Loss-of-function tolerance of enhancers in the human genome

Duo Xu[1,2,3,4], Omer Gokcumen[5], Ekta Khurana[1,2,3,4,*]

[1] Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA

[2] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA

[3] Englander Institute for Precision Medicine, New York Presbyterian Hospital-Weill Cornell Medicine, New York, NY 10021, USA

[4] Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10065, USA

[5] Department of Biological Sciences, University at Buffalo, The State University of New York, Buffalo, New York 14260, USA

***Corresponding author: ekk2003@med.cornell.edu**

Abstract

Previous studies have surveyed the potential impact of loss-of-function (LoF) variants and identified LoF-tolerant protein-coding genes. However, the tolerance of human genomes to losing enhancers has not yet been evaluated. Here we present the catalog of LoF-tolerant enhancers using structural variants from whole-genome sequences. Using a conservative approach, we estimate that each individual human genome possesses at least 28 LoF-tolerant enhancers on average. We assessed the properties of LoF-tolerant enhancers in a unified regulatory network constructed by integrating tissue-specific enhancers and gene-gene interactions. We find that LoF-tolerant enhancers are more tissue-specific and regulate fewer and more dispensable genes. They are enriched in immune-related cells while LoF-intolerant enhancers are enriched in kidney and brain/neuronal stem cells. We developed a supervised learning approach to predict the LoF-tolerance of enhancers, which achieved an AUROC of 96%. We predict 5,677 more enhancers would be likely tolerant to LoF and 75 enhancers that would be highly LoF-intolerant. Our predictions are supported by known set of disease enhancers and novel deletions from PacBio sequencing. The LoF-tolerance scores provided here will serve as an important reference for disease studies.

Introduction

Loss-of-function (LoF) variants in genes are defined as those which impair or eliminate the function of the encoded protein. Despite their protein-coding disruption, it has been shown that some LoF variants can be tolerated in healthy individuals (Ng et al. 2008, Pelak et al. 2010, Genomes Project et al. 2010, Telenti et al. 2016). Genes harboring homozygous LoF variants are called LoF-tolerant genes. Multiple studies have shown the average number of LoF variants ranges from 100~200 per individual (MacArthur et al. 2012, Genomes Project et al. 2015, Lek et al. 2016). In addition, MacArthur et al estimated that on average there are 20 LoF-tolerant genes per human genome (MacArthur et al. 2012). Such lists of LoF variants have greatly aided gene prioritization in disease studies by providing functional references for variants (Kathiresan and Srivastava 2012, Tg and Hdl Working Group of the Exome Sequencing Project et al. 2014, Gilissen et al. 2014, Genovese et al. 2016, Yu et al. 2018). It also enabled estimations of gene indispensability by providing a confident set of LoF variants and LoF-tolerant genes in human genomes (Khurana et al. 2013, MacArthur et al. 2012).

However, in stark contrast to protein-coding genes, our knowledge about the dispensability of non-coding regulatory elements is limited. The atlas of cell- and tissue-specific regulatory elements developed by large-scale efforts, such as ENCODE (Consortium 2012, Davis et al. 2018), Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics et al. 2015), FANTOM (Andersson et al. 2014) and the availability of thousands of whole-genomes makes this an opportune time to ask the same questions that were asked for protein-coding genes and to identify the non-coding elements that can tolerate homozygous LoF.

It has been shown that 'shadow' enhancers, defined as the ones that have similar functions to the proximal primary enhancers but locate at distal locations, can be deleted in Drosophila without affecting the viability (Perry et al. 2010). A recent study showed that deletion of some individual enhancers did not significantly affect the fitness of mice, but deletion of pairs of enhancers regulating the same gene led to abnormal limb development (Osterwalder et al. 2018). Thus, the phenotypic effect stemming from the loss of a single enhancer may be mitigated by the activity of another enhancer, whose function is redundant to the deleted one, and is therefore only apparent if both enhancers are deleted. This apparent redundancy of enhancers is hypothesized to provide robustness in gene expression in response to the fluctuating environmental conditions (Macneil and Walhout 2011, Wunderlich et al. 2015). These studies suggest that enhancers can act redundantly in groups instead of stand-alone units. On the other hand, alterations at single enhancers have been implicated in rare Mendelian diseases (Ghiasvand et al. 2011, Albuisson et al. 2011, Weedon et al. 2014, Kapoor et al. 2014, Kremer et al. 2017) and genome-wide association studies (GWAS) have found that many susceptibility loci for common diseases reside in enhancers (MacArthur et al. 2017, Trynka et al. 2013, Hindorff et al. 2009, Maurano et al. 2012, Wang et al. 2018). It is expected that loss of essential enhancers would have strong fitness consequences, while LoF-tolerant enhancers would lie at the other end of the spectrum and their loss would not elicit substantial phenotypic impact. Thus, it is important to have a prioritization scheme for LoF-tolerance vs. disease-causing potential of enhancers based on their essentiality.

However, due to the redundancy and complexity of tissue-specific regulatory networks, such prioritization of enhancers has long remained a challenging task.

Here we report a systematic computational approach that uses machine learning to predict the LoF-tolerance of all enhancers in the human genome. We built an integrated regulatory network, MegaNet, in which the nodes consist of enhancers and genes. The edges between enhancers and genes correspond to tissue-specific regulation and those between genes include protein-protein (Stark et al. 2006), metabolic (Kanehisa et al. 2010), phosphorylation (Lin et al. 2010) and signaling interactions (Korcsmáros et al. 2010). We used deletions from 2,054 whole-genomes to identify the LoF-tolerant enhancers in this network while taking ultra-conserved enhancers with experimentally validated enhancer activity as LoF-intolerant (Bejerano et al. 2004, Dickel et al. 2018, Visel et al. 2007, Visel et al. 2008). We used the characteristic differences between LoF-tolerant and LoF-intolerant enhancers in MegaNet to build a random forest model to predict the LoF-tolerance of all enhancers in the human genome. The LoF-tolerance scores of enhancers provided in this study can significantly facilitate the interpretation and prioritization of non-coding sequence variants for disease and functional studies.

Results

Construction of MegaNet

Integration of transcription factor (TF) binding profiles, chromatin features and expression data has revealed the architecture of regulatory networks (He et al. 2014, Yip et al. 2012, Zhu et al. 2016, Whalen, Truty and Pollard 2016, Roy et al. 2016). Availability of tissue-specific annotations has also enabled the construction of tissue-specific regulatory networks (Cao et al. 2017). In order to systematically evaluate the LoF-tolerance of enhancers in tissue-specific regulatory networks, we collected 246,028 unique enhancers regulating 19,170 genes from enhancer-target networks (Cao et al. 2017). We constructed an integrated mega network (MegaNet) for joint assessment of the enhancer properties in the enhancer-gene regulation networks (Cao et al. 2017) and gene centrality in the gene-gene interaction networks (Khurana et al. 2013). The gene-gene interactions in MegaNet consist of protein-protein (Stark et al. 2006), metabolic (Kanehisa et al. 2010), phosphorylation (Lin et al. 2010) and signaling interactions (Korcsmáros et al. 2010).

In the MegaNet, enhancers and genes represent the two kinds of nodes. The directed regulation from enhancers to genes and the undirected interactions between genes are the edges. In order to annotate the tissue-specific properties of nodes and edges in the MegaNet, the enhancer->gene regulation edges are weighted by the number of tissues in which they are active and annotated by tissue types (Figure 1a).

LoF-tolerant enhancers

We adopted the enhancers annotated by Cao et al. (Cao et al. 2017, Methods) which were collected from the ENCODE and Roadmap Epigenomics projects (Consortium 2012, Roadmap Epigenomics et al. 2015). Since samples in the 1000 Genomes Project consist of individuals without strong disease phenotypes (Genomes Project et al. 2010, Auton et al. 2015), we define enhancers that can be homozygously deleted in those individuals as LoF-tolerant enhancers. This is similar to the approach used previously for identification of LoF-tolerant genes (Ng et al. 2008, Pelak et al. 2010, MacArthur et al. 2012). More specifically, to identify the LoF-tolerant enhancers, we identified deletions which occur homozygously in at least one individual among the 2,504 from the 1000 Genomes Project (Sudmant et al. 2015) and intersected them with enhancers. In order to avoid bias introduced by protein-coding regions, deletions that overlap coding exons were excluded. While deletion of parts of enhancers may also lead to loss of their activity, we used a conservative estimation of LoF-tolerant enhancers by only including those that are completely deleted in a homozygous manner. In line with this, our approach also does not include LoF of enhancers by SNVs due to the difficulties in predicting their functional impact. In total, 886 enhancers are identified as LoF-tolerant. The number of LoF-tolerant enhancers per individual genome ranges from 8 to 78 (Supplementary Figure 1).

LoF-intolerant enhancers

Efforts to identify enhancers that are intolerant to LoF have relied on evolutionary conservation to identify the ultra-conserved non-coding elements in the genome (Bejerano et al. 2004). 256 ultra-conserved non-exonic elements have been identified by absolute conservation between orthologous regions of the human, rat and mouse genomes. While the initial study to probe the indispensability of ultra-conserved enhancers showed that their deletion does not affect the viability of mice (Ahituv et al. 2007), more recent studies have found that the mice suffer from severe developmental defects (Dickel et al. 2018), indicating that ultra-conserved enhancers are in fact LoF-intolerant as their loss strongly adversely affects organismal fitness. Overall, we compiled 49 LoF-intolerant enhancers, which correspond to the ultra-conserved non-coding elements that have shown enhancer activity by consistent reporter gene expression in at least three transgenic mice embryos (Bejerano et al. 2004, Visel et al. 2007, Visel et al. 2008, Dickel et al. 2018). Furthermore, in agreement with previous studies (Katzman et al. 2007), we observe a depletion of common polymorphisms and an enrichment of rare variants at LoF-intolerant enhancer regions, providing additional support for negative selection preventing mutations accumulating in LoF-intolerant enhancers (Supplementary Figure 2).

Properties of LoF-tolerant and -intolerant enhancers in the MegaNet

We analyzed the properties of enhancers in MegaNet using enhancer out-degree (EOD, number of genes that an enhancer targets), enhancer tissue ubiquity (ETU, total number of tissues the enhancer is active in), and enhancer->gene edge tissue ubiquity (EGTU, the number of tissues in which the edges are active) (detailed feature description provided in Supplementary Table 1). ETU describes the total number of tissues that the enhancer is active in, while EGTU describes the number of tissues that an enhancer->gene regulation edge is active in (Figure 1a). Khurana et al. integrated multiple biological networks to evaluate the functional essentiality of genes in the human genome (Khurana et al. 2013). We assigned the gene indispensability scores generated from that study to genes in our network to integrate the gene indispensability (GIS) in the MegaNet. In order to assess the enhancer-gene interaction landscape in the MegaNet, we also calculated the number of enhancers regulating each gene (Gene In-Degree, GID), and other network centrality metrics as additional gene properties (detailed feature description provided in

Supplementary Table 1). Due to the characteristic architecture of regulatory networks, an enhancer can regulate multiple genes and a gene can be regulated by multiple enhancers as well. Enhancers regulating multiple genes will have multiple values for each gene feature. We consider both the mean and variance to represent their values, and they are represented with an extension "a" (average) or "v" (variance). For example, the enhancer on the left in Figure 1a regulates two genes in three different tissues. The ETU of the enhancer is 3 while the EGTU is a collection of (2,1). The EGTUa for the enhancer will be 1.5 and EGTUv will be 0.25 (Methods).

LoF-tolerant enhancers tend to be tissue-specific and regulate fewer, more dispensable genes

We compared the network properties of LoF-tolerant and LoF-intolerant enhancers and genome-wide expectation (GW, all other enhancers in the MegaNet). We find that LoF-tolerant enhancers regulate significantly fewer genes (i.e., they have lower EOD) compared to genome-wide expectation and are active in fewer tissues (ETU) compared to both genome-wide expectation and LoF-intolerant enhancers (Figure 1b, Supplementary Figure 3a). In addition, genes regulated by LoF-tolerant enhancers are more dispensable (lower average gene indispensability score, GISa) compared to genome-wide expectation and LoF-intolerant enhancers. In order to account for enhancers with the same average EGTU, but different variance, we also analyzed the variance of EGTU. Both average edge tissue ubiquity (EGTUa) and its variance (EGTUv) are lower for LoF-tolerant enhancers, indicating that their interactions tend to be more tissue-specific (Figure 1b). Overall, these observations indicate that LoF-tolerant enhancers are in general less versatile in the genome and tend to target specific genes in specific tissues.

Genes regulated by LoF-tolerant enhancers are regulated by more enhancers

Interestingly, we observe that the genes that LoF-tolerant enhancers regulate, have significantly more enhancers regulating them (higher Average Gene In-degree, GIDa) (Figure 1b). As mentioned in the Introduction, enhancers can act in groups rather than as single units (Perry et al. 2010, Macneil and Walhout 2011, Wunderlich et al. 2015, Osterwalder et al. 2018). Here, we show that this trend exists in a genome-wide manner, such that enhancers targeting genes that are regulated by multiple enhancers tend to be more LoF-tolerant. Thus, LoF-tolerant enhancers potentially function redundantly to prevent severe phenotypic effects when one or more enhancers are lost.

LoF-intolerant enhancers are enriched in kidney and brain/neuronal stem cells while LoF-tolerant enhancers are enriched in immune related cells

Furthermore, to analyze the tissue-specific properties of enhancers, we extracted the tissue-specific sub-networks from the MegaNet. We observe that different tissues exhibit differential enrichment of LoF-tolerant vs. LoF-intolerant enhancers. We calculated the odds ratio of LoF-tolerant and -intolerant enhancers for each tissue compared to the total number of LoF-tolerant and -intolerant enhancers across all other tissues respectively (Figure 2). We found that the proportion of LoF-intolerant enhancers in kidney and neuronal stem cell/brain tissues is significantly enriched (Fisher's exact test P-value = 0.010 and 2.80e-11 respectively, Figure 2). Interestingly, this trend is reversed in cells involved in immune response (Hematopoietic stem cells (HSC) & B-cell and T-cell), where LoF-intolerant enhancers are depleted while LoF-tolerant are enriched (Fisher's exact test P-value = 4.94e-4 and 1.70e-7, Figure 2). Our results are consistent with the previous knowledge that ultra-conserved enhancers are related to brain or developmental function (Dickel et al. 2018). Importantly, they show that enhancers involved in immune response tend to be more LoF-tolerant.

Supervised learning to predict enhancer loss-of-function tolerance

Enhancer->gene regulation occurs in a complex network with interactions between enhancers and genes and among genes. Thus, to systematically predict the LoF tolerance of enhancers, we built a random forest classification model based on the network properties of enhancers and genes in the MegaNet (in total 63 features for 15 tissues as described above and in Supplementary Table 1, Methods). We also included evolutionary conservation (Siepel et al. 2005) and gene dispensability scores (Khurana et al. 2013) (Methods).

In order to avoid the prediction bias introduced by unbalanced positive and negative sample sizes, we randomly chose 50 enhancers from the LoF-tolerant enhancer set and used the 49 LoF-intolerant enhancers as the negative set to train the model. The process was repeated 50 times to sample all the 886 LoF-tolerant enhancers for training, and our model achieved an average area under the receiver operating characteristic curve (AUROC) of 0.9633 +/- 0.0002 evaluated by 10-fold cross validation on the balanced sets (Methods).

In the training process, our model uses the ultra-conserved enhancers as LoF-intolerant enhancers. Therefore, our model may bias towards enhancers residing in regions with low conservation for the predicted LoF-tolerant enhancers. In order to evaluate whether conservation alone can separate LoF-tolerant and -intolerant enhancers, we trained one model using conservation as the only feature and another model using all other features except conservation. We used one positive set with randomly selected 50 LoF-tolerant enhancers to test our models. We obtained an average AUROC of 0.92 +/- 0.0523 for the conservation-only model and 0.79 +/- 0.1801 for the other-features-only model while the final model using all features achieved an average AUROC of 0.98 +/- 0.0256 (Figure 3a). These results confirm that the integrative model gives the best performance and that the network features enable substantially better discrimination between LoF-tolerant and -intolerant enhancers than conservation alone.

Next, we evaluated the importance of features by mean decrease impurity, which measures the decrease in the weighted impurity of the tree by each feature (Breiman 1984, Pedregosa et al. 2011). We find that while conservation is the most important feature (importance=0.3629, Figure 3b), other network properties provide substantial information to the model. On comparison of the feature importance, as expected, the features that show high importance in the model are the ones that show a significant difference between LoF-tolerant and -intolerant enhancers (as discussed above). Average edge tissue ubiquity (EGTUa) and enhancer tissue ubiquity (ETU) are the most important features after conservation, collectively contributing 18.7% of the importance. Gene indispensability scores (GIS), gene closeness centrality (GCC), out-degree of neuronal stem cell/brain enhancers and the average indegrees of genes they target also appear as important features in the model (Figure 3b).

Prediction of novel LoF-tolerant enhancers and validation using PacBio structural variants

We applied our model on all enhancers in the MegaNet, except the ones used in training. Out of 245,093 enhancers tested, 5,677 are predicted to be highly confident tolerant to LoF with high LoF-tolerance probability ($P_{\text{LoF-tol.}} > 0.95$), while 75 are predicted to be highly confident LoF-intolerant candidates with very low LoF-tolerance probability ($P_{\text{LoF-tol.}} < 0.05$, Supplementary Table 3). The predicted LoF-intolerant candidates show similar patterns to the ones in the training set as they tend to be active in more tissues (P-value = 2.408e-72) and regulate genes that are more indispensable (P-value = 1.556e-31) compared to LoF-tolerant candidates (Figure 3c, Methods).

We compared the number of predicted LoF-tolerant enhancers to the predicted number of dispensable genes from Khurana et al. (Khurana et al. 2013). For comparable analysis between genes and enhancers, we took highly confident LoF-tolerant and -intolerant enhancers and used the same cut-off for gene dispensability scores (scores lower than 0.05 for dispensable and higher than 0.95 for indispensable genes). We found that the ratio of predicted LoF-tolerant to LoF-intolerant enhancers (75.7, 5677:75) is significantly higher compared to the ratio of predicted dispensable to indispensable protein-coding genes (0.483, 1259:2606, Fisher's exact test P-value < 2.2e-16). This result is consistent with the hypothesis that LoF of genes would likely cause more severe fitness effects and must be under stronger negative selection than the LoF of regulatory elements.

Thus, in addition to the 886 homozygously deleted LoF-tolerant enhancers used in training, our model predicts additional 5,677 highly confident LoF-tolerant enhancers ($P_{\text{LoF-tol.}} > 0.95$). We postulate that many of these enhancers have not yet been detected as LoF-tolerant because of (a) the limited sample size of whole-genome sequences and (b) undetected deletions by short-read sequencing due to the limited mappability of short reads in repetitive and complex regions. In particular, recent studies have pointed out that the map of genomic deletions with Illumina short-reads is highly incomplete. The longer sequencing reads in PacBio technology enabled the detection of many additional structural variants (SVs, including deletions), particularly in high-repeat regions (24,825 as opposed to 10,884 per human genome) (Chaisson et al. 2015, Chin et al. 2013, Kronenberg et al. 2018, Chaisson et al. 2018). We tested the performance of our method on homozygously deleted enhancers obtained from a combination of PacBio long-reads and Illumina short-reads (Chaisson et al. 2018). We found 21 novel enhancers completely deleted in a homozygous fashion in the three individuals sequenced by Chaisson et al. Our model predicted significantly higher LoF-tolerance probability scores for these enhancers than the genome average (Kolmogorov-Smirnov test P-value = 0.010, Figure 4b). This result shows that the scores predicted by our model can help with identification of LoF-tolerant enhancers even in the absence of large numbers of whole-genomes and incomplete maps of genomic deletions generated using Illumina short-reads.

In order to estimate how many LoF-tolerant enhancers we may expect to obtain as more whole-genomes are sequenced, we randomly chose increasing numbers of genomes in sets of 100 from 2,504 whole-genomes and calculated the number of LoF-tolerant enhancers discovered. Our power calculations using this sub-sampling approach show that the number of LoF-tolerant enhancers is likely to increase exponentially as more genomes are sequenced (Figure 4a). However, sequencing all human genomes to find all the LoF-tolerant enhancers is still infeasible even with short-reads sequencing, let alone more expensive and time-consuming long-reads sequencing. Thus, our model can serve as a practical method to predict which enhancers will be more prone to LoF-tolerance and in the interpretation of disease-associated non-coding variants as discussed below.

Predicted LoF-intolerant enhancers and disease risk

In order to evaluate if our model can be informative for the prediction of disease-associated regulatory elements, we extracted a set of disease enhancers from DiseaseEnhancer database (Zhang et al. 2018). In this database, the authors used manual curation to collect the enhancers for which the related genetic variation has been associated with disease phenotypes or important TF binding changes (Zhang et al. 2018). We examined the LoF-tolerance scores predicted by our model for the 90 disease enhancers matched in MegaNet (Methods). We find that the disease-associated enhancers have significantly lower LoF-tolerance probabilities relative to all the enhancers (Kolmogorov-Smirnov test P-value = 1.150e-6), suggesting that our model correctly predicts their intolerance to loss of function (Figure 4b).

We further categorized these enhancers into different disease groups, for example, obesity, skin diseases, neurological disorders, artery diseases, immune disorders, and developmental diseases. We find that skin disease related enhancers have higher LoF-tolerance probability scores (Wilcoxon rank sum test P-value = 0.025, Supplementary Figure 4a), while psychological disorders related enhancers have lower LoF-tolerant scores (average predicted $P_{\text{LoF-tol}}$ = 0.46, Wilcoxon rank sum test P-value = 0.024, Supplementary Figure 4a).

We also inspected a few prominent individual examples related to severe diseases. Previous studies have shown that a single nucleotide mutation in an enhancer regulating *SLC26A4* can cause decreased enhancer activity leading to repression of gene expression (Fuxman Bass et al. 2015), which in turn is associated with Pendred syndrome (Campbell et al. 2001, Tsukamoto et al. 2003). Pendred syndrome is a disorder associated with hearing loss caused by abnormalities of inner ear. SLC26A4 is an anion transporter and its disablement can cause hearing loss and inner ear malformation (Yang et al. 2007, Lazzereschi et al. 2005). This enhancer (Enhancer A, Figure 4c) is predicted to be LoF-intolerant by our model with $P_{\text{LoF-tol.}}$ = 0.41 ($P_{\text{LoF-tol}} < 0.5$), consistent with its loss of function leading to the disease. In contrast, a neighboring enhancer (Enhancer B), which is 1.2 kbp away is predicted to be LoF-tolerant ($P_{\text{LoF-tol.}}$ = 0.94). This result shows that our model can differentiate between LoF-tolerant and LoF-intolerant enhancers even when they regulate the same gene. Closer inspection of these two enhancers reveals that the reason why these two closely located enhancers are predicted to have different LoF-tolerance by our model is that Enhancer A is active in more tissues (spleen and heart) and shows higher sequence conservation (PhastCon score = 0.36) relative to Enhancer B (active in H1 embryonic stem cells with PhastCon score = 0.026, Supplementary Table 2).

In another prominent example of enhancers related to severe diseases, *ZIC3* is a protein-coding gene in the ZIC family of C2H2-type zinc finger proteins, acting as a transcriptional activator in the early stages of determining body left-right asymmetry. Mutations in *ZIC3* have been found in X-linked heterotaxy syndrome and isolated congenital heart disease (CHD) (Gebbia et al. 1997, Ware et al. 2004). Homozygous mutations in *ZIC3* in mice result in 50% embryonic lethality and live born mice exhibit severe congenital heart defects, pulmonary reversal or isomerism (Purandare et al. 2002). Out of 33 enhancers that regulate this gene, 17 are predicted to be LoF-intolerant by our model with average $P_{\text{LoF-tol.}}$ = 0.25. Previous studies have found 8 LoF mutations in coding regions of *ZIC3* related to the heterotaxy, however, they only explained ~1% of the cases (Ware et al. 2004). Therefore, the LoF-intolerant enhancers predicted by our model may provide potential novel susceptibility loci for the study of X-linked heterotaxy and CHD.

Overall, these results suggest that the LoF-tolerance probability scores predicted by our model can provide a powerful reference for disease and clinical studies.

Non-conserved enhancers may be LoF-intolerant

As noted above, although evolutionary conservation is the feature with the highest importance (0.338) in our model, other network features improve the performance of the model and their integration allows us to achieve an AUROC of 96%. In order to further interpret the relationship between network properties and LoF-tolerance, we examined their contribution to disease enhancers. From the disease enhancer set described in the previous section, there are 11/39 enhancers with conservation < 0.065 (median of all enhancer PhastCon scores) (Pollard et al. 2010) yet they are predicted to be LoF-intolerant by our model. One example is an enhancer regulating the gene *MFS1*. Two SNVs (rs3821943, rs4689397) in this enhancer have been associated with type 2 diabetes and shown to decrease the enhancer activity by lowering the expression of *MFS1* through luciferase reporter activity (Stitzel et al. 2010). The susceptibility loci reported in the study locate in one enhancer in our dataset with $P_{\text{LoF-tol}}$ = 0, hence it is predicted to be a highly confident LoF-intolerant candidate. However, the conservation for this enhancer region is low (PhastCon score = 0.0043), which shows that our model can help prioritize and interpret disease variants beyond the use of evolutionary conservation (Supplementary Figure 4b).

Discussion

In this study, we constructed a unified human regulatory network (MegaNet) by integrating tissue-specific enhancer-target networks and gene-gene interactions. To define enhancers that may be tolerant to LoF in the genome, we used deletions from 1000 Genomes Project. We describe the differences between LoF-tolerant and -intolerant enhancers in the MegaNet. We observe that LoF-tolerant enhancers regulate fewer genes and tend to be more tissue-specific. We also find that the genes regulated by LoF-tolerant enhancers tend to be regulated by more enhancers, indicating enhancer redundancy in the network. The catalogue of LoF-tolerant enhancers allowed us to develop a supervised learning method to predict the LoF-tolerance of all enhancers in the human genome using their properties in MegaNet. Independent data sets obtained using long-read sequences and known sets of disease enhancers provide validation for the LoF-tolerance scores predicted by our model.

GWAS have revealed that the majority of the variants associated with complex diseases reside in non-coding regions of the genome (Hindorff et al. 2009, McCarthy and Hirschhorn 2008, Maurano et al. 2012). Moreover, whole exome sequencing could only find the causal variants for ~25-50% of patients (Wortmann et al. 2015, Kremer et al. 2017). It is likely that regions excluded from exome sequencing, namely non-coding regions, harbor the variants explaining many of the remaining unexplained cases (Valente and Bhatia 2018). Major international efforts such as the UK Biobank and TOPMed (NHLBI Trans-Omics for Precision Medicine) aim to use whole-genome sequencing to uncover disease variants (Bycroft et al. 2018, Turnbull et al. 2018, Sarnowski et al. 2018, He et al. 2019, Telenti et al. 2016, Perkins et al. 2018). The LoF-tolerance scores for enhancers provided here can significantly facilitate the interpretation and prioritization of non-coding sequence variants in whole-genome sequencing studies.

Materials and Methods

Obtaining enhancer-gene networks

Enhancer-gene networks in different tissues were obtained from the ENCODE+Roadmap LASSO dataset in Cao et al. (Cao et al. 2017, http://yiplab.cse.cuhk.edu.hk/jeme/). In Cao et al, they collected ChIP-seq data for H3k4me1, H3K27ac, H3K27me3, DNase-seq together with ChromHMM-predicted active enhancers to generate a union set of enhancers. In total, we collected 246,028 unique enhancers regulating 19,170 genes from enhancer-target networks from all tissue types. We grouped 127 Roadmap tissue types by the given sample group into 19 tissue groups and discarded ungrouped cell types.

In order to identify LoF-tolerant enhancers, we first identified all deletions existing in a homozygous state in any one individual in the 1000 Genomes Phase 3 data (Sudmant et al. 2015). We excluded any deletion overlapping coding exon regions and then intersected the remaining deletions with enhancer coordinates to obtain our list of 886 LoF-tolerant enhancers. Only enhancers that are 100% deleted were included.

In order to identify LoF-intolerant enhancers, we started with ultra-conserved elements and retained only those showing consistent reporter gene expression (Bejerano et al. 2004, Visel et al. 2007, Visel et al. 2008, Dickel et al. 2018). We intersected the remaining elements with enhancer coordinates in our dataset, keeping only those with >50% reciprocal overlap. In total, we define 49 LoF-intolerant enhancers.

We compared the length distributions of enhancers and deletions (Supplementary Figure 5). The average length of deletions is much longer than enhancers. Thus, LoF-tolerant enhancers are likely not biased towards shorter enhancers (shorter enhancers are more likely to be completely deleted). To be more stringent, we still excluded the length of enhancers as a feature in the following analysis.

Tissue-specific subnetworks

To distinguish enhancer activity differences between tissues, we extracted tissue-specific networks from the MegaNet. Enhancers in HSC & B-cell and Epithelial tissues exhibit significant differences in tissue-specific network properties between LoF-tolerant and LoF-intolerant enhancers (Wilcoxon rank sum test P-value < 0.05, Supplementary Figure 3b).

Collecting features for the model

Besides the tissue specificity information of enhancers, we also used the gene centralities and gene indispensability scores (Khurana et al. 2013) as measurements for gene priority in the network. In order to only consider the direct interactions between gene pairs, indirect interactions, genetic interaction and regulatory interactions, were excluded from our integrated network. Enhancer-target network features were calculated using Python networkX package (Hagberg, Swart and S Chult 2008). Conservation scores for sequence were obtained from PhastCons (Pollard et al. 2010).

Detailed information about network features is provided in Supplementary Table 1. The enhancer tissue ubiquity (ETU) is the total number of tissues that the enhancer is active in. The enhancer-gene edge tissue Ubiquity (EGTU) is the number of tissues that the enhancer-gene regulation edge is active in. For enhancers that regulate multiple genes, to transform gene features for those regulated genes into an enhancer feature, we took both the average and variance for each gene features and represented it with extension "a" (average) or "v" (variance). For each enhancer, we denote ETU as n, then EGTU is a list of $(e_1, e_2, \dots, e_n)$. The EGTUa will be $\frac{\sum_{i=1}^{n} e_i}{n}$, and the EGTUv is $\frac{\sum_{i=1}^{n}(e_i - \mathrm{EGTUa})^2}{n}$.

Feature selection

To avoid overfitting introduced by features correlated with each other, we calculated the Spearman distance between each feature. We noticed that features for tissue type adipose/epithelial and digestive are strongly correlated with each other, thus only one of them (adipose) was kept for further model building. In addition, tissue type myosat and mesench are mixed with other tissue clusters, so we eliminated them from the final tissue set. In the end, there are in total 15 tissue types considered and 62 features overall.

Model building and testing

The model was built using tools from Python Scikit-learn package (Pedregosa et al. 2011). Random and grid searches were performed to find the best parameters for the random forest classifier. 10-fold cross validation was performed to evaluate the performance of the model. We selected 50 LoF-tolerant and 49 LoF-intolerant enhancers to train the classifier. To avoid overfitting, we repeatedly sampled through all LoF-tolerant enhancers 50 times. The mean AUROC is 0.9633 +/- 0.0002. Due to the small sample size of LoF-intolerant enhancers, we also randomly chose 50 enhancers from neither the LoF-tolerant nor intolerant set as "LoF-intolerant" to test overfitting of the model. We performed the same parameter searching and cross validation repeatedly 50 times and obtained mean AUROC of 0.6154 +/- 0.0697, indicating that the small sample size for LoF-intolerant enhancers did not lead to overfitting.

We applied the model on all other enhancers in the network and predicted their probability to be LoF-tolerant as their LoF-tolerant scores. The predicted LoF-tolerant probabilities are the mean predicted class probabilities of the trees in the forest (Pedregosa et al. 2011). Among 245,361 enhancers tested, 194,812 ($P_{\text{LoF-tol}}$ >= 0.5) are predicted to be LoF-tolerant enhancers, while 50,281 to be LoF-intolerant ($P_{\text{LoF-tol}}$ < 0.5) and 5,677 are predicted to be highly confident LoF-tolerant ($P_{\text{LoF-tol.}}$ > 0.95), while 75 are predicted to be highly confident LoF-intolerant candidates with very low LoF-tolerance probability ($P_{\text{LoF-tol.}}$ < 0.05).

Validation

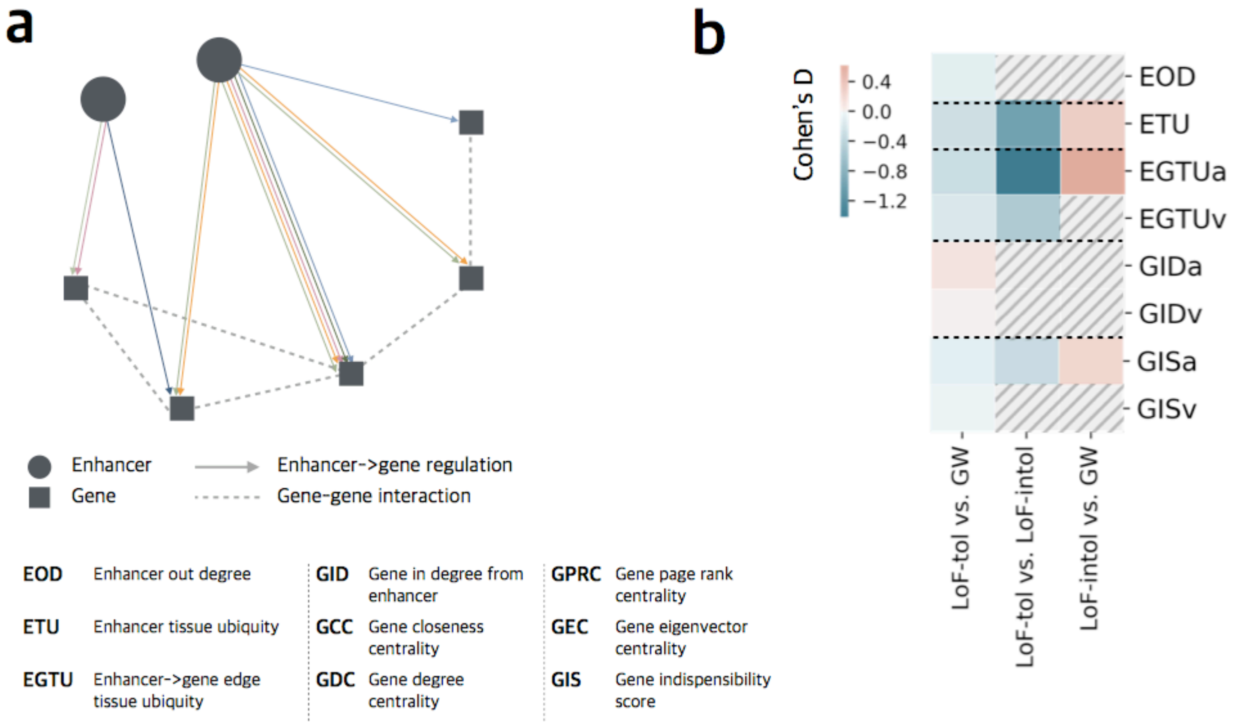
To further validate our observation that there are additional LoF-tolerant enhancers in human genomes, we obtained novel deletions to identify LoF-tolerant enhancers. Those novel deletions were from the 1000 Genomes structural variation consortium where they used integrated structural variation calling methods including both Illumina short reads and PacBio long reads sequencing for three individuals from 1000 Genomes trio studies (Chaisson et al. 2018). In total, we used 12,939 deletions from the PacBio structural variants set that were present in the three children from the trio family and intersected them with 1000 Genomes Phase 3 deletions. There are 11,118 novel deletions with less than 80% overlap with the 1000 Genomes Phase 3 deletions. Out of those novel deletions, 21 of them can delete enhancers completely from our enhancer set.

Disease enhancers

Disease enhancers were collected from Zhang et al. (Zhang et al. 2018). We intersected our enhancers with the 1,059 disease enhancers which defined in Zhang et al., if no overlap found then take the closest neighbored enhancer. After this, keep only the disease enhancers that its target gene from the DiseaseEnhancer matches the enhancer-gene regulation from our dataset. To further filter out the disease enhancers related to somatic variants, we excluded enhancers associated with cancer. In the end, we collected 90 enhancers in our dataset with disease associations.

Figures and Supplementary Tables

Figure 1. MegaNet features. a) Schema of the MegaNet, circle and square represent nodes for enhancers and genes, respectively, and colored directed arrows are enhancer->gene regulation edges. Different colors represent the interactions active in different tissues. Dashed lines represent the gene-gene interactions. b) Only significant comparisons (P-value < 0.05) are shown in color, non-significant ones are marked by dashed lines. Color scale represents Cohen's D for effective size, positive values stand for higher average while negative values stand for lower average. LoF-tol, LoF-intol and GW represent LoF-tolerant, LoF-intolerant and genome-wide respectively. 'a' and 'v' stand for the average and variance for the corresponding features in a).



EOD	Enhancer out degree	**GID**	Gene in degree from enhancer	**GPRC**	Gene page rank centrality
ETU	Enhancer tissue ubiquity	**GCC**	Gene closeness centrality	**GEC**	Gene eigenvector centrality
EGTU	Enhancer->gene edge tissue ubiquity	**GDC**	Gene degree centrality	**GIS**	Gene indispensibility score

Figure 2. Tissue-specific enhancers. Horizontal bars show the percentage of LoF-tolerant and LoF-intolerant enhancers in each tissue type. Vertical bar plots show their odds ratios for enrichment/depletion in each tissue compared to all LoF-intolerant/-tolerant enhancers in all tissues (asterisks mark the statistical significance using Fisher's exact test).

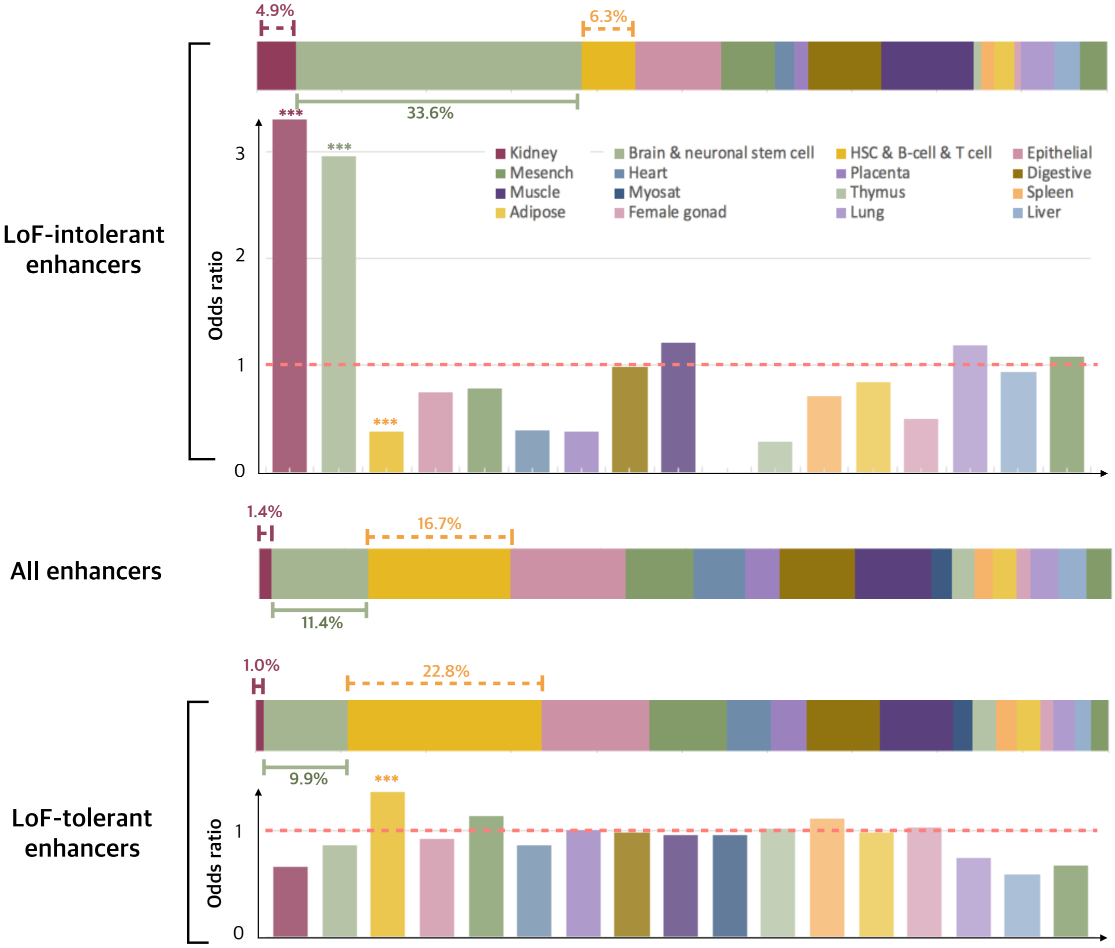


Figure 3. Model performance. a) 10-fold cross validation AUROC of a randomly selected set of 50 positives for the random forest classification model. b) Feature importance for the classification model. X-axis shows the features. c) Enhancer tissue ubiquity (ETU) and average gene indispensability scores (GISa) for LoF-tolerant and -intolerant enhancer candidates.

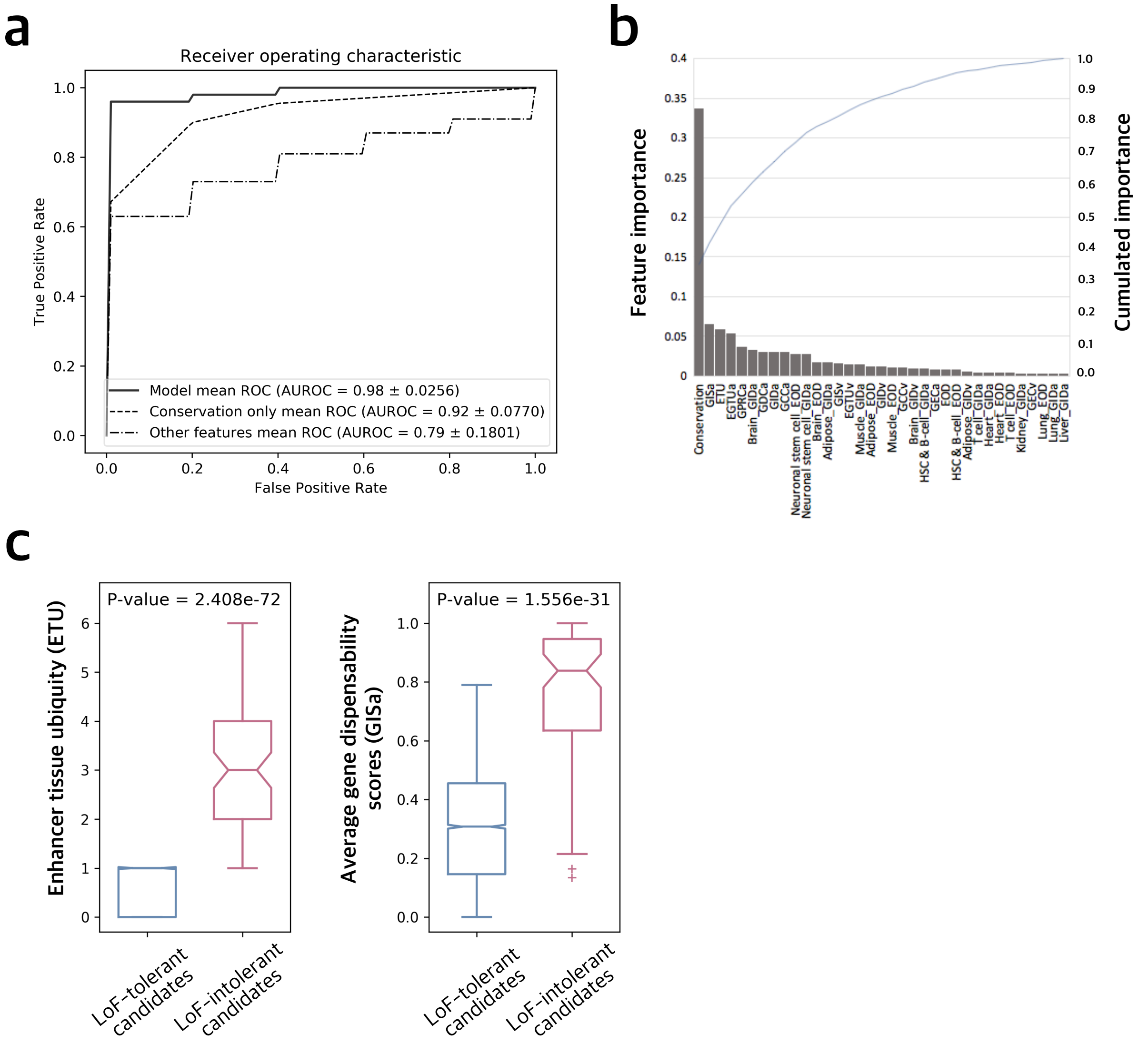
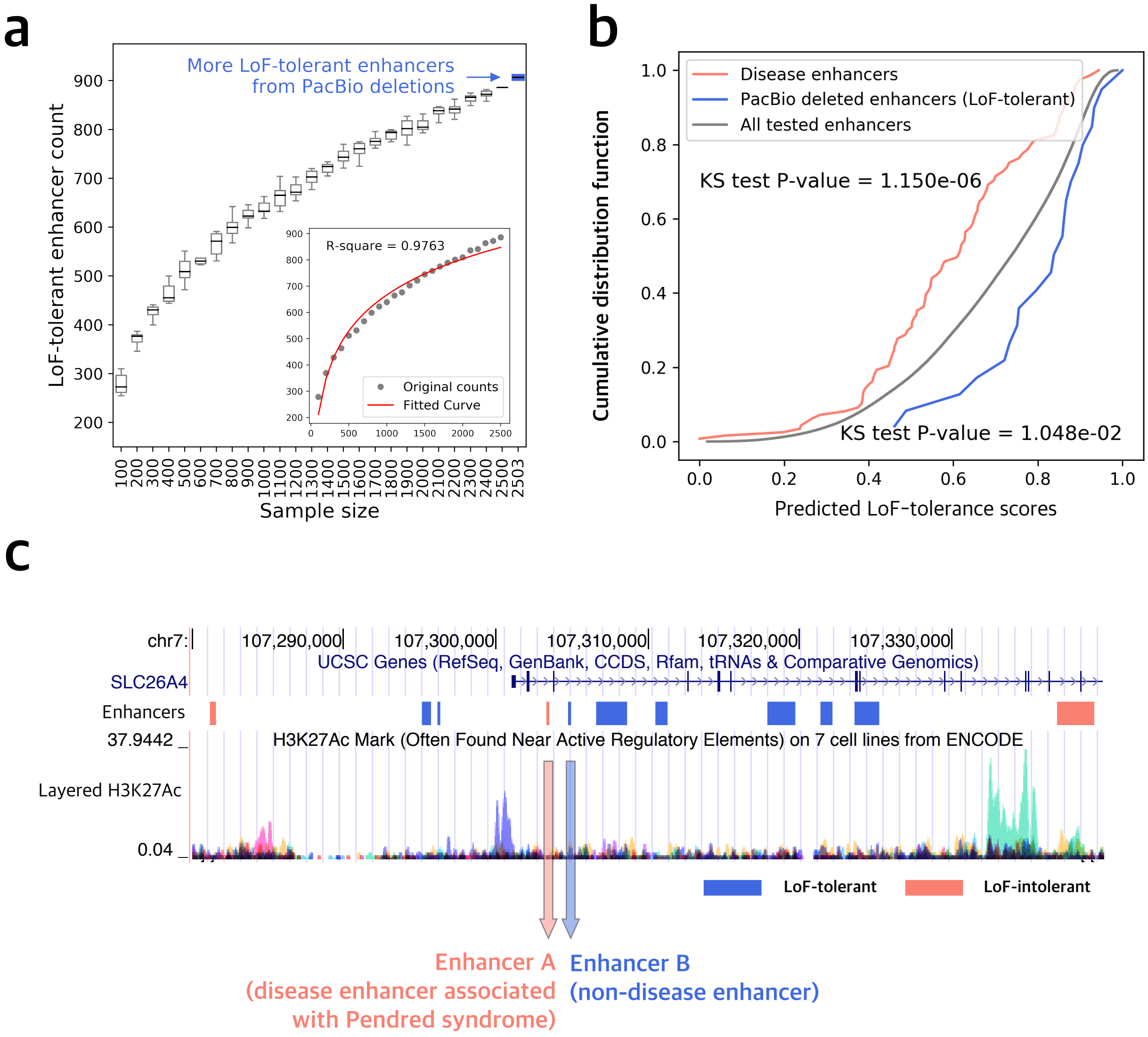
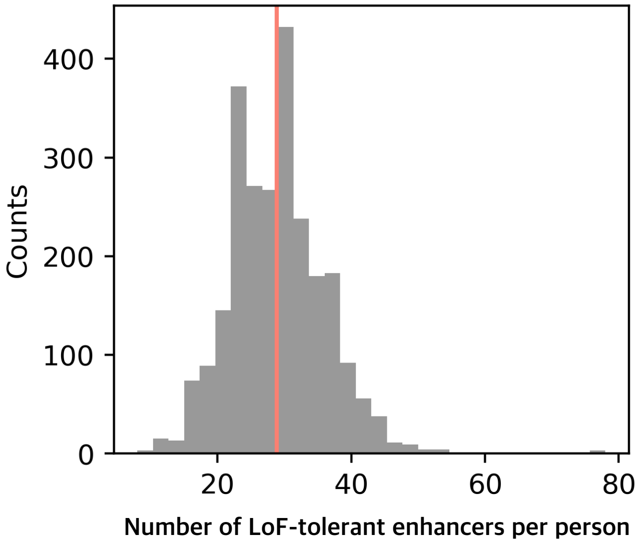


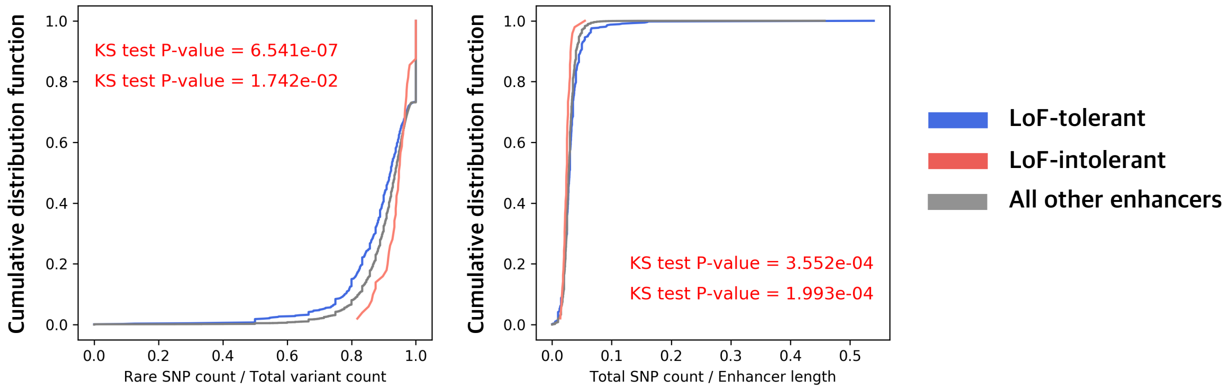
Figure 4. Validation using PacBio SVs and disease enhancers. a) Number of observed LoF-tolerant enhancers with increasing sample size. On the x-axis, 2503 includes the LoF-tolerant enhancers observed from 3 additional individuals sequenced using PacBio. b) Cumulative distribution function for LoF-tolerant scores for disease enhancers (red), all tested enhancers (grey), PacBio deleted enhancers (blue). KS-test P-values are between disease enhancers vs. all tested and PacBio enhancers vs. all tested. c) Genome region of *SLC26A4* and part of the enhancers regulating it. Blue denotes the predicted LoF-tolerant enhancers, while red is for predicted LoF-intolerant enhancers. Layered H3K27Ac is the modification of histone H3 lysine 27 acetylation which is associated with active enhancers.



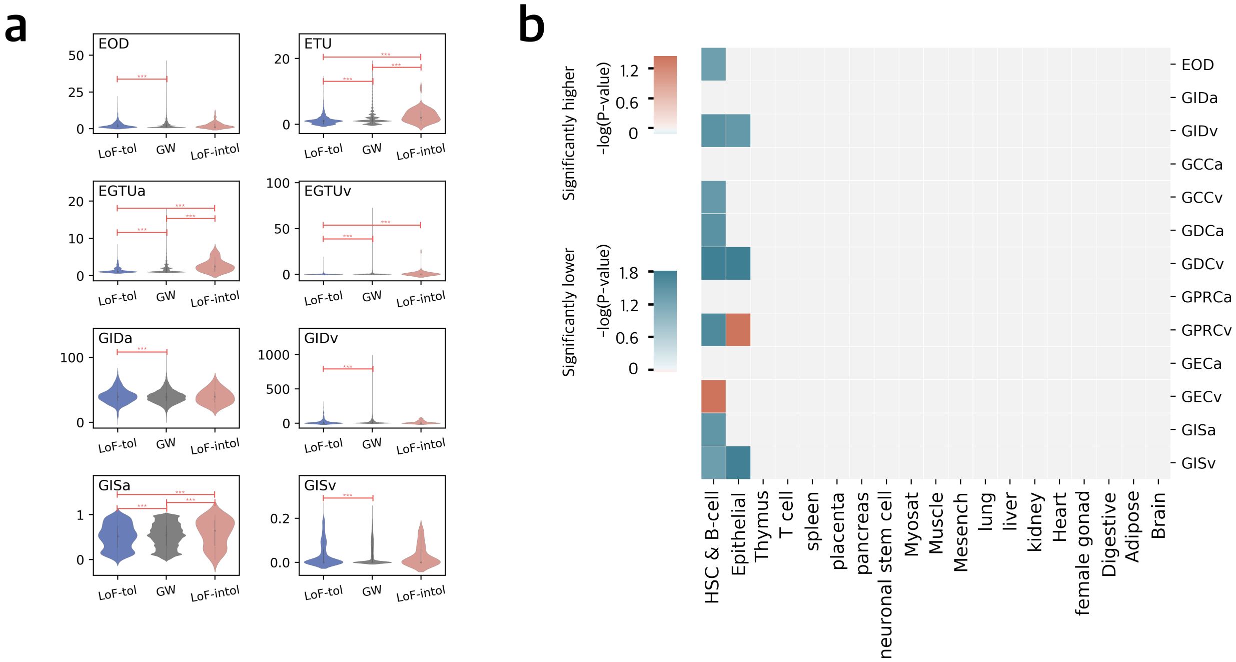
Supplementary Figure 1. Number of LoF-tolerant enhancers per individual from 2,504 genomes. Each individual has on average 28 enhancers (red vertical line) completely and homozygously deleted in the genome.



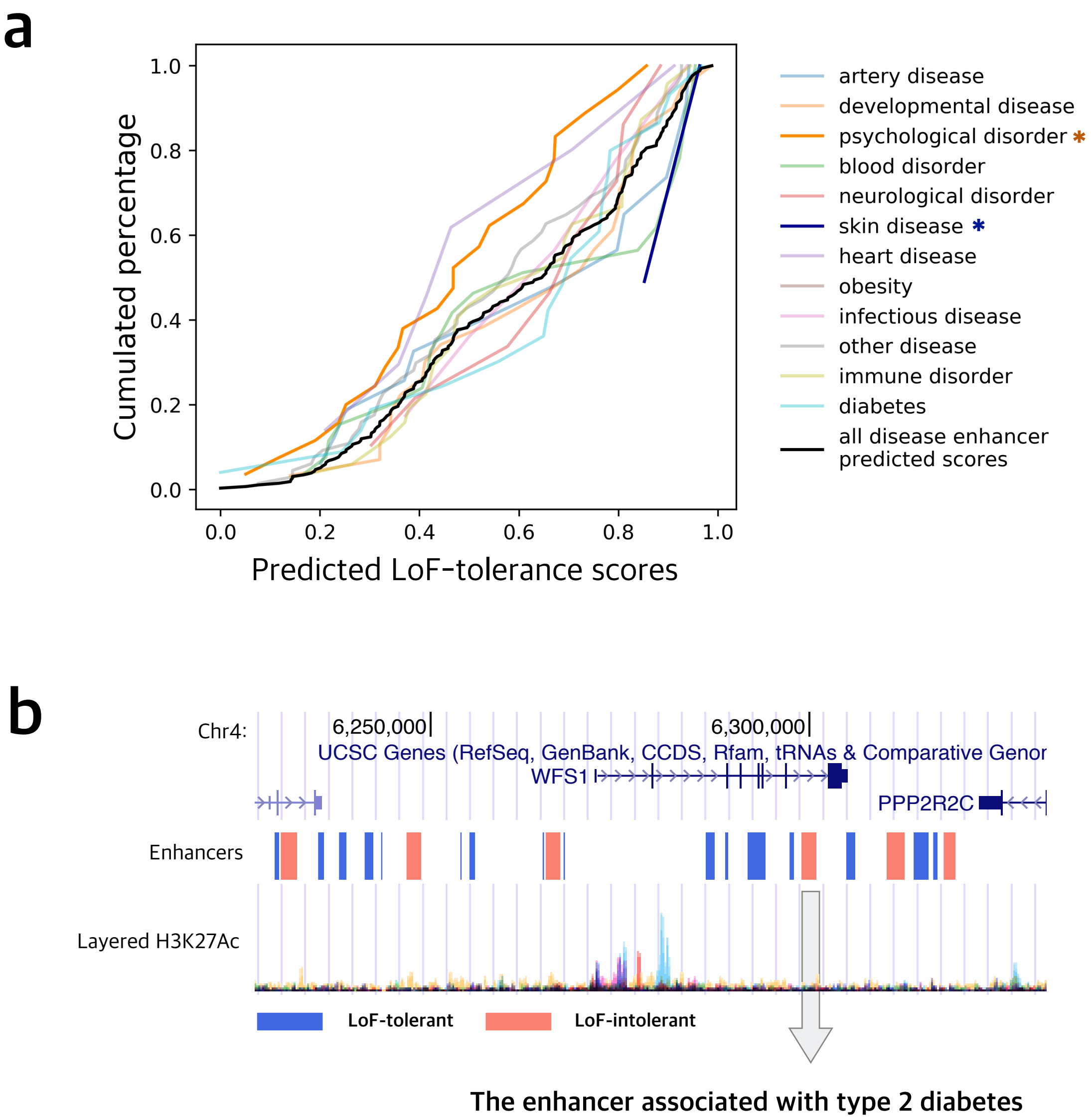
Supplementary Figure 2. Comparison of enrichment of rare variants and all polymorphisms between LoF-tolerant and -intolerant enhancers and all other enhancers (genome-wide, GW). Upper P-value is for LoF-tolerant vs. GW, while lower P-value is for LoF-intolerant vs. GW. The P-values were calculated by Kolmogorov-Smirnov test (KS test).



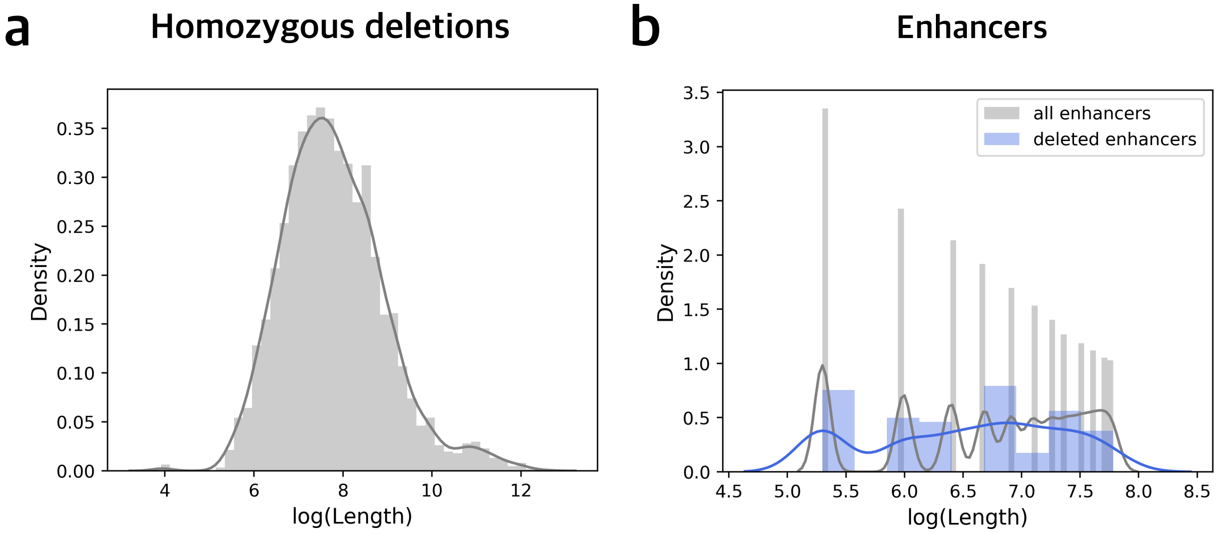
Supplementary Figure 3. Network features in the MegaNet and in tissues-specific networks. a) network features in the MegaNet, significant comparisons are marked by asterisks. b) Each column represents a tissue-specific network comparison between LoF-tolerant vs. LoF-intolerant enhancers (P-values were computed using Wilcoxon rank sum test).



Supplementary Figure 4. Disease enhancers. a) Predicted LoF-tolerance scores for disease enhancers by disease types. Y-axis is the cumulated percentage of enhancers for the corresponding LoF-tolerance scores on x-axis. Disease types are colored as shown, significant ones (P-value < 0.05) are marked by asterisks. b) Genome region of *WFS1* and part of the enhancers regulating it. Blue denotes the predicted LoF-tolerant enhancers, while red is for predicted LoF-intolerant enhancers. Layered H3K27Ac is the modification of histone H3 lysine 27 acetylation which is associated with active enhancers.



Supplementary Figure 5. Length distribution of homozygous deletions, deleted enhancers and all enhancers.



Supplementary Table 1. Summary of network features.

Supplementary Table 2. Categories of ENDODE and Roadmap tissues

Supplementary Table 3. Predicted LoF-tolerance scores for all enhancers in this study.

References

Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal, L. A. Pennacchio & E. M. Rubin (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biol.,* 5**,** e234.

Albuisson, J., B. Isidor, M. Giraud, O. Pichon, T. Marsaud, A. David, C. Le Caignec & S. Bezieau (2011) Identification of two novel mutations in Shh long-range regulator associated with familial pre-axial polydactyly. *Clin Genet,* 79**,** 371-7.

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. R. Forrest, P. Carninci, M. Rehli & A. Sandelin (2014) An atlas of active enhancers across human cell types and tissues. *Nature,* 507**,** 455-461.

Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis & G. P. Consortium (2015) A global reference for human genetic variation. *Nature,* 526**,** 68-74.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick & D. Haussler (2004) Ultraconserved elements in the human genome. *Science,* 304**,** 1321-1325.

Breiman, L. 1984. *Classification and Regression Trees*. Chapman & Hall.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly & J. Marchini (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature,* 562**,** 203-209.

Campbell, C., R. A. Cucci, S. Prasad, G. E. Green, J. B. Edeal, C. E. Galer, L. P. Karniski, V. C. Sheffield & R. J. Smith (2001) Pendred syndrome, DFNB4, and PDS/SLC26A4 identification of eight novel mutations and possible genotype-phenotype correlations. *Hum. Mutat.,* 17**,** 403-411.

Cao, Q., C. Anyansi, X. Hu, L. Xu, L. Xiong, W. Tang, M. T. S. Mok, C. Cheng, X. Fan, M. Gerstein, A. S. L. Cheng & K. Y. Yip (2017) Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.,* 49**,** 1428-1436.

Chaisson, M. J. P., J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach & E. E. Eichler (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature,* 517**,** 608-611.

Chaisson, M. J. P., A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. Hastie, D. Antaki, P. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C.-S. Chin, Z. Chong, N. T. Chuang, C. C. Lambert, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, D. Gorkin, M. Gujral, V. Guryev, W. H. Heaton, J. Korlach, S. Kumar, J. Y. Kwon, J. E. Lee, J. Lee, W.-P. Lee, S. P. Lee, S. Li, P. Marks, K. Viaud-Martinez, S. Meiers, K. M. Munson, F. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stutz, D. C. J. Spierings, A. Ward, A. E. Welch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, E. Lowy, S.

Yakneen, S. McCarroll, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P.-Y. Kwok, P. M. Lansdorp, G. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. Talkowski, R. E. Mills, T. Marschall, J. O. Korbel, E. E. Eichler & C. Lee (2018) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*.

Chin, C.-S., D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner & J. Korlach (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods,* 10**,** 563-569.

Consortium, E. P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

Cowper-Sal lari, R., X. Zhang, J. B. Wright, S. D. Bailey, M. D. Cole, J. Eeckhoute, J. H. Moore & M. Lupien (2012) Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet,* 44**,** 1191-8.

Davis, C. A., B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka & J. M. Cherry (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res,* 46**,** D794-D801.

Dickel, D. E., A. R. Ypsilanti, R. Pla, Y. Zhu, I. Barozzi, B. J. Mannion, Y. S. Khin, Y. Fukuda-Yuzawa, I. Plajzer-Frick, C. S. Pickle, E. A. Lee, A. N. Harrington, Q. T. Pham, T. H. Garvin, M. Kato, M. Osterwalder, J. A. Akiyama, V. Afzal, J. L. R. Rubenstein, L. A. Pennacchio & A. Visel (2018) Ultraconserved Enhancers Are Required for Normal Development. *Cell,* 172**,** 491-499.e15.

Fuxman Bass, J. I., N. Sahni, S. Shrestha, A. Garcia-Gonzalez, A. Mori, N. Bhat, S. Yi, D. E. Hill, M. Vidal & A. J. M. Walhout (2015) Human gene-centered transcription factor networks for enhancers and disease variants. *Cell,* 161**,** 661-673.

Gebbia, M., G. B. Ferrero, G. Pilia, M. T. Bassi, A. Aylsworth, M. Penman-Splitt, L. M. Bird, J. S. Bamforth, J. Burn, D. Schlessinger, D. L. Nelson & B. Casey (1997) X-linked situs abnormalities result from mutations in ZIC3. *Nat. Genet.,* 17**,** 305-308.

Genomes Project, C., G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles & G. A. McVean (2010) A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-1073.

Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean & G. R. Abecasis (2015) A global reference for human genetic variation. *Nature,* 526**,** 68-74.

Genovese, G., M. Fromer, E. A. Stahl, D. M. Ruderfer, K. Chambert, M. Landén, J. L. Moran, S. M. Purcell, P. Sklar, P. F. Sullivan, C. M. Hultman & S. A. McCarroll (2016) Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.,* 19**,** 1433-1441.

Ghiasvand, N. M., D. D. Rudolph, M. Mashayekhi, J. A. Brzezinski, D. Goldman & T. Glaser (2011) Deletion of a remote enhancer near ATOH7 disrupts retinal neurogenesis, causing NCRNA disease. *Nat Neurosci,* 14**,** 578-86.

Gilissen, C., J. Y. Hehir-Kwa, D. T. Thung, M. van de Vorst, B. W. M. van Bon, M. H. Willemsen, M. Kwint, I. M. Janssen, A. Hoischen, A. Schenck, R. Leach, R. Klein, R. Tearle, T. Bo, R. Pfundt, H. G. Yntema, B. B. A. de Vries, T. Kleefstra, H. G. Brunner, L. E. L. M. Vissers & J. A. Veltman (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature,* 511**,** 344-347.

Hagberg, A., P. Swart & D. S Chult. 2008. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

He, B., C. Chen, L. Teng & K. Tan (2014) Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U. S. A.,* 111**,** E2191-9.

He, K. Y., X. Li, T. N. Kelly, J. Liang, B. E. Cade, T. L. Assimes, L. C. Becker, A. L. Beitelshees, A. P. Bress, Y. C. Chang, Y. I. Chen, P. S. de Vries, E. R. Fox, N. Franceschini, A. Furniss, Y. Gao, X. Guo, J. Haessler, S. J. Hwang, M. R. Irvin, R. R. Kalyani, C. T. Liu, C. Liu, L. W. Martin, M. E. Montasser, P. M. Muntner, S. Mwasongwe, W. Palmas, A. P. Reiner, D. Shimbo, J. A. Smith, B. M. Snively, L. R. Yanek, E. Boerwinkle, A. Correa, L. A. Cupples, J. He, S. L. R. Kardia, C. Kooperberg, R. A. Mathias, B. D. Mitchell, B. M. Psaty, R. S. Vasan, D. C. Rao, S. S. Rich, J. I. Rotter, J. G. Wilson, A. Chakravarti, A. C. Morrison, D. Levy, D. K. Arnett, S. Redline, X. Zhu & T. O. P. M. B. P. W. G. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2019) Leveraging linkage evidence to identify low-frequency and rare variants on 16p13 associated with blood pressure using TOPMed whole genome sequencing data. *Hum Genet,* 138**,** 199-210.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins & T. A. Manolio (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.,* 106**,** 9362-9367.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe & M. Hirakawa (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.,* 38**,** D355-60.

Kapoor, A., R. B. Sekar, N. F. Hansen, K. Fox-Talbot, M. Morley, V. Pihur, S. Chatterjee, J. Brandimarto, C. S. Moravec, S. L. Pulit, Q. T. I.-I. G. Consortium, A. Pfeufer, J. Mullikin, M. Ross, E. D. Green, D. Bentley, C. Newton-Cheh, E. Boerwinkle, G. F. Tomaselli, T. P. Cappola, D. E. Arking, M. K. Halushka & A. Chakravarti (2014) An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval. *Am. J. Hum. Genet.,* 94**,** 854-869.

Kathiresan, S. & D. Srivastava (2012) Genetics of human cardiovascular disease. *Cell,* 148**,** 1242-1257.

Katzman, S., A. D. Kern, G. Bejerano, G. Fewell, L. Fulton, R. K. Wilson, S. R. Salama & D. Haussler (2007) Human genome ultraconserved elements are ultraselected. *Science,* 317**,** 915.

Khurana, E., Y. Fu, J. Chen & M. Gerstein (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.,* 9**,** e1002886.

Korcsmáros, T., I. J. Farkas, M. S. Szalay, P. Rovó, D. Fazekas, Z. Spiró, C. Böde, K. Lenti, T. Vellai & P. Csermely (2010) Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics,* 26**,** 2042-2050.

Kremer, L. S., D. M. Bader, C. Mertes, R. Kopajtich, G. Pichler, A. Iuso, T. B. Haack, E. Graf, T. Schwarzmayr, C. Terrile, E. Koňaříková, B. Repp, G. Kastenmüller, J. Adamski, P. Lichtner, C. Leonhardt, B. Funalot, A. Donati, V. Tiranti, A. Lombes, C. Jardel, D. Gläser, R. W. Taylor, D. Ghezzi, J. A. Mayr, A. Rötig, P. Freisinger, F. Distelmaier, T. M. Strom, T. Meitinger, J. Gagneur & H. Prokisch (2017) Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun,* 8**,** 15824.

Kronenberg, Z. N., I. T. Fiddes, D. Gordon, S. Murali, S. Cantsilieris, O. S. Meyerson, J. G. Underwood, B. J. Nelson, M. J. P. Chaisson, M. L. Dougherty, K. M. Munson, A. R. Hastie, M. Diekhans, F. Hormozdiari, N. Lorusso, K. Hoekzema, R. Qiu, K. Clark, A. Raja, A. E. Welch, M. Sorensen, C. Baker, R. S. Fulton, J. Armstrong, T. A. Graves-Lindsay, A. M. Denli, E. R. Hoppe, P. Hsieh, C. M. Hill, A. W. C. Pang, J. Lee, E. T. Lam, S. K. Dutcher, F. H. Gage, W. C. Warren, J. Shendure, D. Haussler, V. A. Schneider, H. Cao, M. Ventura, R. K. Wilson, B. Paten, A. Pollen & E. E. Eichler (2018) High-resolution comparative analysis of great ape genomes. *Science,* 360.

Lazzereschi, D., F. Nardi, A. Turco, L. Ottini, C. D'Amico, R. Mariani-Costantini, A. Gulino & A. Coppa (2005) A complex pattern of mutations and abnormal splicing of Smad4 is present in thyroid tumours. *Oncogene,* 24**,** 5344-5354.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur & E. A. Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-91.

Lin, J., Z. Xie, H. Zhu & J. Qian (2010) Understanding protein phosphorylation on a systems level. *Brief. Funct. Genomics,* 9**,** 32-42.

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, C. Genomes Project, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein & C. Tyler-Smith (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science,* 335**,** 823-828.

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham & H. Parkinson (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.,* 45**,** D896-D901.

Macneil, L. T. & A. J. M. Walhout (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res.,* 21**,** 645-657.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul & J. A. Stamatoyannopoulos (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science,* 337**,** 1190-1195.

McCarthy, M. I. & J. N. Hirschhorn (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.,* 17**,** R156-65.

Ng, P. C., S. Levy, J. Huang, T. B. Stockwell, B. P. Walenz, K. Li, N. Axelrod, D. A. Busam, R. L. Strausberg & J. C. Venter (2008) Genetic variation in an individual human exome. *PLoS Genet.,* 4**,** e1000160.

Osterwalder, M., I. Barozzi, V. Tissières, Y. Fukuda-Yuzawa, B. J. Mannion, S. Y. Afzal, E. A. Lee, Y. Zhu, I. Plajzer-Frick, C. S. Pickle, M. Kato, T. H. Garvin, Q. T. Pham, A. N. Harrington, J. A. Akiyama, V. Afzal, J. Lopez-Rios, D. E. Dickel, A. Visel & L. A. Pennacchio (2018) Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature,* 554**,** 239-243.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & É. Duchesnay (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.,* 12**,** 2825-2830.

Pelak, K., K. V. Shianna, D. Ge, J. M. Maia, M. Zhu, J. P. Smith, E. T. Cirulli, J. Fellay, S. P. Dickson, C. E. Gumbs, E. L. Heinzen, A. C. Need, E. K. Ruzzo, A. Singh, C. R. Campbell, L. K. Hong, K. A. Lornsen, A. M. McKenzie, N. L. M. Sobreira, J. E. Hoover-Fong, J. D. Milner, R. Ottman, B. F. Haynes, J. J. Goedert & D. B. Goldstein (2010) The characterization of twenty sequenced human genomes. *PLoS Genet.,* 6**,** e1001111.

Perkins, B. A., C. T. Caskey, P. Brar, E. Dec, D. S. Karow, A. M. Kahn, Y. C. Hou, N. Shah, D. Boeldt, E. Coughlin, G. Hands, V. Lavrenko, J. Yu, A. Procko, J. Appis, A. M. Dale, L. Guo, T. J. Jönsson, B. M. Wittmann, I. Bartha, S. Ramakrishnan, A. Bernal, J. B. Brewer, S. Brewerton, W. H. Biggs, Y. Turpaz & J. C. Venter (2018) Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proc Natl Acad Sci U S A,* 115**,** 3686-3691.

Perry, M. W., A. N. Boettiger, J. P. Bothma & M. Levine (2010) Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.,* 20**,** 1562-1567.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom & A. Siepel (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.,* 20**,** 110-121.

Purandare, S. M., S. M. Ware, K. M. Kwan, M. Gebbia, M. T. Bassi, J. M. Deng, H. Vogel, R. R. Behringer, J. W. Belmont & B. Casey (2002) A complex syndrome of left-right axis, central nervous system and axial skeleton defects in Zic3 mutant mice. *Development,* 129**,** 2293-2302.

Roadmap Epigenomics, C., A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang & M. Kellis (2015) Integrative analysis of 111 reference human epigenomes. *Nature,* 518**,** 317-330.

Roy, S., A. F. Siahpirani, D. Chasman, S. Knaack, F. Ay, R. Stewart, M. Wilson & R. Sridharan (2016) A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.,* 44**,** 1977-1978.

Sarnowski, C., C. L. Satizabal, C. DeCarli, A. N. Pitsillides, L. A. Cupples, R. S. Vasan, J. G. Wilson, J. C. Bis, M. Fornage, A. S. Beiser, A. L. DeStefano, J. Dupuis, S. Seshadri, N. T.-O. f. P. M. T. Consortium & T. N. W. Group (2018) Whole genome sequence analyses of brain imaging measures in the Framingham Study. *Neurology,* 90**,** e188-e196.

Sekiya, T. & K. S. Zaret (2007) Repression by Groucho/TLE/Grg proteins: genomic site recruitment generates compacted chromatin in vitro and impairs activator binding in vivo. *Mol Cell,* 28**,** 291-303.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J.

Kent, W. Miller & D. Haussler (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.,* 15**,** 1034-1050.

Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz & M. Tyers (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.,* 34**,** D535-9.

Stitzel, M. L., P. Sethupathy, D. S. Pearson, P. S. Chines, L. Song, M. R. Erdos, R. Welch, S. C. Parker, A. P. Boyle, L. J. Scott, E. H. Margulies, M. Boehnke, T. S. Furey, G. E. Crawford, F. S. Collins & N. C. S. Program (2010) Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab,* 12**,** 443-55.

Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, C. Genomes Project, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler & J. O. Korbel (2015) An integrated map of structural variation in 2,504 human genomes. *Nature,* 526**,** 75-81.

Telenti, A., L. C. T. Pierce, W. H. Biggs, J. di Iulio, E. H. M. Wong, M. M. Fabani, E. F. Kirkness, A. Moustafa, N. Shah, C. Xie, S. C. Brewerton, N. Bulsara, C. Garner, G. Metzker, E. Sandoval, B. A. Perkins, F. J. Och, Y. Turpaz & J. C. Venter (2016) Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U. S. A.,* 113**,** 11901-11906.

Tg and Hdl Working Group of the Exome Sequencing Project, N. H. L. a. B. I., J. Crosby, G. M. Peloso, P. L. Auer, D. R. Crosslin, N. O. Stitziel, L. A. Lange, Y. Lu, Z.-Z. Tang, H. Zhang, G. Hindy, N. Masca, K. Stirrups, S. Kanoni, R. Do, G. Jun, Y. Hu, H. M. Kang, C. Xue, A. Goel, M. Farrall, S. Duga, P. A. Merlini, R. Asselta, D. Girelli, O. Olivieri, N. Martinelli, W. Yin, D. Reilly, E. Speliotes, C. S. Fox, K. Hveem, O. L. Holmen, M. Nikpay, D. N. Farlow, T. L. Assimes, N. Franceschini, J. Robinson, K. E. North, L. W. Martin, M. DePristo, N. Gupta, S. A. Escher, J.-H. Jansson, N. Van Zuydam, C. N. A. Palmer, N. Wareham, W. Koch, T. Meitinger, A. Peters, W. Lieb, R. Erbel, I. R. Konig, J. Kruppa, F. Degenhardt, O. Gottesman, E. P. Bottinger, C. J. O'Donnell, B. M. Psaty, C. M. Ballantyne, G. Abecasis, J. M. Ordovas, O. Melander, H. Watkins, M. Orho-Melander, D. Ardissino, R. J. F. Loos, R. McPherson, C. J. Willer, J. Erdmann, A. S. Hall, N. J. Samani, P. Deloukas, H. Schunkert, J. G. Wilson, C. Kooperberg, S. S. Rich, R. P. Tracy, D.-Y. Lin, D. Altshuler, S. Gabriel, D. A. Nickerson, G. P. Jarvik, L. A. Cupples, A. P. Reiner, E. Boerwinkle & S. Kathiresan (2014) Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.,* 371**,** 22-31.

Trynka, G., C. Sandor, B. Han, H. Xu, B. E. Stranger, X. S. Liu & S. Raychaudhuri (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.,* 45**,** 124-130.

Tsukamoto, K., H. Suzuki, D. Harada, A. Namba, S. Abe & S.-I. Usami (2003) Distribution and frequencies of PDS (SLC26A4) mutations in Pendred syndrome and nonsyndromic hearing loss associated with enlarged vestibular aqueduct: a unique spectrum of mutations in Japanese. *Eur. J. Hum. Genet.,* 11**,** 916-922.

Turnbull, C., R. H. Scott, E. Thomas, L. Jones, N. Murugaesu, F. B. Pretty, D. Halai, E. Baple, C. Craig, A. Hamblin, S. Henderson, C. Patch, A. O'Neill, Devereau, K. Smith, A. R. Martin, A. Sosinsky, E. M. McDonagh, R. Sultana, M. Mueller, D. Smedley, A. Toms, L.

Dinh, T. Fowler, M. Bale, T. Hubbard, A. Rendon, S. Hill, M. J. Caulfield & G. Project (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ,* 361**,** k1687.

Valente, E. M. & K. P. Bhatia (2018) Solving Mendelian Mysteries: The Non-coding Genome May Hold the Key. *Cell,* 172**,** 889-891.

Visel, A., S. Minovitsky, I. Dubchak & L. A. Pennacchio (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.,* 35**,** D88-D92.

Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin & L. A. Pennacchio (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.,* 40**,** 158-160.

Wang, Z., Q. Zhang, W. Zhang, J. R. Lin, Y. Cai, J. Mitra & Z. D. Zhang (2018) HEDD: Human Enhancer Disease Database. *Nucleic Acids Res,* 46**,** D113-D120.

Ware, S. M., J. Peng, L. Zhu, S. Fernbach, S. Colicos, B. Casey, J. Towbin & J. W. Belmont (2004) Identification and functional analysis of ZIC3 mutations in heterotaxy and related congenital heart defects. *Am. J. Hum. Genet.,* 74**,** 93-105.

Weedon, M. N., I. Cebola, A. M. Patch, S. E. Flanagan, E. De Franco, R. Caswell, S. A. Rodríguez-Seguí, C. Shaw-Smith, C. H. Cho, H. L. Allen, J. A. Houghton, C. L. Roth, R. Chen, K. Hussain, P. Marsh, L. Vallier, A. Murray, S. Ellard, J. Ferrer, A. T. Hattersley & I. P. A. Consortium (2014) Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet,* 46**,** 61-64.

Whalen, S., R. M. Truty & K. S. Pollard (2016) Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.,* 48**,** 488-496.

Wortmann, S. B., D. A. Koolen, J. A. Smeitink, L. van den Heuvel & R. J. Rodenburg (2015) Whole exome sequencing of suspected mitochondrial patients in clinical practice. *J Inherit Metab Dis,* 38**,** 437-43.

Wunderlich, Z., M. D. Bragdon, B. J. Vincent, J. A. White, J. Estrada & A. H. DePace (2015) Krüppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers. *Cell Rep,* 12**,** 1740-7.

Yang, T., H. Vidarsson, S. Rodrigo-Blomqvist, S. S. Rosengren, S. Enerback & R. J. H. Smith (2007) Transcriptional control of SLC26A4 is involved in Pendred syndrome and nonsyndromic enlargement of vestibular aqueduct (DFNB4). *Am. J. Hum. Genet.,* 80**,** 1055-1063.

Yip, K. Y., C. Cheng, N. Bhardwaj, J. B. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder & M. Gerstein (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.,* 13**,** R48.

Yu, Y., Y. Lin, Y. Takasaki, C. Wang, H. Kimura, J. Xing, K. Ishizuka, M. Toyama, I. Kushima, D. Mori, Y. Arioka, Y. Uno, T. Shiino, Y. Nakamura, T. Okada, M. Morikawa, M. Ikeda, N. Iwata, Y. Okahisa, M. Takaki, S. Sakamoto, T. Someya, J. Egawa, M. Usami, M. Kodaira, A. Yoshimi, T. Oya-Ito, B. Aleksic, K. Ohno & N. Ozaki (2018) Rare loss of function mutations in N-methyl-D-aspartate glutamate receptors and their contributions to schizophrenia susceptibility. *Transl. Psychiatry,* 8**,** 12.

Zhang, G., J. Shi, S. Zhu, Y. Lan, L. Xu, H. Yuan, G. Liao, X. Liu, Y. Zhang, Y. Xiao & X. Li (2018) DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Res.,* 46**,** D78-D84.

Zhu, Y., Z. Chen, K. Zhang, M. Wang, D. Medovoy, J. W. Whitaker, B. Ding, N. Li, L. Zheng & W. Wang (2016) Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.,* 7**,** 10812.