Prioritization of non-coding variants based on human intolerance to variation and primary sequence context using deep learning

Dimitrios Vitsios[1], **Ryan S. Dhindsa**[1], **Ayal B. Gussow**[2], **Slavé Petrovski**[1]

[1] Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK

[2] National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA

Abstract

Elucidating functionality in non-coding regions is a key challenge in human genomics. It has been shown that intolerance to variation of coding and proximal non-coding sequence is a strong predictor of human disease relevance. Here, we integrate intolerance to variation, functional genomic annotation (such as methylation and chromatin accessibility) and primary genomic sequence to build "Junk Annotation" Residual Variation Intolerance Score (JARVIS): a comprehensive deep learning model to prioritize non-coding regions. JARVIS outperforms comparable human lineage-specific scores in inferring pathogenicity of non-coding variants. Furthermore, despite not incorporating information on evolutionary conservation, JARVIS performs comparably, and in some cases, outperforms, other conservation-based scores in classifying pathogenic single-nucleotide and structural variants. This provides a unique and complementary prioritization paradigm to the heavily relied upon phylogenetic conservation-based predictions. In constructing JARVIS, we introduce a new intolerance metric: the genome-wide Residual Variation Intolerance Score (gwRVIS), which uses a sliding-window approach applied to Whole Genome Sequencing (WGS) data from 62,784 individuals. gwRVIS is among the most important features in JARVIS, and we verified that gwRVIS can distinguish sequence occupied by human Mendelian disease genes from more tolerant CCDS regions and intergenic sequence. Both JARVIS and gwRVIS capture previously inaccessible human-lineage constraint information to help prioritize genetic variants found in the human non-coding regulatory sequence and will enhance our understanding of the non-coding genome.

1 Introduction

The growing collection of human whole genome sequencing data has allowed researchers to identify stretches of the genome that are preferentially intolerant to genetic variation. For the protein-coding component of the human genome, we now have multiple metrics that capture disease potential at the level of the gene[1,2] and regions within a gene[3,4,5] with high confidence. These scores have transformed our ability to identify disease-causing mutations in the exome[1,6]. However, the majority of human genetic variation resides in non protein-coding regions of the genome[7,8], and our ability to interpret variants has been limited because the functional importance of these regions is largely unknown.

Early studies have attempted to introduce methods that assess intolerance to mutation in the non-coding genome to improve our understanding of variation in these regions. While these metrics have shown promise[6,9,10], their resolution has been limited due to small sample sizes of whole-genome sequencing (WGS) reference cohorts and may be biased due to strong dependence on evolutionary conservation in addition to human constraint in constructing non-coding intolerance scores. However, regulatory elements have high evolutionary turnover[11], which can obfuscate the use of conservation to interpret variation for many regions in the non-coding genome. The increasing sizes of WGS reference cohorts now offers an opportunity to assess intra-species human variation at an unprecedented resolution.

Here, we introduce intolerance metrics that examine regions of the non-coding genome that may be purged of extensive variation due to purifying selection within the human lineage, adopting a larger reference set and a machine learning based approach. We have previously introduced ncRVIS[6], which quantifies constraint in

proximal non-coding regions such as promoters and untranslated regions. We first expand this method to the entire genome using a sliding window approach (with single nucleotide resolution) to create genome-wide RVIS (gwRVIS). We then integrate genome-wide intolerance with information on primary genomic sequence and additional functional genomic annotation to build a novel comprehensive pathogenicity prediction framework for non-coding variants in the human genome. We intentionally do not employ any conservation information for the construction of our novel scores. This allows us to pinpoint regions that are more likely to be human-specific in terms of their functional relevance and provide a complementary human-lineage score to the many established phylogenetic conservation-based scores. Our metrics aim to facilitate prioritizing regions among the non-coding human genome which when mutated may be more likely to correlate with a clinically relevant effect.

2 Methods

2.1 Genome-wide Intolerance to Variation Score

We first sought to construct a score that captures the genome-wide intolerance to variation profile. We applied a tiled genome-wide Residual Variation Intolerance Score (gwRVIS) to whole genome sequencing (WGS) data from 62,784 individuals available in the TOPMed dataset[12] (*Freeze5* release). We scan the entire genome with a sliding-window approach (using a 1-nucleotide step), recording the number of all variants and common variants, irrespective of their predicted effect, within each window, to eventually calculate a single-nucleotide resolution genome-wide intolerance score. Taking into consideration the largest segregation achieved between the most intolerant and tolerant genomic classes we selected a window length of 3kb and Minor Allele Frequency (MAF) threshold of 0.1%. We then fit an ordinary linear regression model to predict common variants based on the total number of all variants found in each window. Eventually, we define the studentised residuals of this regression model as genome-wide Residual Variation Intolerance Score (gwRVIS), with lower gwRVIS values corresponding to greater intolerance.

2.2 A multi-module deep learning framework for non-coding variant pathogenicity inference

Equipped with a novel human-lineage specific constraint score that spans the entire human genome we next sought to further improve our ability to prioritize noncoding sequence by integrating additional information beyond gwRVIS. Thus, we integrate two additional layers of information: a) primary genomic sequence context around each variant (unstructured data) and b) genomic annotations such as methylation, chromatin accessibility or other structured features extracted from raw genomic sequence, such as GC content and mutability rate. By combining this information (gwRVIS, primary genomic sequence context and additional genomic annotations) we developed “Junk Annotation” RVIS or JARVIS: a multi-module deep learning framework for pathogenicity inference of non-coding regions that still remains naïve to existing phylogenetic conservation metrics in its score construction. We trained four different models for JARVIS: a) Gradient Boosting using structured features (i.e. without raw sequence information) b) feed-forward Deep Neural Net (DNN) using structured features, c) Convolutional Neural Net (CNN) with raw sequence input and d) the multi-module neural network model that combines information from both structured features and raw sequences (**Fig. 1**).

As our training set, we adopted all non-coding variants annotated in ClinVar as “Pathogenic” or “Likely pathogenic” and a random subset of “control” variants from denovo-db[13], considered to be benign. To build a generic non-coding variant classification model, we integrated variant data from five non-coding regions during training: intergenic regions, UTRs, lincRNAs, UCNEs and VISTA enhancers. The multi-module model outperformed all other models used for training JARVIS on the ClinVar pathogenic variant set , achieving an

AUC of 0.940 with 5-fold cross-validation (compared to AUC scores of 0.930, 0.929 and 0.872, from the DNN, Gradient Boosting and CNN models, respectively). Thus, we define as JARVIS the scores extracted by the multi-module model, which comprises of a CNN module for information inference from underlying sequence and a feed-forward DNN to assess structured feature data such as gwRVIS, sequence-derived features and external annotations.

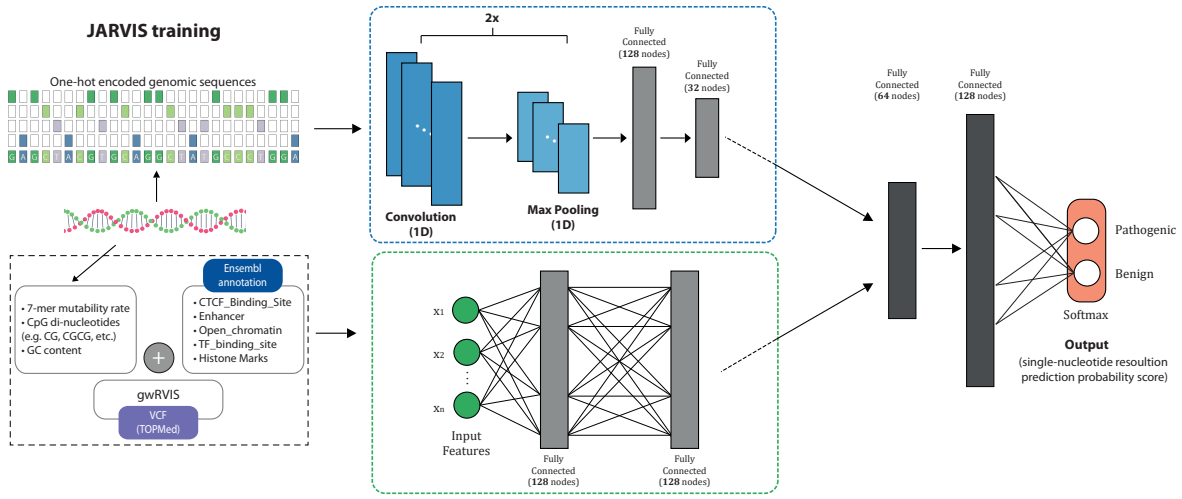

Fig 1. JARVIS: a multi-module deep-learning based score for non-coding variants pathogenicity inference with single-nucleotide resolution. Deep-learning framework for non-coding variants pathogenicity inference based on different types of annotation: genome-wide Residual Variation Intolerance Score (gwRVIS), primary genomic sequence, structured features extracted by raw genomic sequence (e.g. mutability rate, GC content, etc.) and additional annotations from Ensembl (e.g. histone marks, chromatin accessibility, CTCF binding sites, etc). All structured features are initially passed onto a 2-layer Deep Neural Net (DNN). Primary genomic sequences (in windows of 3kb) are fed into a deep Convolutional Neural Net and then flattened prior to merge with the higher representations of the structured features previously processed by the DNN. The combined higher representations of features are processed by two additional fully connected layers, followed by a '*softmax*' output, which gives the pathogenicity likelihood for each input variant as a probability score.

3 Results

3.1 Stratifying the human genome based on intolerance to variation

To determine biological relevance of gwRVIS we first sought to confirm the ability of gwRVIS to differentiate between different classes of protein-coding CCDS (Consensus Coding Sequence) windows based on their disease relevance. Despite not incorporating functional protein-coding annotations, gwRVIS manages to correctly stratify different CCDS sets based on their expected levels of constraint, grouping them in order of decreasing intolerance to variation as follows: OMIM-Haploinsufficient, 25% most intolerant CCDS, rest of CCDS and 25% most tolerant CCDS. We then studied seven major genomic classes: intergenic regions, lincRNAs, introns, CCDS, UTRs, VISTA enhancers[6] and UCNEs, listed here in ascending order of inferred intolerance to variation. The intergenic gwRVIS score distribution emerges as the most tolerant class with a median gwRVIS=-0.0014. This median gwRVIS closely aligns with the theoretical null distribution defined by gwRVIS=0, reflecting an equal number of observed and expected common variants.

Surprisingly, the CCDS protein-coding region of the genome was not the most intolerant functional class. We observed that Ultra Conserved Non-coding Elements[14] (UCNEs; highly conserved non-coding regions between human and chicken) are ranked as the most intra-species intolerant class (median gwRVIS: -0.99; Mann-Whitney U vs intergenic: $p<1x10^{-320}$), and this is despite gwRVIS not using any information about species conservation in its construction. VISTA enhancers (a class of highly conserved enhancers active during

embryonic development) and CCDS follow with the next highest intolerance to variation profile, very similar to UTRs (median gwRVIS: -0.77, -0.55 and -0.51; Mann-Whitney U vs intergenic: $p<1x10^{-320}$ for VISTA enhancers, CCDS and UTRs, respectively). Finally, introns and lincRNAs have a more tolerant gwRVIS score distribution that more closely resembles the distribution from intergenic regions, but due to sheer size of the corresponding score lists remains highly significant (median scores: -0.050, -0.0015; Mann-Whitney U vs intergenic: $p<1x10^{-320}$ and $p=2.6x10^{-168}$, respectively).

3.2 Classification of non-coding pathogenic variants based on their intolerance to variation

Overall, non-coding variants represent a small fraction of all pathogenic classified variants residing among curated variant-level resources such as ClinVar[15]. Here, we examine the properties of gwRVIS in context of ClinVar clinically-classified pathogenic non-coding variants. We compiled two lists of non-coding variants: a pathogenic set based on ClinVar and a set of benign variants based on the "control" variants from denovo-db[13] (spanning across intergenic regions, lincRNAs, VISTA enhancers, UCNEs or UTRs). We then trained a logistic regression model with 5-fold Cross-Validation, using gwRVIS or another genome-wide score (CADD[16], phastCons46way[17], phyloP46way[18] and Orion) as the only independent variable predictor. Remarkably, we observed that gwRVIS outperforms all other scores in pathogenic variant classification from lincRNA regions (AUC=0.937; **Fig. 2a**) and intergenic regions (AUC=0.763; **Fig. 2b**). In UTRs, gwRVIS's performance drops but remains considerably higher than Orion (AUCs: 0.732 versus 0.597; **Fig. 2c**).

In order to estimate the novel contribution of gwRVIS information in non-coding variant detection, we also trained a multiple logistic regression model using gwRVIS and CADD as the independent variables. We observe, that gwRVIS boosts CADD's performance (**Fig. 2**) in lincRNAs (AUC: increased to 0.937 from 0. 895), intergenic regions (AUC: increased to 0.809 from 0.741) and UTRs (AUC: increased to 0.835 from 0.777). This indicates that gwRVIS captures novel information that is not represented among the 63 distinct annotations/features employed by CADD. Moreover, although ncRVIS is the top performing single score in UTRs (AUC=0.823), it is lower than the combined gwRVIS and CADD score (AUC=0.835).

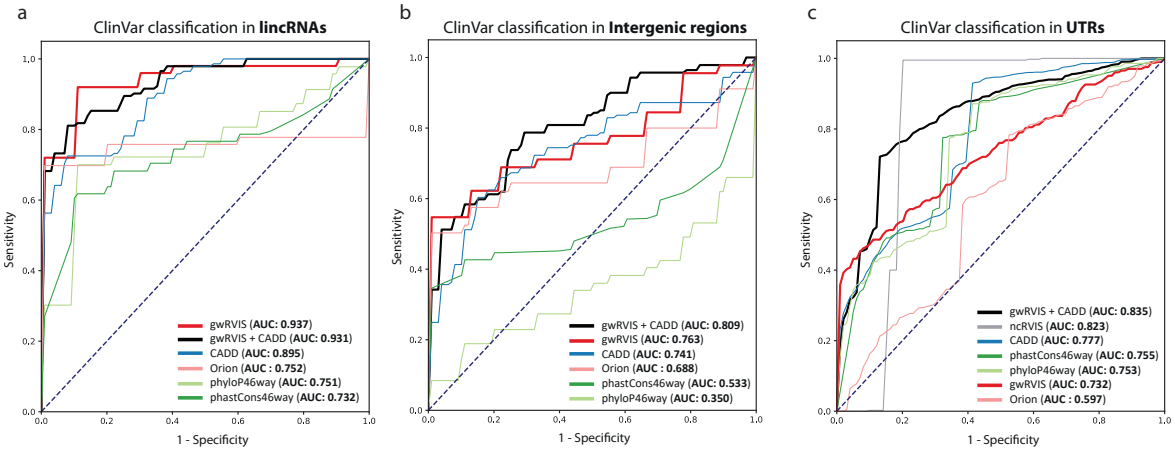

Figure 2. Predictive power of gwRVIS for pathogenic variant classification. Mean ROC curves (with 5-fold cross-validation) from gwRVIS benchmarking against CADD, phastCons (46-way), phyloP (46 way) and Orion, during ClinVar pathogenic-vs-benign variants classification for three non-coding genomic classes: **a)** lincRNAs, **b)** Intergenic regions and **c)** UTRs. The combined performance of gwRVIS with CADD is also shown. ncRVIS is included in the benchmarking of the UTR regions (**c**), as a robust score specifically designed for the UTR genomic class.

3.3 JARVIS performance on pathogenicity likelihood inference

We compared JARVIS against eight other popular genome-wide scores: CADD[16], phastCons[17] (46way), phylop[18]

(46way), DANN[19], LINSIGHT[20], ncER[21], CDTS[22] and Orion[9]. It is important to note that ncER, LINSIGHT, CADD and DANN incorporate multiple phylogenetic conservation metrics (e.g. phyloP, phastCons, SiPhy and CEGA) in their score constructions. We trained the deep learning-based multi-module JARVIS model using all ClinVar non-coding pathogenic variants across all chromosomes with 5-fold cross-validation and compared its performance against the rest of the scores. JARVIS outperforms all other scores (AUC=0.940; **Fig. 3a**), despite some of them including conservation information. Two scores, LINSIGHT and ncER, achieved better performance on this dataset (AUC: 0.961 and 0.977, respectively), however, they are either overfit on the JARVIS training set or potentially biased with additional information, such as distance from the closest TSS. When integrating the TSS distance in the JARVIS model, this version of JARVIS indeed exceeds the performance of both LINSIGHT and ncER (AUC=0.984). However, we don't eventually include TSS distance as a feature in the final model as we want to avoid overfitting JARVIS predictions towards variants residing closer to protein-coding regions.

Based on the performance of JARVIS when using different models for training, we observed that deep learning models are superior to Gradient Boosting and also that inclusion of raw sequences features provides the highest predictive power in the pathogenic variant classification task from non-coding regions. It is, however, difficult to infer the real contribution from each feature employed by the full deep learning-based JARVIS model. Thus, we employ an impurity-based feature extraction algorithm with a Gradient Boosting classifier as a proxy to infer the relative contribution of each of the structured features. We observe that gwRVIS ranks second in feature importance while the sequence-derived features, specifically CpG density, are the other most dominant subset from the entire feature set (**Fig. 3b**). Functional annotations follow with lower contribution but still carrying a considerable burden, especially for certain types of histone marks. These findings demonstrate that the genome-wide intolerance score, as it is captured by gwRVIS, adds considerable value to the predictive power of JARVIS.

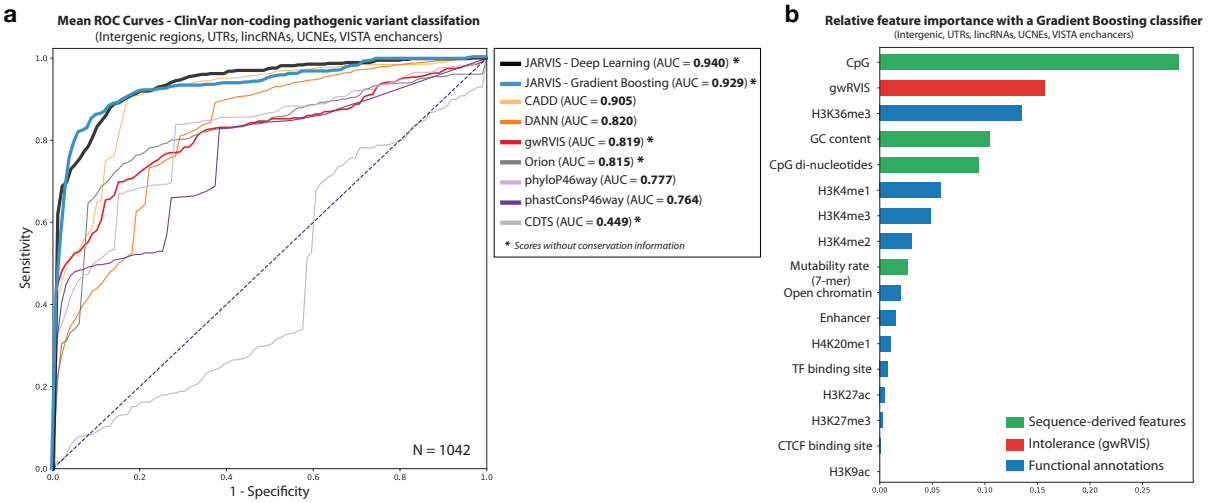

Fig 3. JARVIS cross-validation performance and feature importance. a) JARVIS performance with 5-fold Cross-Validation after training with a multi-module deep neural network, using all non-coding ClinVar-based pathogenic variants and a matched set of putative benign variants from denovo-db. Variants used for training belong to any of the following genomic classes: intergenic regions, UTRs, lincRNAs, UCNEs or VISTA enhancers. A total of 521 non-coding pathogenic variants have been used for this classification task, thus N=1042 represents the total size of the training set (using a set of control variants of equal size). Performance for the rest of genome-wide scores shown here has been calculated using a logistic regression model with 5-fold cross-validation on the JARVIS training set. **b)** An overview of the relative importance of the structured features integrated within JARVIS, as they are extracted by a Gradient Boosting classifier following an impurity-based feature selection algorithm.

3.4 Prioritization of pathogenic structural variants

Finally, we sought to estimate the ability of JARVIS and gwRVIS to distinguish large structural variants based on their inferred clinical impacts. We employ for this task a rich set of structural variants (SV) called from 14,891 whole genome sequences in the gnomAD dataset (v2.1). Structural variants have been annotated with various functional consequences with regards to coding or proximal-coding sequences (Copy Gain, Duplication-LoF, Intronic, LoF, Partial-Duplication, Promoter, UTR and Whole-Gene inversion) or otherwise classified to have an effect on intergenic regions. We consider the latter case (SV in intergenic regions) as our benign set and try to classify it against all other sets of putative pathogenic structural variants. Using a 10-fold cross-validation approach with a logistic regression classifier for all benchmarked scores, JARVIS achieved the highest performance in six out of eight comparisons (**Fig. 4**; AUC=0.684-0.844), outperforming all other scores that follow with lower AUC ranges (Orion: 0.591-0.755; LINGISHT AUC: 0.605-0.747; ncER: 0.448-0.709 and gwRVIS: 0.542-0.667, ordered by the highest AUC in each range).

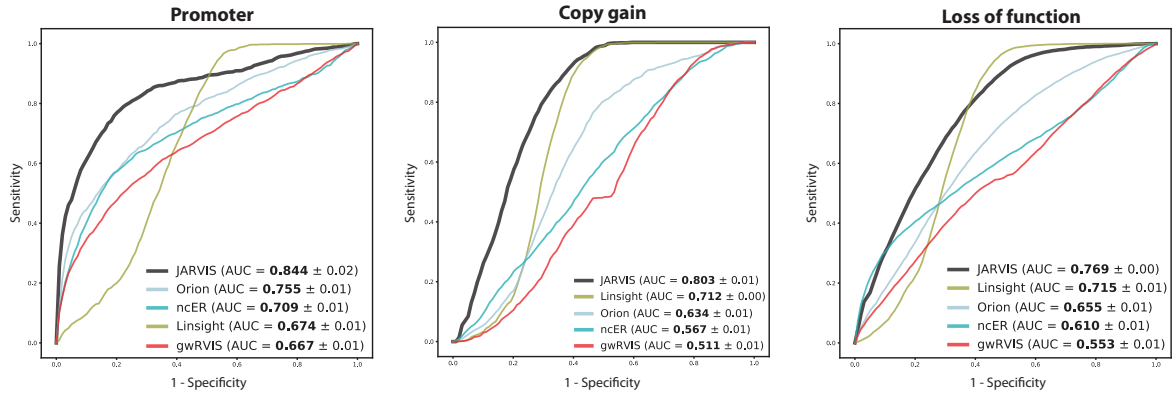

Figure 4. JARVIS and gwRVIS performance on structural variants. ROC curves from classification of benign structural variants (intergenic) against different sets of putative pathogenic ones, using a 10-fold cross-validation approach with a logistic regression model on five scores: JARVIS, gwRVIS, LINSIGHT, ncER and Orion.

4 Conclusion

Several methods have been developed in recent years that attempt to address the challenge of prioritizing non-coding variants. Most of these methods employ a combination of functional annotation but also cross-species conservation information. Here, we presented JARVIS and gwRVIS, two scores that encompass exclusively human lineage-specific information but still manage to perform comparably or even better than conservation-informed scores. Pathogenicity likelihood of non-coding regions cannot be efficiently inferred by cross-species conservation based metrics due to the high evolutionary turnover of these regions. Thus, the two human-lineage specific metrics we introduce may allow us to reduce dependence on conservation-derived metrics and increasingly rely on human genomic constraint in our search for human disease variants.

References

1. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet.* **9**, (2013).
2. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
3. Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* (2016). doi:10.1186/s13059-016-0869-4
4. Traynelis, J. *et al.* Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. doi:10.1101/gr.226589.117
5. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017). doi:10.1101/148353
6. Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLOS Genet.* **11**, e1005492 (2015).
7. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* (2002). doi:10.1038/nature01262
8. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* (2009). doi:10.1073/pnas.0903103106
9. Gussow, A. B. *et al.* Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLoS One* **12**, e0181604 (2017).
10. di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. doi:10.1038/s41588-018-0062-7
11. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (80-.).* (2010). doi:10.1126/science.1186176
12. NHLBI. NHLBI Trans-Omics for Precision Medicine Whole Genome Sequencing Program. *https://www.nhlbiwgs.org/* (2016).
13. Turner, T. N. *et al.* denovo-db: a compendium of human de novo variants . *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw865
14. Dimitrieva, S. & Bucher, P. UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.* **41**, D101-9 (2013).
15. Landrum, M. J. *et al.* ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1153
16. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
17. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* (2005). doi:10.1101/gr.3715005
18. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* (2010). doi:10.1101/gr.097857.109
19. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
20. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* (2017). doi:10.1038/ng.3810
21. Wells, A. *et al.* Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* (2019). doi:10.1038/s41467-019-13212-3
22. Di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0062-7
23. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* (2020). doi:10.1038/s41586-020-2493-4