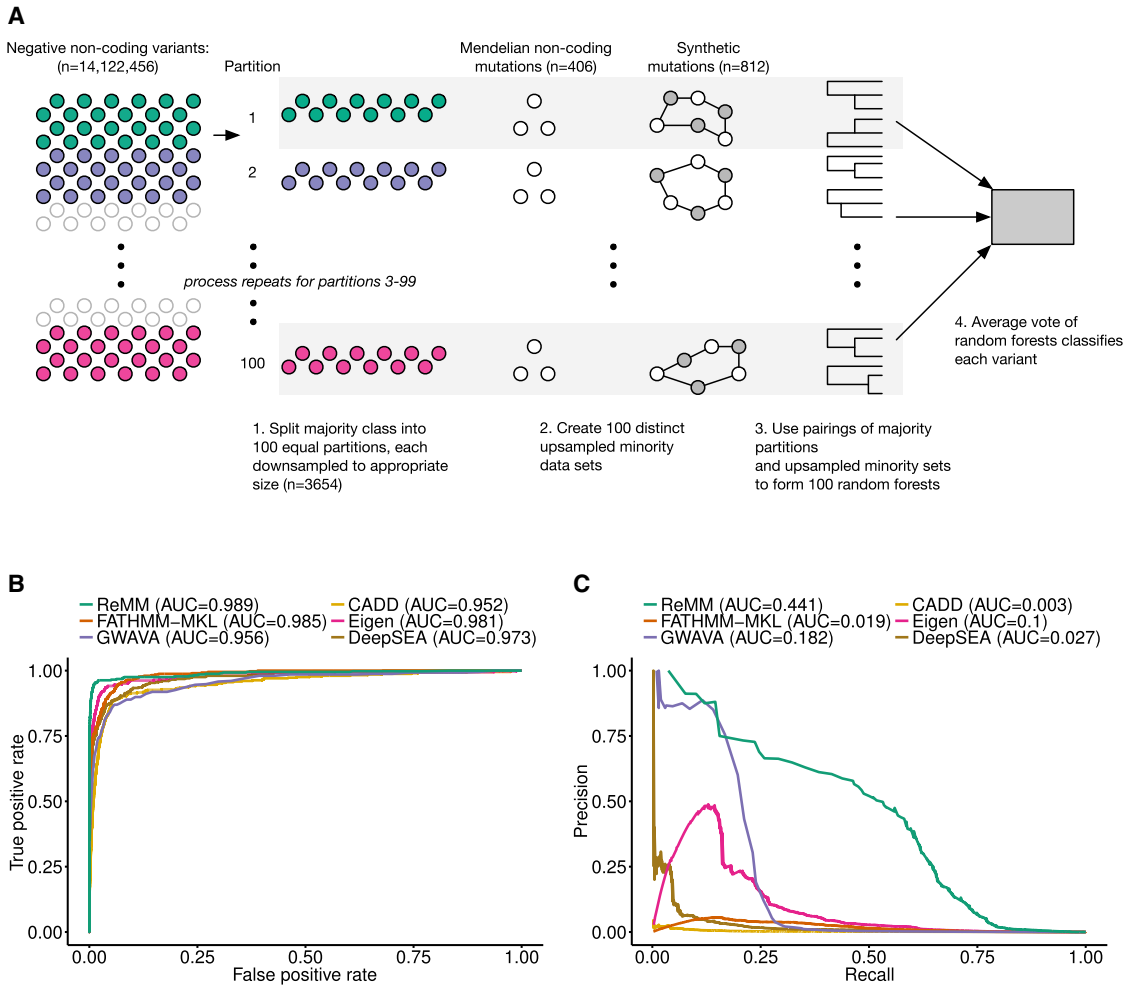

Figure 2. Regulatory Mendelian Mutation-Deleteriousness Score
(A) Summary of the algorithm for deriving the ReMM score.
(B and C) Performance comparison between ReMM and other state-of-the-art genome-wide deleteriousness score.
(B) Receiver operating characteristic curves.
(C) Precision-recall curves.

disruptions of coding sequences in the same gene.[30] We therefore developed a framework, Genomiser, that combines phenotypic, regulatory, and genotypic information for the prioritization of non-coding variants associated with a specific Mendelian disease. Genomiser integrates our existing hiPHIVE algorithm[55,56,63] to exploit phenotypic information from human and model organisms, with the ReMM score to exploit genotypic information, and relevant distal regulatory sequences into the prioritization process. Genomiser is available as an extension to our existing, freely downloadable Exomiser[19] software suite. The input consists of either a single-sample variant call format (VCF) or multi-sample VCF with associated pedigree (PED) file, representing the variations in either an entire human genome or portions thereof. For example, instead of whole genomes, one could use clinical exome data containing at least some part of the regulatory genome such as UTR or proximal promoter sequences. Additionally, the software requires at least one Human Phenotype Ontology[64] (HPO) term that describes the clinical abnormalities of the individual being investigated.

As shown in Figure 3, Genomiser first of all identifies and scores the genes that have the most similar phenotypes to the phenotypic profile under investigation represented using the HPO terms. This scoring makes use of the hiPHIVE algorithm to calculate phenotypic similarity based on either existing phenotypic knowledge from human disease (OMIM, Orphanet), mouse (MGI, IMPC), and zebrafish (ZFIN) sources or using guilt-by-association based on proximity in the STRING-DB protein-protein association network to assign similarity where no phenotypes exist for a gene. Variants associated with the most phenotypically similar genes are then retained if they either (1) lie within a gene including all promoter, UTR, and intronic regions, (2) are within 25 kb up- or down-stream of a gene, or (3) within predicted regulatory features from FANTOM5[46] or the Ensembl regulatory build.[57] Candidate variants are assigned to the most phenotypically similar gene within

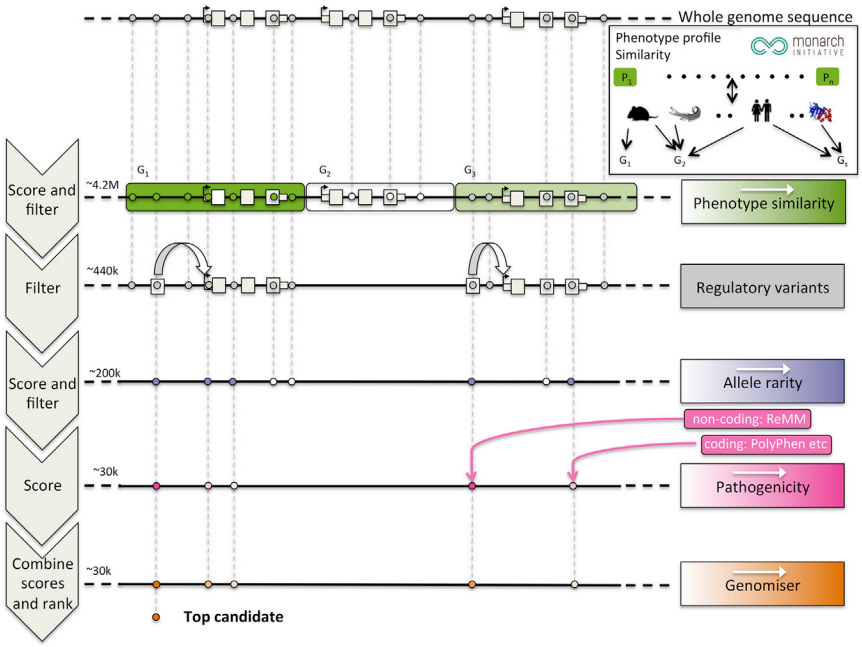

Figure 3. The Genomiser Analysis Framework

Genomiser takes as input a whole-genome variant call format (VCF) file, a list of human phenotype ontology (HPO) terms representing the clinical signs and symptoms observed in the individual being investigated by WGS, and optional user parameters that control the filtering and prioritization steps. See text for details of the prioritization procedure.

the chromosomal topological domain,[31] rather than simply taking the closest gene. We have previously shown that an assessment of regulatory boundaries formed by topological domains can improve the identification of candidate pathogenic copy-number variants.[65]

After this, any common variants (>1% MAF by default) are removed and, optionally, the known/suspected inheritance model used to remove any variants that don't fit the expected pattern. The remaining coding or regulatory variants are then scored according to the allele frequency and predicted deleteriousness (using the ReMM score for non-coding and the existing hiPHIVE method for coding and splice sequences). A composite score based on the phenotypic similarity of the gene to the observed phenotypic profile and the best scoring variant in that gene (or mean of the best two under a compound heterozygous model) is then used to rank the genes and their associated variants.

Genomiser was evaluated by analyzing its capability to recover a known regulatory Mendelian mutation among the about 4 million variants included in simulated disease genomes. To do so, we randomly added one of the regulatory Mendelian mutations to a randomly chosen, unaffected whole-genome sequence from the 1000 Genomes Project[14] and ran the resulting genome file through Genomiser using the default parameters and known inheritance model. We checked whether Genomiser was able to prioritize the spiked-in regulatory variant as the top candidate. We repeated this prioritization experiment on 10,419 simulated disease genomes. We tested Genomiser in three different experimental conditions, with either (1) no phenotype information, (2) the full phenotypic profile of the disease associated with the regulatory variant taken from our public dataset of HPO disease annotations, or (3) a more realistic clinical phenotype profile. For the latter more realistic clinical scenario, we (1) simulated *incomplete phenotyping* by randomly limiting the profile to three HPO terms, (2) simulated *imprecise phenotyping* by changing two of these terms to less specific parental term, and (3) simulated atypical/confounding presentation by adding a further two random HPO terms from the whole of HPO.[1] The recently published Phen-Gen tool[61] also has the capacity to process HPO-encoded phenotypic information and whole-genome data, so we additionally compared our performance against this using the same genomic and phenotypic profiles and identical allele frequency and inheritance model filtering.

Genomiser was able to prioritize the causative, regulatory variant as the top-scoring candidate in 77% of the genomes when using the full phenotypic profile (Figure 4). There is a slight reduction in performance to 68% when using the restricted phenotypic profile that is more likely to represent the type of phenotype annotations collected in realistic clinical settings. In both scenarios, our results represent a substantial improvement over Phen-Gen, which achieved performances of 19% and 14% using the full or restricted phenotypes, respectively. Performance did vary by variant category; the 5′ UTR, RNA gene, and microRNA gene mutations were the easiest to prioritize and the 3′ UTR mutations were particularly difficult. When phenotype data were not used by Genomiser, the performance dropped substantially to 23%. Genomiser offers a flexible framework where other non-coding deleteriousness prediction methods such as CADD can be used instead of using our ReMM score. With CADD, the performance was 71% and 61% when using the full or restricted phenotypic profiles, respectively, and without phenotype data the causative variant was not seen as the top scoring hit in any samples.

We note that the ReMM scores used in the Genomiser experiment were computed by 10-fold cross validation: the scores for the mutations included in each fold were obtained through a model trained on mutations not included in that fold, but only on those of the remaining nine. In other words, the ReMM score used in Genomiser for a

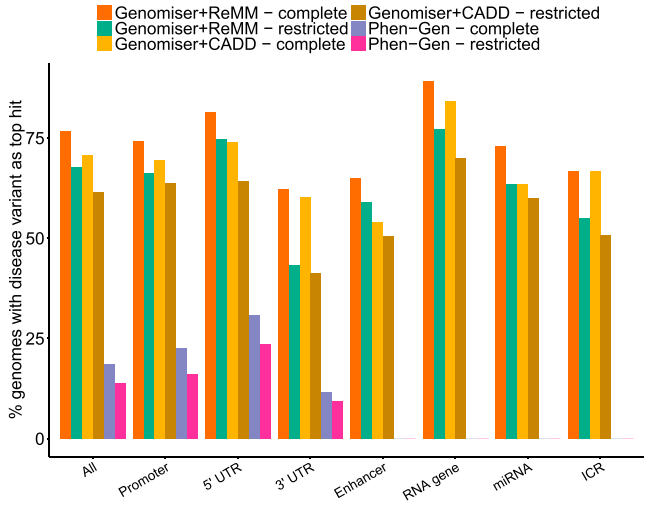

Figure 4. Performance Evaluation of Genomiser
The curated Mendelian regulatory mutations were added one at a time to unaffected genomes from the 1000 Genomes Project to generate 10,419 simulated disease genomes (see Material and Methods). As an additional test, the same simulations were performed using the CADD score instead of the ReMM score. The genomes were also run under the same frequency, inheritance, and phenotype conditions through Phen-Gen. Bars show percentage of genomes in which the true variant was prioritized as the top hit when assessing all the genomes or the subcategories involving promoter, UTR, enhancer, RNA gene, microRNA gene (miRNA), and imprinting control region (ICR) variants.

specific mutation was obtained by a model not trained on this variant.

Finally, although our focus in this manuscript is on methods to prioritize non-coding Mendelian mutations, we note that the Genomiser software application also makes use of our previously published methods for coding variants.[56,66] To assess the performance of Genomiser on 22 published cases of compound heterozygosity in which one causal mutation is regulatory and the other is coding or splice site (Table S5), we spiked both mutations into a genome VCF file and ran our analysis as above. The causative gene was ranked top in 18 (84%) of samples, demonstrating the ability of Genomiser to integrate information about coding and non-coding variants into the prioritization process.

Discussion

In this work we have presented a complete framework, Genomiser, for the prioritization of non-coding variants in Mendelian disease that offers a quick and effective means to identify such variants from whole-genome sequences. The final framework combines ReMM with other measures of variant candidacy, predicted regulatory regions, and a measure of a regulated gene's candidacy based on similarity of the phenotypic profile observed in an individual under investigation by WGS to existing knowledge, making use of the integrated cross-species genotype-phenotype knowledge base developed by the Monarch Initiative.[67] In 77% of samples, Genomiser is able to identify the causative regulatory Mendelian mutation as the top candidate out of the 4 million plus variants in a whole genome. This approach has the potential to substantially accelerate the detection of pathogenic, non-coding Mendelian variants by NGS and to explore the role of this currently understudied category of mutations. Although our focus here is on regulatory variants, Genomiser can still identify causative coding variants with high accuracy as in the original Exomiser application.

In order to computationally predict the consequences of NCVs in the human genome, two types of training data are required: disease associated (positive) and disease unassociated (negative). Previous methods have been designed to detect functional NCVs in general rather than solely those NCVs that cause Mendelian disease;[7–9,68] the latter set is difficult to find due to the fact that available databases contain errors[69] and conflate Mendelian and GWAS-associated variants. Therefore, for this work, we performed extensive and detailed literature curation to identify mutations that are associated with Mendelian disease and whose pathogenicity was judged to be plausible based on cosegregation, experimental evidence, or similar considerations. Our analysis of this collection of such mutations showed that they do in fact differ substantially from background positions in the genome (Figure 1).

This collection allowed us to train a ML classifier using only Mendelian disease-associated mutations. The methodologies used by CADD, FATHMM-MKL, GWAVA, DeepSEA, and EIGEN were designed for substantially larger positive sets and hence we developed our own ML strategy to overcome the challenges posed by our unique use case and make no claim as to the superiority of our method across all scenarios. However, for the prioritization of Mendelian mutations in whole-genome sequencing data, the ReMM score performed much better, as shown in Figure 2 and in the detailed results provided in the Supplemental Note and Figures S1–S3. The synergy of different factors explains the success of ReMM in scoring regulatory Mendelian mutations. At first, ReMM has been designed to deal with highly imbalanced data (we have a ratio of about 1:36,000 between positive and negative examples) and is a supervised machine learning algorithm that is designed to be used when reliable training data is available, such as our manually curated dataset of non-coding Mendelian mutations. Second, the oversampling of synthetic positive examples with SMOTE and the undersampling of negative examples are key factors to balance training set data and to avoid biased predictions toward the majority class (negative variants).[50] Third, the adoption of a hyperensemble strategy allows reliable base learners to be used (each base learner is a random forest) and also allowed most of the available search space to be covered, while maintaining a good balancing between positive and negative examples. Finally, taking the average of the scores computed by the hyperensemble of RFs can reduce the variance component of the error.[52]

Importantly, none of the competing methods was available within a start-to-finish application for phenome-driven WGS analysis such as the Genomiser. The modular software architecture of Genomiser allows different scoring methods to prioritize pathogenic variants to be used. By using CADD instead of ReMM on the same test data, we were able to rank 71% percent of the pathogenic variants in first place (versus 77% of top-ranked variants when ReMM is used; Figure 4). Future work will be needed to determine whether the ReMM score, or future versions of the ReMM score or one of the competing scores, will be useful for the full spectrum of non-coding Mendelian variation, which could conceivably differ in many ways from the small set of currently known noncoding Mendelian variants.

Whichever scoring methodology is used, variant analysis alone is unlikely to be useful to identify Mendelian disease-associated variants in WGS data, which typically contain more than 4 million variants, approximately 40,000 of which are locating in protein coding sequences. We benchmarked the recall of our regulatory Mendelian mutations when Genomiser was not used and just the variant scores alone were used to prioritize the 10,419 simulated whole genomes used in our main experiment. Using only simple filtering to remove any common variants with a MAF greater than 1%, prioritization by CADD or ReMM scores alone was not able to identify the causative variant as the top hit in any of the samples. Even looking at the top 10 or 100 variants, the causative variant was seen in only 0.2% and 4% of samples by CADD, and 7% and 18% by ReMM. This is not surprising because neither CADD nor ReMM use phenotypic information; rather, they are designed to assess the potential deleteriousness/pathogenicity of genetic variants irrespective of a specific Mendelian disease. Here, we use our phenotype-driven approach for prioritizing disease genes that we have developed for and previously applied for CNVs, clinical exome analysis, and whole-exome analysis.[1,66,70] We then examine regulatory sequences assigned to the genes that have been prioritized in this way and rank the associated regulatory variants based on a combination of their ReMM score, allele frequency, and the similarity of the observed phenotypic features and existing knowledge of the gene (Figure 3). We show that our performance is approximately four times as good as the only previous algorithm able to prioritize WGS data for Mendelian disease (Phen-Gen[61]), being able to detect the causative, non-coding variant as the top candidate in 68%–77% of cases depending on whether a full or more realistic, restricted phenotypic profile is used (Figure 4). In contrast, Phen-Gen was able to identify the causative variant as the top hit in only 14%–19% of samples, depending on whether the full or restricted phenotypic profile was used. Even looking at the top 100 variants returned by Phen-Gen, the causative one was identified in only 31%–34% of samples. Phen-Gen uses its own model for predicted non-coding pathogenicity and is trained on positive sets of HGMD regulatory variants and GWAS hits and a neutral set of common (>30% allele frequency) variants using evolutionary conservation, function signals from ENCODE, and proximity to coding regions as properties. The issues discussed above with these positive sets not fully representing true disease-causing variants probably accounts for some of the reduced performance. Given the effort required in pursuing candidate variants to establish causality, especially for regulatory variants, computational prioritization routines need to regularly place the true causal variant near the top of the list to be effective. We would therefore argue that Genomiser offers an effective solution for identifying causative, non-coding Mendelian variants.

The inclusion of phenotype data is critical for the effective prioritization of the regulatory variants, with performance dropping from 68%–77% to 23% when Genomiser is run without any input HPO IDs with a consequent removal of filtering and prioritization by phenotypic similarity score but retention of frequency and regulatory feature filtering and prioritization by ReMM score and allele rarity. Although collecting a full and detailed phenotypic profile of the individual being investigated by WGS will certainly improve the chance of prioritizing the correct causative variant,[71] the semantic algorithms underlying Genomiser are robust in that they allow for partial and non-exact matching between the observed phenotypes and previous disease and model organism phenotypic features associated with the gene. Genomiser offers the prospect of discovering novel disease-gene association through the inclusion of model organism data that extends the phenotypic coverage across the human proteome, along with the guilt-by-association approach covering any remaining genes without phenotype data. Throughout our analysis, coding and non-coding variants were simultaneously assessed and Genomiser can effectively identify both, as shown by the 84% performance for identifying compound heterozygous variants involving a coding and non-coding variant in the disease-associated gene. Genomiser can be freely downloaded as part of the Exomiser suite[55,66] and will process a whole genome in around 10 min on a standard desktop computer.

Where Genomiser failed to prioritize one of the regulatory Mendelian mutations as the top candidate, this was for a number of reasons. 8% of the Mendelian regulatory variants were lost during the filtering steps, with half lost because they are distal to a gene but do not yet fall into a predicted enhancer. In the remaining 16% of cases, the causative variant is detected but not as the top scoring candidate due to allele frequencies approaching 1% and/or a low ReMM score. Finally we note that although Genomiser performs well in the analysis of genomes simulated to contain non-coding Mendelian mutations, it is currently unclear how common non-coding Mendelian mutations are and thus how much of a performance boost can be expected by approaches such as the one presented here.

Initiatives such as the UK 100,000 Genomes project, the Precision Medicine Initiative, and many others are

poised to make genomic medicine part of health care for individuals with rare and common disease. However, to date, a tiny minority of published mutations in Mendelian disease have been found in non-coding sequences. A major but currently unanswerable question is whether this class of mutations (which comprise at least ten major pathomechanistic categories; Table 1) are more common than currently appreciated but simply have not been detected because of the historical focus on protein-coding exons and the fact that they were rarely sought in the Sanger sequencing era, and indeed still within most bioinformatics analysis of NGS data. Non-coding sequences, such as enhancer elements, have been poorly investigated[30] and the challenge of understanding how to interpret non-coding variants in diagnostic settings or in projects dedicated to the characterization of novel disease-associated genes is only beginning to be tackled. The answer to this question is pressing, because only about 25%–40% of individuals with suspected Mendelian disease who are investigated in large-scale whole-exome screening programs actually receive a diagnosis.[72–75] Although non-coding mutations not detected by whole-exome analysis are unlikely to be the only cause for the lack of a diagnosis in these individuals, WGS puts us for the first time in the position to test this hypothesis. In this work, we have presented effective algorithmic approaches designed especially to address this question, and we provide an application called the Genomiser that can be used on mid-range consumer hardware to analyze VCF files derived from WGS. We have focused in this report on small (<25 nt) non-coding mutations for many classes of mutation. A similar approach could be applied to the analysis of deep intronic splicing mutation and to "silent" changes in coding sequences that lead to misregulation. It is also likely that as more data on non-coding Mendelian mutations becomes available, it will be possible to improve the performance of ML approaches further and to develop bespoke classifiers for specific categories of mutation.

Supplemental Data

Supplemental Data include three figures, six tables, and a supplemental note and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.07.005.

Acknowledgments

This work was supported by grants from the European Union Seventh Framework Programme (FP7/2007-2013) ("SYBIL" grant No. 602300), NIH (1 U54 HG006370-01), the NIH Office of the Director (#5R24OD011883), the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy (Contract No. DE-AC02-05CH11231), the Bundesministerium für Bildung und Forschung (BMBF project numbers 0313911 and 01EC1402B), the Deutsche Forschungsgemeinschaft (DFG SP1532/2-1), and the DAAD Funding programme Research Stays for University Academics and Scientists (ID 57210259).

Received: April 4, 2016
Accepted: July 1, 2016
Published: August 25, 2016

Web Resources

1000 Genomes, http://www.1000genomes.org
CADD, http://cadd.gs.washington.edu/
DeepSEA, http://deepsea.princeton.edu/media/code/deepsea_train_bundle.v0.9.tar.gz
ExAC Browser, http://exac.broadinstitute.org/
FANTOM5 consortium, http://enhancer.binf.ku.dk/presets/permissive_enhancers.bed
FATHMM-MKL, https://github.com/HAShihab/fathmm-MKL
Genomiser download, ftp://ftp.sanger.ac.uk/pub/resources/software/exomiser/downloads/exomiser
Genomiser manual, https://exomiser.github.io/Exomiser
Genomiser source code, https://github.com/exomiser/Exomiser
GWAVA, ftp://ftp.sanger.ac.uk/pub/resources/software/gwava
Human Phenotype Ontology (HPO), http://www.human-phenotype-ontology.org/
Monarch Initiative, http://monarchinitiative.org
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
OMIM, http://www.omim.org/

References

1. Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci. Transl. Med. *6*, 252ra123.
2. Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA *312*, 1880–1887.
3. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.
4. Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.
5. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.
6. Ritchie, G.R., and Flicek, P. (2014). Computational approaches to interpreting genomic sequence variation. Genome Med. *6*, 87.
7. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
8. Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296.
9. Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict

the impact of regulatory variants from DNA sequence. Nat. Genet. *47*, 955–961.
10. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.
11. Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. *48*, 214–220.
12. Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., and Birney, E. (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. *18*, 1829–1843.
13. Paten, B., Herrero, J., Beal, K., Fitzgerald, S., and Birney, E. (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. *18*, 1814–1828.
14. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
15. Jäger, M., Wang, K., Bauer, S., Smedley, D., Krawitz, P., and Robinson, P.N. (2014). Jannovar: a java library for exome annotation. Hum. Mutat. *35*, 548–555.
16. Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756–D763.
17. Cazzola, M., and Skoda, R.C. (2000). Translational pathophysiology: a novel molecular mechanism of human disease. Blood *95*, 3280–3288.
18. Scheper, G.C., van der Knaap, M.S., and Proud, C.G. (2007). Translation matters: protein synthesis defects in inherited disease. Nat. Rev. Genet. *8*, 711–723.
19. Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.
20. Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106.
21. Jarinova, O., and Ekker, M. (2012). Regulatory variations in the era of next-generation sequencing: implications for clinical molecular diagnostics. Hum. Mutat. *33*, 1021–1030.
22. Jones, B.L., and Swallow, D.M. (2011). The impact of cis-acting polymorphisms on the human phenotype. HUGO J. *5*, 13–23.
23. Ma, M., Ru, Y., Chuang, L.-S., Hsu, N.-Y., Shi, L.-S., Hakenberg, J., Cheng, W.-Y., Uzilov, A., Ding, W., Glicksberg, B.S., and Chen, R. (2015). Disease-associated variants in different categories of disease located in distinct regulatory elements. BMC Genomics *16* (*Suppl 8*), S3.
24. Pickering, B.M., and Willis, A.E. (2005). The implications of structured 5′ untranslated regions on translation and disease. Semin. Cell Dev. Biol. *16*, 39–47.
25. Chen, J.-M., Férec, C., and Cooper, D.N. (2006). A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes I: general principles and overview. Hum. Genet. *120*, 1–21.
26. Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. USA *106*, 7507–7512.
27. Chatterjee, S., and Pal, J.K. (2009). Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. Biol. Cell *101*, 251–262.
28. Chuzhanova, N., Cooper, D.N., Férec, C., and Chen, J.-M. (2007). Searching for potential microRNA-binding site mutations amongst known disease-associated 3′ UTR variants. Genomic Med. *1*, 29–33.
29. Wethmar, K., Smink, J.J., and Leutz, A. (2010). Upstream open reading frames: molecular switches in (patho)physiology. BioEssays *32*, 885–893.
30. Gordon, C.T., and Lyonnet, S. (2014). Enhancer mutations and phenotype modularity. Nat. Genet. *46*, 3–4.
31. Epstein, D.J. (2009). Cis-regulatory mutations in human disease. Brief. Funct. Genomics Proteomics *8*, 310–316.
32. Sakabe, N.J., Savic, D., and Nobrega, M.A. (2012). Transcriptional enhancers in development and disease. Genome Biol. *13*, 238.
33. Khan, I.A., Mort, M., Buckland, P.R., O'Donovan, M.C., Cooper, D.N., and Chuzhanova, N.A. (2006). In silico discrimination of single nucleotide polymorphisms and pathological mutations in human gene promoter regions by means of local DNA sequence context and regularity. In Silico Biol. (Gedrukt) *6*, 23–34.
34. Savinkova, L.K., Ponomarenko, M.P., Ponomarenko, P.M., Drachkova, I.A., Lysova, M.V., Arshinova, T.V., and Kolchanov, N.A. (2009). TATA box polymorphisms in human gene promoters and associated hereditary pathologies. Biochemistry (Mosc.) *74*, 117–129.
35. Meola, N., Gennarino, V.A., and Banfi, S. (2009). microRNAs and genetic diseases. PathoGenetics *2*, 7.
36. Kawahara, Y. (2014). Human diseases caused by germline and somatic abnormalities in microRNA and microRNA-related genes. Congenit. Anom. (Kyoto) *54*, 12–21.
37. Cammaerts, S., Strazisar, M., De Rijk, P., and Del Favero, J. (2015). Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. Front. Genet. *6*, 186.
38. Hrdlickova, B., de Almeida, R.C., Borek, Z., and Withoff, S. (2014). Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochim. Biophys. Acta *1842*, 1910–1922.
39. Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet. *94*, 677–694.
40. Giardine, B., van Baal, S., Kaimakis, P., Riemer, C., Miller, W., Samara, M., Kollia, P., Anagnou, N.P., Chui, D.H.K., Wajcman, H., et al. (2007). HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. Hum. Mutat. *28*, 206–206.
41. Podlevsky, J.D., Bley, C.J., Omana, R.V., Qi, X., and Chen, J.J.-L. (2008). The telomerase database. Nucleic Acids Res. *36*, D339–D343.
42. Siepel, A., Pollard, K., and David, H. (2006). New methods for detecting lineage-specific selection. Proc. 10th Int. Conf. Res. Comput. Mol. Biol. (RECOMB 2006) 190–205.
43. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. *43*, D670–D681.
44. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

45. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493–D496.
46. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al.; FANTOM Consortium (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.
47. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.
48. Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., et al. (2013). DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. *41*, D936–D941.
49. Riggs, E.R., Jackson, L., Miller, D.T., and Van Vooren, S. (2012). Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience. Hum. Mutat. *33*, 787–796.
50. He, H., and Garcia, E. (2009). Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. *21*, 1263–1284.
51. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. *16*, 321–357.
52. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.
53. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA data mining software. ACM SIGKDD Explor. Newsl. *11*, 10.
54. Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536–1543.
55. Smedley, D., Jacobsen, J.O.B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat. Protoc. *10*, 2004–2015.
56. Smedley, D., and Robinson, P.N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Med. *7*, 81.
57. Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P.R. (2015). The ensembl regulatory build. Genome Biol. *16*, 56.
58. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.
59. Le Cessie, S., and van Houwelingen, J. (1992). Ridge estimators in logistic regression. Appl. Stat. *41*, 191–201.
60. Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE *10*, e0118432.
61. Javed, A., Agrawal, S., and Ng, P.C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat. Methods *11*, 935–937.
62. Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat. Methods *12*, 841–843.
63. Bone, W.P., Washington, N.L., Buske, O.J., Adams, D.R., Davis, J., Draper, D., Flynn, E.D., Girdea, M., Godfrey, R., Golas, G., et al. (2016). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet. Med. *18*, 608–617.
64. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C.M., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. *42*, D966–D974.
65. Ibn-Salem, J., Köhler, S., Love, M.I., Chung, H.-R., Huang, N., Hurles, M.E., Haendel, M., Washington, N.L., Smedley, D., Mungall, C.J., et al. (2014). Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol. *15*, 423.
66. Robinson, P.N., Köhler, S., Oellrich, A., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., et al.; Sanger Mouse Genetics Project (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. *24*, 340–348.
67. Mungall, C.J., Washington, N.L., Nguyen-Xuan, J., Condit, C., Smedley, D., Köhler, S., Groza, T., Shefchek, K., Hochheiser, H., Robinson, P.N., et al. (2015). Use of model organism and disease databases to support matchmaking for human disease gene discovery. Hum. Mutat. *36*, 979–984.
68. Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. *47*, 276–283.
69. Bell, C.J., Dinwiddie, D.L., Miller, N.A., Hateley, S.L., Ganusova, E.E., Mudge, J., Langley, R.J., Zhang, L., Lee, C.C., Schilkey, F.D., et al. (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci. Transl. Med. *3*, 65ra4.
70. Köhler, S., Schoeneberg, U., Czeschik, J.C., Doelken, S.C., Hehir-Kwa, J.Y., Ibn-Salem, J., Mungall, C.J., Smedley, D., Haendel, M.A., and Robinson, P.N. (2014). Clinical interpretation of CNVs with cross-species phenotype data. J. Med. Genet. *51*, 766–772.
71. Washington, N.L., Haendel, M.A., Köhler, S., Lewis, S.E., Robinson, P.N., Smedley, D., and Mungall, C.J. (2013). How good is your phenotyping? Methods for quality assessment. In Phenoday2014.Biol.-Lark.Org, pp. 1–4.
72. de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.
73. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. JAMA *312*, 1870–1879.
74. Zhu, X., Petrovski, S., Xie, P., Ruzzo, E.K., Lu, Y.-F., McSweeney, K.M., Ben-Zeev, B., Nissenkorn, A., Anikster, Y., Oz-Levi, D., et al. (2015). Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet. Med. *17*, 774–781.
75. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med. *369*, 1502–1511.