ARTICLE

Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network

Souhrid Mukherjee,[1] Joy D. Cogan,[2] John H. Newman,[3] John A. Phillips III,[2] Rizwan Hamid,[2] Undiagnosed Diseases Network, Jens Meiler,[4,5,6,7,8,9,10,*] and John A. Capra[1,6,7,11,12,*]

Summary

Rare diseases affect millions of people worldwide, and discovering their genetic causes is challenging. More than half of the individuals analyzed by the Undiagnosed Diseases Network (UDN) remain undiagnosed. The central hypothesis of this work is that many of these rare genetic disorders are caused by multiple variants in more than one gene. However, given the large number of variants in each individual genome, experimentally evaluating combinations of variants for potential to cause disease is currently infeasible. To address this challenge, we developed the digenic predictor (DiGePred), a random forest classifier for identifying candidate digenic disease gene pairs by features derived from biological networks, genomics, evolutionary history, and functional annotations. We trained the DiGePred classifier by using DIDA, the largest available database of known digenic-disease-causing gene pairs, and several sets of non-digenic gene pairs, including variant pairs derived from unaffected relatives of UDN individuals. DiGePred achieved high precision and recall in cross-validation and on a held-out test set (PR area under the curve $> 77\%$), and we further demonstrate its utility by using digenic pairs from the recent literature. In contrast to other approaches, DiGePred also appropriately controls the number of false positives when applied in realistic clinical settings. Finally, to enable the rapid screening of variant gene pairs for digenic disease potential, we freely provide the predictions of DiGePred on all human gene pairs. Our work enables the discovery of genetic causes for rare non-monogenic diseases by providing a means to rapidly evaluate variant gene pairs for the potential to cause digenic disease.

Introduction

Causal genetic variants have been identified for thousands of Mendelian diseases.[1–3] However, in spite of the advent of cheaper and more accurate sequencing technologies, causal variants have not been identified for approximately half (~3,000) of known rare genetic diseases.[4–6] To help address this challenge, the NIH established the Undiagnosed Diseases Network (UDN) in 2014. Comprising teams of researchers and clinicians from 12 sites across the United States, the UDN integrates whole-exome/genome sequencing with expert clinical evaluation to develop diagnoses and treatment plans for individuals who could not be diagnosed by conventional clinical approaches.[7–9] Although this approach has yielded much success,[10–21] more than half of all UDN cases remain undiagnosed. We hypothesize that many of these unsolved, rare cases might involve variants in multiple genes that only result in a disease phenotype when combined, which complicates diagnosis.

Variants in multiple genes can synergistically lead to disease via different mechanisms.[22–25] Digenic inheritance was first demonstrated in 1994, when concurrent mutations in two genes were found to cause retinitis pigmentosa.[26] Digenic inheritance is the simplest form of oligogenic inheritance in which variants in multiple genes lead to disease.[27–29] There are various classifications of digenic disease,[30] but in all cases of digenic inheritance, the phenotype results from variants in two genes. In isolation, the individual variants that form a digenic pair are benign or lead to a simpler phenotype. However, upon simultaneous mutation, the variants either interact to produce disease or combine to produce a more complex, and usually more severe, phenotype that cannot be explained by variants in one gene alone.

The Digenic Diseases Database (DIDA)[28] has chronicled several hundred cases of digenic disease. Analyses of DIDA have revealed that digenic-disease-causing gene pairs are more likely to functionally and/or physically interact with one another than expected by chance.[28] Machine learning approaches have been developed for distinguishing between different types of digenic disease pairs[31] and identification of disease-causing variant combinations,[32,33] including oligogenic combinations of greater than two genes.[34]

We hypothesize that the disease phenotype in some unresolved rare-disease-affected individuals is most likely a

[1]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA; [2]Department of Pediatrics, Division of Medical Genetics and Genomic Medicine, Vanderbilt University School of Medicine, Nashville, TN 37232, USA; [3]Pulmonary Hypertension Center, Division of Allergy, Pulmonary, and Critical Care Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [4]Department of Chemistry, Vanderbilt University, Nashville, TN 37235, USA; [5]Department of Pharmacology, Vanderbilt University, Nashville, TN 37235, USA; [6]Center for Structural Biology, Vanderbilt University, Nashville, TN 37235, USA; [7]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [8]Institute for Drug Discovery, Leipzig University Medical School, Leipzig 04103, Germany; [9]Department of Chemistry, Leipzig University, Leipzig 04109, Germany; [10]Department of Computer Science, Leipzig University, Leipzig 04109, Germany; [11]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [12]Bakar Computational Health Sciences Institute and Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94143, USA
*Correspondence: jens@meilerlab.org (J.M.), tony@capralab.org (J.A.C.)
https://doi.org/10.1016/j.ajhg.2021.08.010.

result of digenic inheritance and develop the digenic predictor (DiGePred), a high-throughput machine learning tool for evaluating the likelihood that dysfunction of gene pairs leads to digenic disease. We focus on the specific challenge of identifying gene pairs that have functional or phenotypic potential to cause a digenic disease when both are disrupted in an individual. We consider all cases in DIDA, which includes cases where both variants are required for disease, and cases in which having the variants simultaneously modifies disease presentation or severity. Our approach is based on supervised machine learning with a random forest classifier trained on diverse functional, network, and evolutionary properties of known digenic gene pairs versus realistic sets of non-digenic gene pairs, including variant pairs from healthy individuals. We evaluate the accuracy of DiGePred and demonstrate that it has a low false positive rate, which is essential for clinical applications. To aid in rapid screening of patients for potential digenic disease variants, we provide a classification of the digenic disease potential for all human gene pairs.

Subjects and methods

Digenic gene pairs

We obtained known digenic disease gene pairs from the Digenic Diseases Database (DIDA; with the latest version as of April 2021, which was updated in July 2017).[28] There were 140 unique gene pairs in DIDA. These pairs served as the "positive" training data for the machine learning classifier and were termed the "digenic" set of gene pairs. DIDA provides information about the genes mutated together in cases of digenic disease, the variants in the genes, the number of variants on both alleles, as well as information concerning the connectivity of the genes forming a gene pair, such as distance on protein-protein interaction (PPI) network, whether expressed in same tissue, whether members of the same biochemical pathway, and whether annotated to have the same function. The additional list of digenic pairs discussed in a follow-up paper by the group that produced DIDA[31] were not used for training.

Non-digenic gene pairs

We generated several sets of putative non-digenic gene pairs that served as the "negative" data in training different classifiers. The "unaffected" non-digenic set was created from genes with variants in the sequenced exomes or genomes of UDN individuals' relatives deemed unaffected by the UDN. Thus, we consider any combination of genes observed to be mutated simultaneously in any one "unaffected" individual to be non-digenic. Combining gene pairs from 55 individuals, the unaffected set contains 1.8 million gene pairs. We considered validation sets both with and without gene-level overlap with the training/validation sets. The "random" non-digenic set was created by selection of random pairs from the list of all human genes. The "permuted" non-digenic set was created by generation of all possible pairs of two genes from the DIDA genes, excluding actual DIDA pairs; this resulted in 13,390 permuted gene pairs. We created the "matched" non-digenic gene pair set from the random gene pairs by selecting gene pairs such that the distribution of the six network and functional features (NFFs) match those of the digenic set. We binned the digenic gene pairs by dividing the distribution of features into equal-sized intervals such that every feature value data interval had an equal number of gene pairs. We selected random gene pairs for the matched set such that the distributions of feature values for all the selected pairs recapitulated the overall distribution for all features of the digenic set, simultaneously.

Six network and functional features

Pathway similarity

The pathway annotations for the genes were derived from KEGG[35] and Reactome.[36] We used the Jaccard similarity metric[37] to calculate the proportion of pathway overlap between the two genes. The Jaccard similarity is measured by the ratio between the intersection of two sets and the union of two sets. In this case, we calculated the pathway similarity by taking the ratio of pathways annotations in common with both genes and pathway annotations associated with either. If both genes did not have pathway annotation, the similarity value was 0.

Phenotype similarity

The phenotype annotations from the Human Phenotype Ontology (HPO)[38] for the genes were used as features. The phenotypic overlap between the two genes was calculated with the Jaccard similarity metric, similar to as described above. The value for missing phenotype annotations was 0.

Co-expression

The co-expression data were derived from the COXPRESdb web server version 7.3.[39] The data are in the form of a mutual co-expression rank, which indicated how likely it was for a pair of genes to be co-expressed in the same tissue and the same level compared to other gene pairs. A lower rank indicated high co-expression. The inverse of the rank was used as the feature, and if either gene was not found in the co-expression database, the value used was 0.

PPI distance

The network data were downloaded from the UCSC gene and pathway interaction browser,[40] which in turn was derived from other sources of data, such as PPI databases,[41–44] functional annotation databases,[45] and others. The PPI network was based on experimental data regarding protein interactions. The inverse of the shortest path between a pair of genes on this network was used as the PPI distance feature.

Pathway distance

The pathways interaction network was based on interactions between the various curated biochemical pathways. The inverse of the shortest path between a pair of genes on this network was used as the pathway distance feature.

Literature distance

The literature-mined interaction network was made up of interactions derived from reported interactions or predicted associations in published biomedical literature. The inverse of the shortest path between a pair of genes on this network was used as the literature distance feature. For each network ("PPI," "pathway," and "literature"), a value of 0 indicates the absence of a path between the gene pair in the network and was thus assigned to pairs with missing data.

Five evolutionary features

Evolutionary age

We obtained the evolutionary ages of the proteins coded by the genes by using ProteinHistorian.[46] This estimates the ancestral branch on which the gene first appeared and the age in millions

of years. The quadratic mean of the values for each gene in a pair was used as a combined feature.

Gene essentiality

The gene essentiality scores provide a rank of how important and vital a gene is for normal physiology, viability, and survival. They were derived from the OGEE web server.[47,48] The essentiality scores are based on knockout (KO) experiments in model organisms and cell-based assays. The quadratic mean of the values for each gene was used as a combined feature.

Loss-of-function intolerance (pLI)

We added the loss-of-function intolerance (pLI) scores,[49] obtained from the Exome Aggregation Consortium (ExAC). These scores were based on the difference between actual mutation incidence and expected mutation frequency. A depletion of mutation incidence, compared to expected frequency, could mean the inability of the organism to survive if the gene was mutated. The quadratic mean was used as a combined feature.

Selection pressure (dN/dS)

We used measures of selection pressure in the form of dN/dS scores for the genes. These were derived from the EVOLA web server.[50] dN/dS ratios give a measure of the ratio between the non-synonymous mutations and synonymous mutations during evolution. This ratio tells us whether the gene has been evolving under strong positive, negative, or neutral selection. The quadratic mean was used as a combined feature.

Haploinsufficiency

We used the haploinsufficiency scores,[51] which were in the form of predictions of which genes were haploinsufficient on the basis of observed mutations. The quadratic mean was used as a combined feature.

Gene-focused network and functional features

Several additional gene-level attributes in the network and functional data sources described above were used as features.

Number of pathways

The feature used for the classifier was the quadratic mean of the number of pathways associated with gene A and the number of pathways associated with gene B.

Number of phenotypes

Similar to the pathways, the feature used for the classifier was the quadratic mean of number of phenotypes associated with gene A and with gene B, individually.

Network neighbors

The numbers of shared network neighbors, defined as the number of genes directly connected to both gene A and B, were also considered. For each gene pair, we computed the quadratic mean of the number of genes in the network directly connected to gene A and to gene B. These features were defined for all three types of interaction networks.

Number co-expressed

The number of genes highly co-expressed with both gene A and gene B were identified as the top 500 co-expressed genes (out of possible 20,000) for each. The feature used in the classifier was the quadratic mean number of genes highly co-expressed with gene A and gene B, individually.

Encoding gene-level features

Several of the evolutionary, genomic, and network features are attributes of individual genes rather than gene pairs. We combined these gene-level features into a single feature for each gene pair by computing their quadratic mean. Results were similar when we used the arithmetic mean (Figure S5).

Performance quantification

We computed receiver operating characteristic (ROC) and precision-recall (PR) curves to evaluate the performance of the classifiers. The ROC curve plots the false positive rate (FPR) on the x axis and the true positive rate (TPR) on the y axis. We used the area under each curve (AUC) to summarize performance.

Training and testing the DiGePred random forest models

We trained several random forest (RF) classifiers to distinguish digenic and non-digenic gene pairs. We selected RFs because they can integrate diverse features, perform well on unbalanced positive and negative sets, and provide interpretable models. The scikit learn (sklearn) python module was used for all training, evaluation, and prediction.[52] Hyper-parameters were selected by nested cross-validation on 80% of the labeled gene pairs. A stratified shuffle split was used for 10-fold cross-validation. This method involved splitting the data into ten equal parts; each part of the data contained approximately the same ratio of positives and negatives as the other parts. The optimum number of trees was found to be 500, and the maximum depth was found to be 15. On the basis of these analyses, we selected the classifier trained with the unaffected negative pairs and all features as the best model, and we refer to this as the DiGePred classifier.

The remaining 20% of the combined labeled data was held out for final performance validation of this best model from the cross-validation. These pairs had not been previously evaluated by the classifier. We also considered held-out test sets that had no overlap with the training/validation sets at the gene level ("no gene overlap" classifiers). In addition to the held-out positive digenic pairs, we generated 100 sets of held-out non-digenic pairs for evaluation. This enabled us to evaluate the best classifier 100-fold by using the same positive digenic pairs in every iteration but also a unique non-overlapping set of held-out non-digenic pairs in every iteration.

Evaluation with additional digenic pairs not in DIDA

The classifier was further evaluated with an external set made up of gene pairs considered to be digenic that were reported after DIDA was compiled. The external evaluation set was used in the previously published variant combination pathogenicity predictor (VarCoPP/ORVAL [Oligogenic Resource for Variant AnaLysis]).[32,34] This set had three unique gene pairs, which did not overlap with DIDA pairs. These gene pairs, (*AHI1*, *CEP290*), (*CEP290*, *CRB1*), and (*CEP290*, *RPE65*), were labeled Papadimitriou et al., 2019 validation set. We included recently discovered novel digenic inheritance of profound non-syndromic hearing impairment caused by (*PCDH15*, *USH1G*).[53] In addition, three recently reported cases of digenic inheritance in immune disorders were used. Ameratunga et al., 2017 identified epistatic interactions between *TACI* and *TCF3* (or *TNFRSF13B*) resulting in severe primary immunodeficiency disorder and systemic lupus erythematosus.[54] Hoyos-Bachiloglu et al., 2017 discussed how human immunodeficiency was caused by mutations in *IFNAR1* and *IFNGR2*.[55] We used more recent digenic findings such as (*LAMA4*, *MYH7*), linked to infantile dilated cardiomyopathy,[56] from Abdallah et al., 2019; (*KCNE2*, *KCNH2*), linked to long QT syndrome types 2 and 6,[57] from Heida et al., 2019; (*CLCNKB*, *SLC12A3*), linked to Gitelman syndrome,[58] from Kong et al., 2019; (*CACNA1C*, *SCN5A*), linked to long QT phenotype,[59] from Nieto-Marín et al., 2019; (*FGFR1*, *KLB*), linked to insulin resistance[60] and diabetes, from Stone

et al., 2019; (*CLCNKA*, *CLCNKB*), linked to Bartter syndrome with sensorineural deafness,[61] from Nozu et al., 2008; and (*CLCN7*, *TCIRG1*), linked to osteoporosis,[62] from Yang et al., 2018 to assess the classifier as well.

We also included gene pairs not characterized as digenic but displaying functional synergy associated with disease or adverse phenotypes. We derived the gene pair from the previously reported UDN study that found mutations in *TRPS1* and *FBN1* to be responsible for the patient phenotype, and it was labeled Zastrow et al., 2017 (UDN).[63]

Feature importance

To identify the most important features, we used the classifier feature importance function in sklearn, which uses the Gini impurity approach to quantify the relative feature importance for all features. Owing to possible biases in the Gini-based approaches when diverse features are considered,[64] we also used a permutation approach to calculate feature importance. This involved scrambling the feature values and comparing the error in classification between using the actual and permuted values for each individual feature.[65]

Prediction score thresholds

We determined a digenic score threshold for the DiGePred classifier for classifying gene pairs as digenic on the basis of the $F_{0.5}$ metric. This is a modification of the F_1 statistic, designed to attenuate the effect of false negatives. It is calculated as $F_\beta = ((1 + \beta^2)\ x\ TP) / ((1 + \beta^2)\ x\ TP + \beta^2\ x\ FN + FP)$, where $\beta = 0.5$, TP = true positives, and FP = false positives. The score that yielded the highest $F_{0.5}$ value was 0.496.

Estimating the false positive rate at various score thresholds

We evaluated the DiGePred classifier with an external set of non-digenic gene pairs as well. These gene pairs were obtained from 38 unaffected relatives of UDN individuals. The genes were preliminarily selected if the variant in the gene had an ExAC[66,67] minor allele frequency of <1%. A gene was further selected if it received a pathogenicity score of "D" ("probably damaging") from PolyPhen2 (Kircher et al., 2014).[68] Only genes passing this PolyPhen2 filter were selected to limit the predictions to pairs of genes with variants that most likely affected molecular function.

Additionally, genes with rare variants were selected on the basis of a consensus pathogenicity approach if at least two out of PolyPhen2, SIFT,[69,70] CADD (Kircher et al., 2014[68]; Rentzsch et al., 2019[71]), and PhyloP[72] agreed that the variant(s) in the gene was pathogenic. A PolyPhen2 selection criterion was similar to above. A variant was deemed pathogenic by SIFT if the score was ≤ 0.05. A CADD score $\geq$ 30 was considered pathogenic, while a PhyloP score of ≤ -10 for a variant deemed it pathogenic. All possible gene pairs were used as the consensus pathogenic gene pairs for an individual.

The fraction of gene pairs predicted to be digenic was compared for individuals with undiagnosed disease versus unaffected members of UDN cohorts. The comparison of these fractions was done for the most confident DiGePred thresholds ($F_{0.5}$ and higher), and the Mann-Whitney U (MWU) test p value was calculated for each and every threshold.

Comparison with ORVAL

We submitted the list of gene pairs for all the unaffected individuals to the ORVAL (Oligogenic Resource for Variant AnaLysis)[32,34] server. We compared the number of pairs predicted to be digenic by ORVAL, according to its highest confidence threshold, to the number predicted by our method to be digenic at the $F_{0.5}$ threshold. We obtained the list of genes for each unaffected individual as mentioned in the previous section. We evaluated the statistical significance of the number of digenic pairs predicted as false positives per individual between DiGePred and ORVAL by using a MWU test.

Furthermore, 20% of all genes with rare variants in the individual were chosen at random. We generated all possible gene pairs to constitute the random set of gene pairs for each individual. We calculated the number of digenic pairs predicted per individual at different score thresholds. We did this to compare the number of false positives between ORVAL and DiGePred fairly. Because ORVAL includes variant effects as a feature, selecting for genes with variants that were predicted pathogenic by PolyPhen2 or by a consensus of several predictors of variant effect could bias against ORVAL, although it reflects common clinical practice. Therefore, we also compared DiGePred and ORVAL on pairs of genes selected at random.

For the purpose of comparing ORVAL predictions on individuals with undiagnosed disease and unaffected members of UDN cohorts, we further ranked ORVAL predictions by using the ORVAL classification score as a prediction threshold. According to the authors, pairs with a classification score of >0.74 with a support score of 100 were scored in the 99% confidence zone. We compared the fraction of gene pairs predicted to be digenic at varying ORVAL classification score thresholds, ranging from 0.74 and higher, for diseased versus unaffected individuals. We calculated the MWU test p value for the distributions at each and every threshold.

Gene Ontology (GO) enrichment

The GO enrichment was computed with a web resource, WebGestalt (WEB-based GEne SeT AnaLysis Toolkit).[73] A list of genes was prepared for each selected set of predicted digenic pairs on the basis of highest score, highest average score, or most predicted pairs. This list of genes was ranked on the basis of the selection criteria, and the GO enrichment for biological process, cellular component, and molecular function categories was performed with the online tool.

Results

Digenic disease gene pairs have different attributes than non-digenic disease gene pairs

Our goal in this study is to develop a machine learning classifier for identifying gene pairs that cause disease when both are disrupted simultaneously but produce no or less severe phenotypes when disrupted in isolation. To this end, we consider all unique known digenic disease pairs curated by the DIDA and contrast them with several informative sets of non-digenic disease pairs. Because our ultimate application is the detection of potential digenic diseases in patients, most of our results focus on comparisons of known digenic gene pairs and gene pairs with variants in "unaffected" parents, siblings, and other relatives

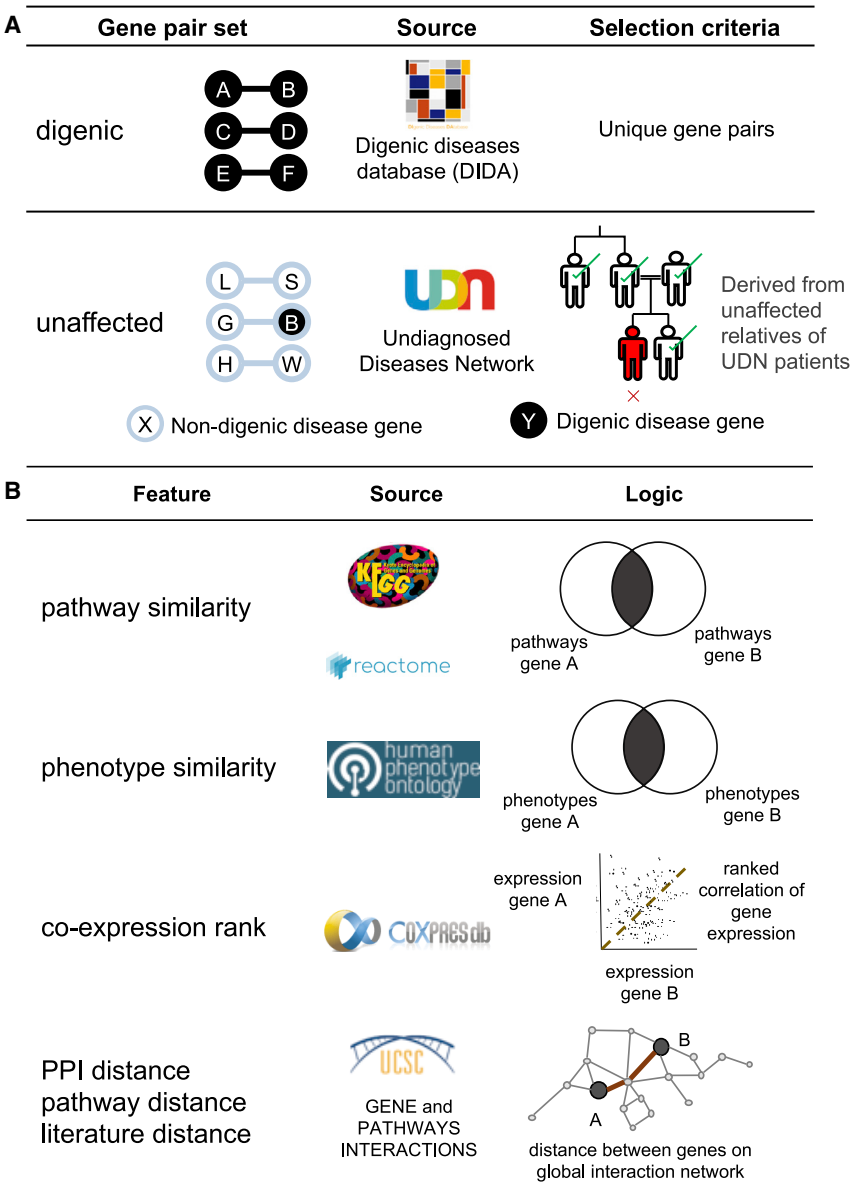

Figure 1. Training sets and features used for machine-leaning-based identification of digenic disease gene pairs

(A) Digenic gene pairs (positives) were derived from the Digenic Diseases Database (DIDA). Unique gene pair combinations (n = 140) were used for training and testing. Non-digenic gene pairs (negatives) were derived from unaffected relatives of UDN individuals. Genes with rare variants in the same individual were used as an unaffected non-digenic gene pair. We also considered several other negative training sets, including random gene pairs, permuted pairs of genes involved in digenic pairs, and gene pairs matched to attributes of digenic pairs (Figure S1).

(B) We considered six network and functional features (NFFs) for training the digenic disease classifiers: (1) "pathway similarity," Jaccard similarity of pathway annotations from KEGG and Reactome for both genes; (2) "phenotype similarity," Jaccard similarity of phenotype annotations from HPO for both genes; (3) "co-expression rank," co-expression rank of gene pair compared to all other gene pairs across multiple tissues from COXPRESdb; (4–6) "network distances" between the genes on protein-protein, pathway, and literature-mined interaction networks from UCSC gene and pathway interaction browser database. We also trained classifiers considering additional evolutionary and functional features (Figure S2).

of 25 UDN individuals (Figure 1A). However, as we show below, our results are similar with other strategies for defining non-digenic disease gene pairs.

Pairs of genes harboring mutations known to cause digenic disease have distinct biological properties when compared with random gene pairs.[28] Previous work has shown that digenic disease pairs have high protein interaction network connectivity and proximity. More than 35% of known digenic disease pairs directly interact on a PPI network, and ~60% of digenic gene pairs are one gene away on the interaction network. Similarly, ~20% of digenic pairs are in the same biochemical pathway, and ~40% are expressed in the same tissues.[28]

Based on this prior knowledge we devised a list of six network and functional features (NFFs) to use as attributes for distinguishing between digenic and non-digenic gene pairs (Figure 1B): (1) "pathway similarity," defined as the Jaccard similarity[37] between the genes' membership in ~1,800 pathways from KEGG[35] and Reactome;[36,74] (2) "phenotype similarity," the Jaccard similarity between the ~6,000 phenotypes from Human Phenotype Ontology (HPO)[38] associated with the genes; (3) "co-expression rank," defined as the rank of the co-expression of the genes across 23 co-expression platforms from 11 species compared to other gene pairs from COXPRESdb;[39] (4) "PPI distance," the distance on a global PPI network; (5) "pathway distance," the distance on an annotated biochemical pathway network; and (6) "literature distance," the distance on a literature-mined interaction network, derived from the UCSC gene and pathway interaction database.[40]

We compared the distribution of the NFFs for known digenic pairs and for non-digenic gene pairs from unaffected relatives of UDN individuals. As expected from previous work, the distribution of each NFF was significantly different between digenic and non-digenic pairs (Figure S3; $p < 10^{-20}$ for each, MWU test). This suggests that a machine learning approach may enable distinguishing digenic from non-digenic disease pairs.

To further explore the properties of digenic disease genes and the ability of a classification approach to recognize them, we defined three additional sets of non-digenic

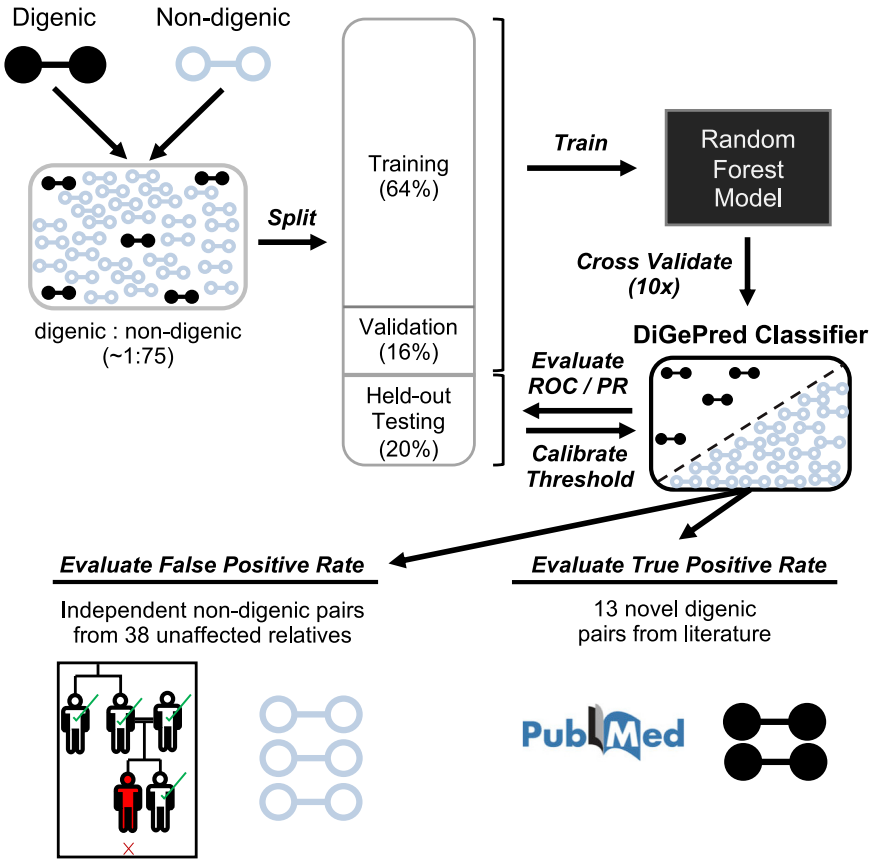

Figure 2. Schematic of the protocol for training and evaluating the DiGePred digenic disease pair classifier

Known digenic pairs (positives) and variant gene pairs from healthy individuals (negatives) were combined at ~1:75 ratio. The combined pairs were divided into training (64%), validation (16%), and held-out test datasets (20%). The DiGePred random forest classifier was trained and cross-validated with the training and validation sets. The final performance estimate for the trained DiGePred classifier was quantified by the receiver operator characteristic (ROC) area under the curve (AUC) and precision-recall (PR) AUC on the held-out test set. This set was also used for establishing suggested thresholds on the continuous DiGePred score. DiGePred's potential clinical utility was further demonstrated by applying it to an additional positive set of 13 novel digenic pairs from the recent literature, one novel gene pair in a resolved UDN individual, and an external set of non-digenic gene pairs from 38 unaffected relatives of UDN individuals.

disease gene pairs (Figure S1). First, we created a "permuted" non-digenic set by generating all possible gene pairs from genes known to be involved in a digenic gene pair and removing the pairs known to be digenic. Second, we constructed a "random" set of non-digenic gene pairs by randomly selecting gene pairs from all possible human genes (excluding known digenic pairs). Third, we created a "matched" non-digenic gene pair set that closely matched the NFF distributions of the digenic gene pairs; however, we were not able to match the distribution of all NFFs perfectly given the skewed distribution of the digenic disease pairs (Figure S3). Nonetheless, the matched set enables exploration of how well our classification approach can identify digenic pairs among non-digenic pairs with similar NFF distributions. To be conservative, we also constructed non-digenic gene pair sets with no overlap between the individual genes present in the training and the held-out test datasets.[75] These are subsets of the unaffected and random sets and will be referred to as "unaffected no gene overlap" and "random no gene overlap," respectively (Figure S1).

Random forest classifiers accurately identify digenic pairs via network and functional features

We divided the available gene pairs into training (64%), validation (16%), and held-out test sets (20%). We trained, evaluated, and compared different models by using 10-fold cross-validation within the training and validation sets (Figure 2). The test set was only analyzed after models had been finalized. Comprehensive studies of genetic interactions have found that one in approximately 40 gene pairs interact.[76] This suggests that digenic interactions are most likely rare; only a very small fraction of all possible gene pairs is likely to produce digenic disease. We trained the random forest machine learning classifier by using the six NFFs to distinguish 140 digenic disease gene pairs (positives) from ~8,400 negative gene pairs. Unless otherwise specified, we focus in the main text on the "unaffected no gene overlap" negative set and present others in the supplemental information. The large class imbalance (~1:75) reflects our expectation few digenic gene pairs among all possible pairs to be evaluated; this exact ratio was selected because of data availability. We evaluated performance by using ROC and PR curves.

The random forest classifier distinguished digenic and non-digenic gene pairs very accurately with the six NFFs. It achieved an average ROC area under the curve (AUC) of 0.90 and a PR AUC of 0.698 on average over 10-folds of cross-validation on the training and validation sets (Figures 3 and S4). The algorithm retains near perfect precision at recall above 60% (Figure 3B). Because we are evaluating multiple classification approaches, the held-out test set was not considered in this analysis.

Including additional features improves ability to identify digenic disease genes

The performance of the classifier based on the six NFFs alone was strong; however, there are many other sources of biological information beyond the NFFs that could

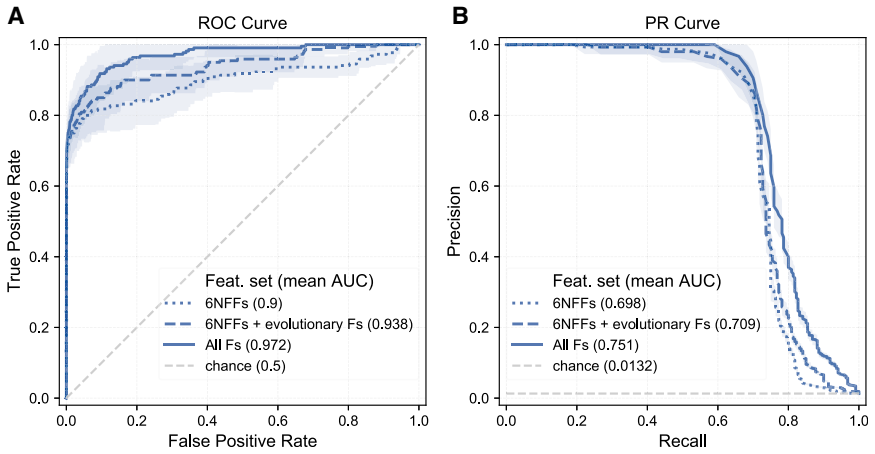

Figure 3. Random forest classifiers can accurately distinguish digenic and non-digenic gene pairs via different feature sets

(A and B) Performance of classifiers at distinguishing between known digenic pairs from DIDA (positives) and gene pairs from 25 healthy individuals (negatives) trained via different feature sets as evaluated by receiver operating characteristic (ROC) curves (A) and precision-recall (PR) curves (B). Classifiers trained on three sets of features are compared: (!) six network and functional features (NFFs) (dotted line); (2) the six NFFs and evolutionary genomics features; and (3) the six NFFs, evolutionary genomics features, and gene-level network and functional features. The mean curves across 10-fold cross-validation on the training and validation sets are plotted with shaded areas representing the standard deviation. Because this analysis is developing and evaluating multiple possible classifiers, we held out the test set for final evaluation (Figure 4).

potentially inform either the nature of the relationship between genes or the relative likelihood and risk of a gene's being mutated and causing disease. We tested whether including additional features in training the classifier would increase performance of the classifier.

First, we trained classifiers by using the six NFFs and five additional evolutionary features that reflect the evolutionary history and constraint on the genes (Figure S2A). These features were as follows: (1) the evolutionary ages of the genes, (2) their essentiality, (3) their intolerance to loss-of-function mutations, (4) the selection pressure acting on them through mammalian evolution (dN/dS), and (5) their haploinsufficiency scores. The addition of evolutionary features, as the quadratic mean of the values associated with both genes, substantially improved classifier performance: average ROC AUC of 0.938 and PR AUC of 0.709 (Figures 3 and S4).

Next, we considered additional features derived from network and functional annotations of the gene pairs (Figure S2B). These features were designed to add additional gene-focused (rather than gene-pair-focused) context and explore the sufficiency of the six NFFs. These features were as follows: (1) the number of pathways, (2) phenotypes, (3) network neighbors, and (4) genes co-expressed for each individual gene in a candidate pair. As above, we used the quadratic mean to combine these gene-level features. Considering these features also further improved classifier performance; there was an average ROC AUC of 0.972 and PR AUC of 0.751 for all features (Figures 3 and S4).

Digenic disease genes can be distinguished from many non-digenic gene sets

We used the same training and evaluation approach as described in the previous section for the unaffected no gene overlap negative set to train random forest classifiers to distinguish digenic disease gene pairs from each of the additional negative sets (random, permuted, and matched) by using all the network, functional, and evolutionary features. In each case, the classifiers performed very well (Figure S4). The classifiers trained to distinguish digenic pairs from random and random no gene overlap pairs performed the best (mean ROC AUC of 0.972 and 0.968 and PR AUC of 0.696 and 0.741, respectively) with all features included for training. As expected, given their similar attributes to the digenic pairs, the permuted and matched negative sets are more challenging for the classifiers, but they still achieved very strong performance with average ROC AUCs of 0.964 and 0.977 and PR AUCs of 0.54 and 0.597, respectively.

Feature importance varies for classifiers trained on different non-digenic sets

We estimated the importance of the features to the classifiers by using the mean decrease in node impurity approach (Figure S6). For the classifier trained with variant gene pairs from unaffected relatives, the Jaccard similarity of phenotypes associated with each gene for a gene pair was the highest weighted feature (37%). The pathway similarity and the mean number of phenotypes for the gene pair were among the other important features (10% and 7% of the weight, respectively). The feature importance values were similar for the classifiers trained with random gene pairs and permuted digenic gene pairs (Figure S6).

The feature importance values were most different for the matched classifier; it placed significantly lower feature importance on the NFFs. This was expected because the differences between the positive and negative training examples in individual NFFs were minimal for this classifier by design. Instead, a range of evolutionary and individual gene-level functional features took on similar levels of importance (Figure S6). This indicates that information in gene-level features related to evolution, gene importance, and relevance to physiology contain useful information about the likelihood of gene pairs interacting to produce digenic disease.

The impurity approach for feature importance calculation can be biased, especially when the classification task includes features with both continuous and discrete

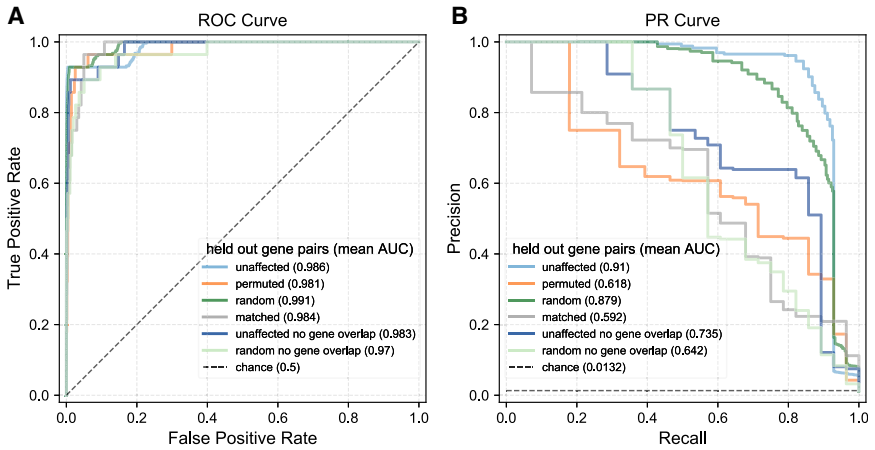

Figure 4. Classifiers accurately distinguish digenic pairs from non-digenic pairs on held-out test sets

(A and B) ROC (A) and PR (B) curves for random forest classifiers trained with all features on digenic gene pairs and various negative sets (indicated in the legend) and evaluated on the appropriate held-out test sets. These test sets consisted of DIDA held-out pairs as positives and six different held-out negative sets: (1) "unaffected," derived from healthy relatives of UDN individuals (light blue); (2) "permuted," derived by generating permutations of known digenic pairs (orange); (3) "random," derived by randomly selecting pairs of genes (dark green); (4) "matched," derived by matching the distribution of network and functional features observed among the digenic pairs (gray); (5) "unaffected no gene overlap," derived from healthy relatives of UDN individuals and no genes in common between the training and test datasets (dark blue); (6) "random no gene overlap," derived by randomly selecting pairs of genes with no genes in common between the training and test datasets (light green). The ROC AUCs were >0.97 in all cases, while the PR AUCs were >0.6 in all cases. In all subsequent analyses, the "unaffected no gene overlap" classifier will be referred to as "DiGePred."

values.[64] Therefore, we also used a permutation approach to calculate the importance for each feature on the basis of the error in classification after the feature values were scrambled. Phenotype similarity was still the most important feature (Figure S7A), and the feature importance values calculated on the basis of the impurity and the permutation approach generally agreed (Spearman rho = 0.404, Figure S7B).

DiGePred accurately identifies held-out digenic pairs

To obtain an unbiased estimate of the best classifiers' performance, we evaluated them by using held-out test sets of digenic and non-digenic pairs. These sets were not used for training or validating the classifier and maintained the 1:75 ratio used during training. The classifiers trained with gene pairs observed in unaffected relatives of UDN individuals as negatives most closely reflect the distribution of gene mutations likely to be seen in real clinical applications. Based on the previous results, the best balance between performance and stringency in selecting the negatives was achieved for the unaffected no gene overlap model with all the features used during training. We focus on this model going forward but report results for all classifiers.

The ROC AUC for the unaffected no overlap classifier on the held-out sets was 0.983, while the mean PR AUC was 0.742 (Figure 4). The classifiers trained on the other non-digenic gene pair sets also performed well on their corresponding held-out sets: the ROC AUCs were better than 0.97 and PR AUCs were better than 0.59 in all cases (Figure 4).

To establish thresholds for predicting potential digenic gene pairs on the basis of the output of the unaffected no gene overlap classifier, we computed thresholds that maximize the F_1 and $F_{0.5}$ scores. The F_1 is maximized at a digenic score of 0.156, and the $F_{0.5}$ is maximized at a digenic score of 0.496. Because we anticipate that precision is more important than recall in most applications, we suggest use of the $F_{0.5}$-based threshold. At this threshold, the classifier correctly identified 13 of 28 digenic gene pairs in the held-out test set and had with a false positive rate of 0.14% (Figure 4, Dataset S1). We refer to this model as the digenic predictor (DiGePred).

DiGePred identifies novel digenic pairs from the recent literature

Although the test set was not seen by the classifier prior to evaluation, it was still obtained from DIDA, the source of digenic pairs for training and testing. Thus, we further applied our classifier to 13 digenic pairs obtained from recent literature not included in DIDA (Table S1). We derived three digenic pairs [(*CEP290*, *RPE65*), (*AHI1*, *CEP290*), (*CEP290*, *CRB1*)] from the validation set used by a recently published digenic classifier.[32] The other digenic gene pairs [(*CLCNKA*, *CLCNKB*), (*TCF3*, *TNFRSF13B*), (*IFNAR1*, *IFNGR2*), (*PCDH15*, *USH1G*), (*LAMA4*, *MYH7*), (*KCNE2*, *KCNH2*), (*CLCNKB*, *SLC12A3*), (*CACNA1C*, *SCN5A*), (*FGFR1*, *KLB*), (*CLCN7*, *TCIRG1*)] were derived from recently reported cases of digenic disease (Abdallah et al., 2019; Ameratunga et al., 2017; Heida et al., 2019; Hoyos-Bachiloglu et al., 2017; Kong et al., 2019; Nieto-Marín et al., 2019; Nozu et al., 2008; Schrauwen et al., 2018; Stone et al., 2019; Yang et al., 2018, respectively). We note that these pairs include some similar phenotypes and overlapping genes and so should not be viewed as 13 independent tests.

DiGePred correctly identified 11 of the 13 novel digenic pairs at the $F_{0.5}$ threshold. (Figures 5, S8, and S9). Two of the gene pairs missed at the $F_{0.5}$ threshold, *IFNAR1* and *IFNGR2* (Hoyos-Bachiloglu et al., 2017) and *LAMA4* and *MYH7* (Abdallah et al., 2019) were identified as digenic at the F_1 threshold (expected FPR of 0.5%) (Figures S8 and S9).

We also evaluated a gene pair from a solved UDN case in which variants in *FBN1* and *TRPS1* caused independent autosomal-dominant conditions with some overlapping

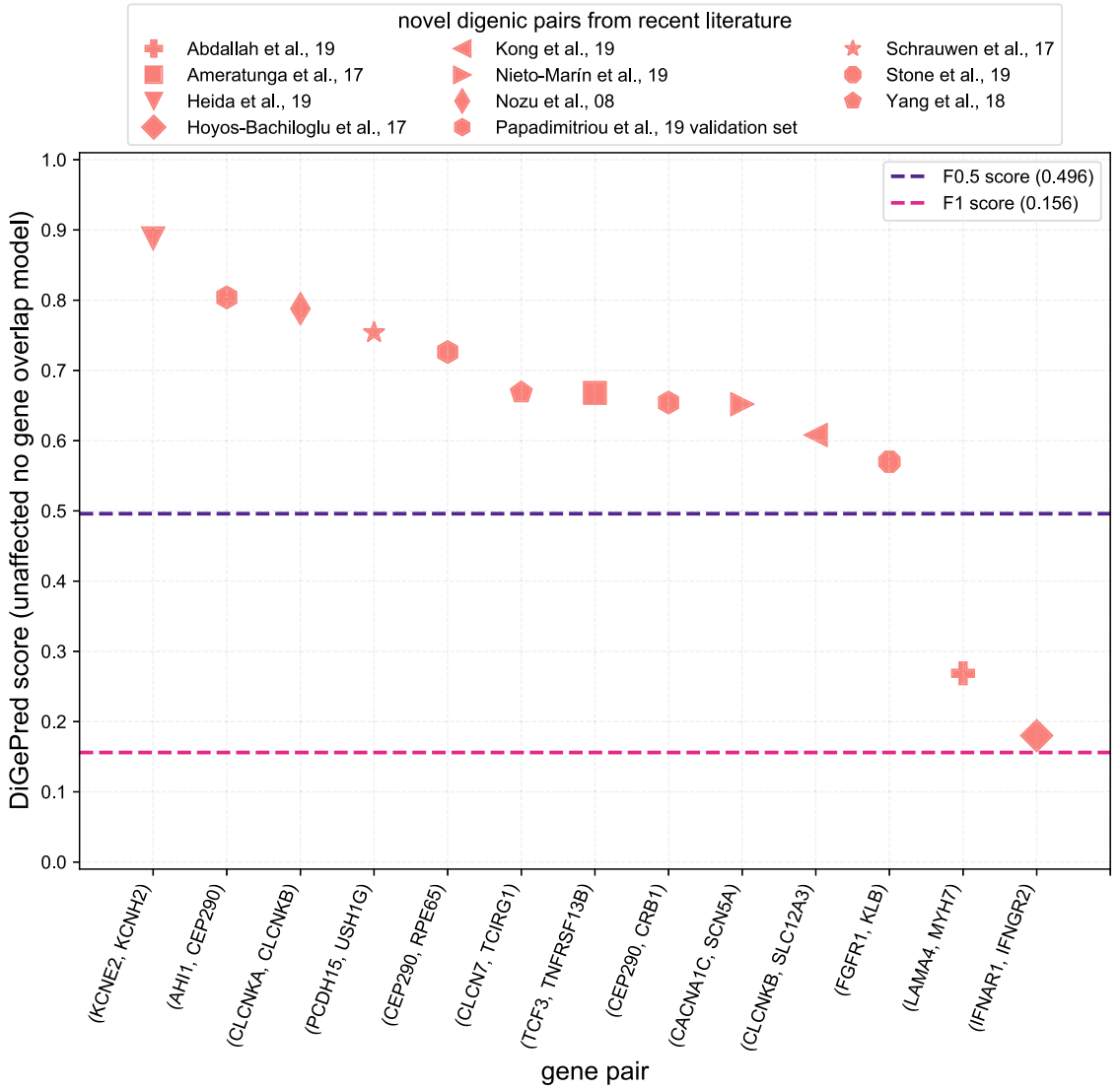

Figure 5. DiGePred accurately identifies novel digenic pairs from the recent literature
Geometric shapes in red indicate the DiGePred scores assigned to 13 novel digenic pairs reported in the recent literature. The dashed pink and purple lines represent the DiGePred score thresholds that maximize the F_1 (0.156) and the $F_{0.5}$ (0.496) metrics (Figure S8). Given the importance of precision in clinical applications, we propose the score maximizing the $F_{0.5}$ metric or higher as a threshold for calling a gene pair digenic. At this threshold, 11 of the 13 novel digenic pairs are predicted to be digenic with a low expected false positive rate ($\leq$0.14%). All digenic pairs score above the F_1 threshold. The DiGePred classifier was trained with all features and the unaffected no gene overlap set as negatives.

symptoms that produced a unique phenotype in the affected individual (Zastrow et al., 2017).[63] As a result of the lack of interaction, this pair does not meet the strict criteria for digenic pairs used here. DiGePred predicted that this gene pair was not digenic at the $F_{0.5}$ threshold. Nonetheless, it was predicted at the F_1 threshold (Figure S8, Dataset S2), suggesting the potential of the classifier to highlight pairs of functionally related genes.

DiGePred has a low false positive rate in real-world applications

Individuals often carry hundreds of protein-coding variants of unknown significance, which results in thousands of potential digenic disease pairs per individual. Thus, when considering the application of classifiers to individuals' genomes, it is essential to understand and control the false positive rate. To this end, we evaluated DiGePred on gene pairs with rare variants predicted to disrupt protein function in 38 human genomes from unaffected parents and relatives of UDN individuals not used in training the algorithm. These healthy individuals should not contain any true digenic disease pairs, so any positive predictions on gene pairs from these individuals are very likely to be false positives. The gene pairs from these individuals were not used in the training, validation, or held-out test sets.

At the $F_{0.5}$ threshold, 8% of unaffected individuals had no predicted candidate digenic pairs and 29% had only

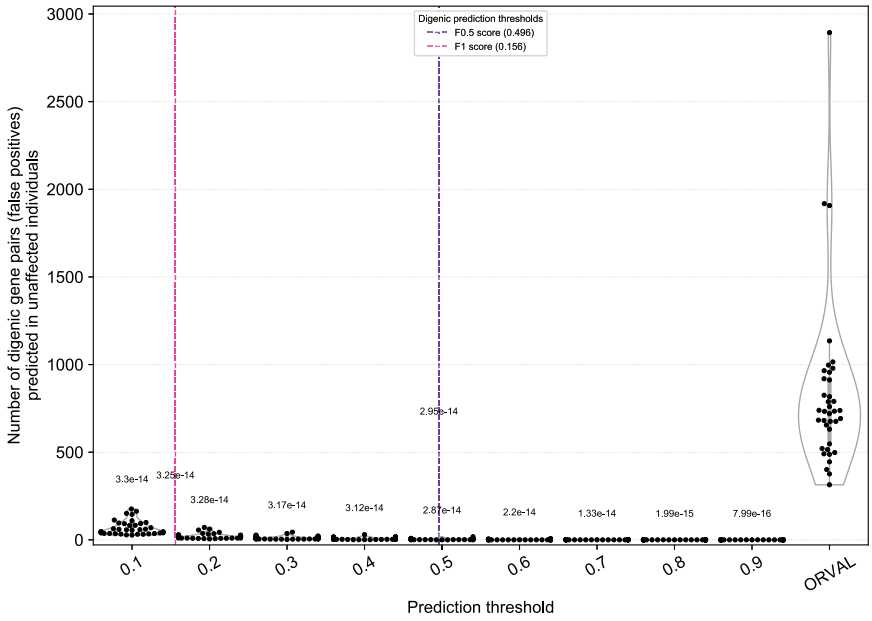

Figure 6. DiGePred has a low false positive rate and outperforms a recent digenic gene prediction method

The number of digenic pairs identified for each of 38 healthy relatives of UDN individuals is plotted at a range of DiGePred thresholds (x axis) and for the highest confidence predictions (99% threshold) of the ORVAL/VarCoPP method. The DiGePred score thresholds that maximize the F_1 and $F_{0.5}$ metrics on the held-out data are shown in pink and purple, respectively. Because the individuals considered are healthy, any predicted digenic disease pairs are very likely false positives. DiGePred predicts significantly fewer digenic pairs at each threshold than ORVAL (Mann-Whitney U test, p values above each bar). At the $F_{0.5}$ threshold, DiGePred predicts an average of under four digenic pairs per healthy individual and none above the 0.9 threshold, while ORVAL predicts an average of 830 digenic pairs per healthy individual at its strictest threshold (Figure S10). Results were similar for classifiers trained on other negative sets (Figures S10–S18).

one candidate digenic pair. On average, less than four digenic pairs were predicted per individual, and only six had more than five pairs (Figures 6 and S10). Furthermore, we emphasize that users can adjust the score threshold to reflect their tolerance for false positives in different applications; for example, the fraction of individuals with no digenic gene pairs predicted was 31%, 66%, and 92% at score thresholds of 0.6, 0.7, and 0.8, respectively (Figure S10).

In contrast, we applied the ORVAL[32,34] method for identifying digenic disease pairs to variants from these same individuals. At its highest confidence threshold, ORVAL predicted that all these healthy individuals have digenic disease pairs and had an average of 830 highest confidence digenic pairs per individual. All individuals were predicted to have >300 digenic pairs, and five (~13%) had more than a thousand digenic pairs predicted (Figure 6). This is a significantly larger number of candidate digenic disease pairs per individual than DiGePred ($p = 2.95 \times 10^{-14}$, MWU test), and these are very likely to be false positives given that these are healthy individuals. This difference in number of false positives was recapitulated for all gene selection criteria, variant pathogenicity prediction approaches, and all models of training considered (Figures S11–S18).

We found that 11.8% of false positives in unaffected individuals (gene pairs incorrectly predicted as digenic by DiGePred) had at least one gene as a member of a known digenic pair in DIDA. Only 3.2% of all gene pairs evaluated by DiGePred had at least one gene overlapping with known digenic pairs from DIDA. This is an approximately 4-fold enrichment of such gene pairs among false positives compared to the genome-wide expectation ($p = 1.31 \times 10^{-5}$).

Prediction of digenic pairs among individuals with undiagnosed disease

To illustrate the application of DiGePred in patients with rare undiagnosed genetic disorders, we applied it to (1) affected individuals from the UDN site at Vanderbilt and (2) a cohort of 111 individuals with Mayer–Rokitansky–Küster–Hauser (MRKH) syndrome.

We first considered variants from ~50 UDN cases and identified several candidate digenic pairs based on DiGePred score integrated with analyses of variant effect, variant inheritance, and similarity of the gene's functions to the patient phenotype. Because these cases are still being actively evaluated, we cannot report full details here. Instead, we describe a representative example. We predicted a candidate digenic pair of *ATXN2* (ataxin 2) and *FUS* (fused in sarcoma) for an individual with ALS (amyotrophic lateral sclerosis)- and Parkinsonism-like phenotypes. The variant in *ATXN2* was a polyglutamine (polyQ) repeat expansion variant, and there is evidence in literature for a functional interaction between these two genes.[77–79]

To explore the performance of DiGePred on UDN individuals more quantitatively, we compared the predictions on variants from 24 affected individuals with 38 available unaffected relatives that were not used in the training of DiGePred. We tested whether the rare disease patients had a higher median of fraction of high-confidence predicted digenic pairs compared to related individuals without rare disease. At all thresholds considered ($F_{0.5}$ or higher), DiGePred predicted a greater fraction of gene pairs with variants to be digenic for individuals with undiagnosed disease than for the unaffected individuals. The difference between the distributions was significant ($p = 6.74 \times 10^{-12}$, Kolmogorov–Smirnov test; Figure S19). In

contrast, the fraction of predicted digenic pairs was similar for the individuals with undiagnosed disease compared to unaffected individuals across a range of ORVAL classification scores within the 99% confidence zone ($p = 0.482$; Figure S20).

Next, we applied DiGePred to variants from a cohort of 111 individuals with MRKH syndrome,[80] a developmental disorder primarily affecting the female reproductive system, often characterized by a congenital absence of a uterus or vagina.[81,82] We identified a potential digenic pair between *LAMC1* (laminin subunit gamma 1), an extracellular matrix (ECM) glycoprotein that is a member of the integrin pathways and plays a role in cell adhesion and signaling, and *MMP14* (matrix metallopeptidase 14), a protein involved in breaking down the extracellular matrix during embryonic development and tissue remodeling. The DiGePred prediction was driven by the two proteins' being highly co-expressed with one another, directly interacting along the integrin pathway, being only one protein away on the global PPI network, and having ~5% phenotype similarity. Furthermore, there is evidence in literature of functional interaction between *LAMC1* and *MMP14* that affects ECM remodeling via fibronectin deposition in zebrafish.[83]

Prediction of digenic pairs among all human gene pairs at various confidence thresholds

To aid in the rapid evaluation of digenic disease potential for a pair of genes of interest, we trained a new DiGePred classifier by using all digenic pairs from DIDA (to maximize use of available data) and variant gene pairs from healthy relatives of UDN individuals. We applied DiGePred to all possible human gene pairs. A gene pair was deemed a candidate digenic pair if the digenic score met the $F_{0.5}$ threshold as described above. As expected, the percentage of all possible gene pairs that were identified as digenic at our most confident threshold was very low (54,318 out of 155.33 million gene pairs, 0.035%). These predictions and the raw digenic scores are available in Dataset S3.

Overall, 7,970 unique genes are involved in at least one predicted digenic pair. This illustrates that DiGePred is not just prioritizing gene pairs that include a gene in a known digenic pair. In fact, only three of the top 100 genes with the most predicted digenic pairs occur in a DIDA pair. These genes are enriched for several essential developmental and molecular GO functional annotations, including "maintenance of cell number" (7.5× expected, false discovery rate [FDR] = 0.005), "chromatin remodeling" (7.3×, FDR = 0.005), and "membrane docking" (7.0× expected, FDR = 0.004; Table S2). For example, *FGF5*, a growth factor important for cell proliferation and differentiation and tissue development and repair, had the highest number of predicted digenic pairs above the $F_{0.5}$ threshold with 370. *ARID1B*, which had the 2nd highest number of predicted digenic pairs, with 262, encodes a component of the SWI/SNF chromatin remodeling complex with broad regulatory functions across the genome. CEP290, a centrosome protein with essential roles in centrosome and cilia development in many cell types, had the 6th most predicted digenic interactions with 232. The genes with the most predicted digenic pairs were also enriched for several organ development and cell cycle processes. The top 100 gene pairs with the highest average DiGePred scores were enriched for tissue and organ development, ciliary function, and electron transfer activity (Figures S21–S23, Tables S3 and S4).

We found that 19,325 (35%) of predicted digenic gene pairs had at least one recessive phenotype associated in OMIM.[84–86] In almost a fifth of these cases (3,697; 19%), at least one phenotype was in common or with high semantic similarity[87] between the two genes. For most of these gene pairs (3,601; 97%), the two genes had different MIM numbers in OMIM. This indicates that the two genes have not been previously annotated as causing a digenic disease[84] and, thus, suggests that they are novel associations.

Existing knowledge provides plausible mechanisms underlying many of these predicted novel digenic gene pairs. For example, a digenic pair comprising *STIM1* and *ORAI1* had the 4th highest score over all human gene pairs. It has been previously reported that *STIM1* and *ORAI1* function together to form Ca^{2+} release-activated Ca^{2+} (CRAC) channels, which are responsible for Ca^{2+} influx called store-operated Ca^{2+} entry (SOCE).[88] The proper functioning of these channels is necessary for maintaining the normal physiology of several cell types, including T cell receptors and human lymphocytes.[89–91] Missense variants in *STIM1* and *ORAI1*, individually, cause diseases with a great degree of phenotypic homogeneity.[92] Loss-of-function variants in *STIM1* and *ORAI1* have also been known to cause immunodeficiency,[93–96] under autosomal-recessive conditions, as reported by OMIM. Therefore, it is possible that single loss of variants in both genes occurring simultaneously could lead to the autosomal-recessive immunodeficiency.

Discussion

In this paper, we describe DiGePred, a high-throughput machine learning approach for identification of gene pairs with the potential to cause digenic disease. We demonstrate the accuracy and robustness of our approach in several realistic scenarios. We were motivated to create DiGePred by the challenge of identifying causal variants in individuals with rare disease that cannot be explained by a single variant. It is not feasible to experimentally evaluate all candidate pairs of variants in an individual of interest. Thus, to facilitate the rapid identification of candidate digenic gene pairs in affected individuals, we provide DiGePred predictions for all pairs of human genes at several confidence thresholds (Datasets S4A–S4D).

The DiGePred classifier trained with negatives derived from unaffected relatives is most likely best suited to the purpose of identifying digenic pairs in individuals with rare disease because it reflects the baseline distribution of gene pairs with variants identified via clinical sequencing

pipelines in individuals without severe disease. Moreover, classifiers trained with these negative sets performed well. However, our approach performs well at distinguishing digenic pairs from several additional sets of candidate non-digenic gene pairs, and the features used by these classifiers are similar unless the prediction problem is explicitly engineered to make them different (Figures 4 and S4).

Nonetheless, there is still much to learn about the mechanisms underlying digenic diseases. The features prioritized by our models support previous work[28,31] in that phenotypic similarity, number of phenotypes, and involvement in the same molecular pathways are the most important predictors. They also suggest that these may be more specific predictors of digenic gene pairs than similar co-expression profiles or close interaction network distance. Our results via the use of negatives that match the network and functional features between positives and negatives sets indicate that digenic gene pairs also have differences in their evolutionary attributes.

Our analyses are based on the examples available in DIDA, but there are most likely hundreds or even thousands of undiscovered digenic diseases. The strong performance of DiGePred on the test set with no gene overlap with the training set and DiGePred's ability to identify new digenic pairs from the recent literature (Figure 5) suggest that the algorithm will generalize. However, we note that our performance estimates may be optimistic; known digenic pairs are unlikely to be an unbiased sample of the full spectrum of digenic mechanisms. We anticipate that our algorithms will further improve as more digenic diseases and their causal molecular mechanisms are determined.

DiGePred is based on functional, biological network, and evolutionary features in a random forest model. Phenotype similarity and other phenotype-related features, such as the mean number of phenotypes associated with each individual gene, were the most important features. Given that our understanding of function of most genes is incomplete, the high reliance on a phenotype-based features could lead to a high-performing model that does not generalize when these features are missing. We retrained and evaluated DiGePred by leaving out either phenotype similarity or all phenotype-related features. There was a decrease in performance on the held-out test set ($p < 1.34 \times 10^{-24}$) (Figure S24); however, the classifiers maintained substantial accuracy and had ROC AUCs $>$ 0.93 and PR AUCs $>$ 0.585. Thus, although our models are most likely somewhat biased by existing knowledge, the strong performance is not only due to overlapping phenotypic annotations.

Using other machine learning approaches and integrating additional features could further improve performance. For example, we have used GO functional annotation enrichment as a way of categorizing our most confident digenic predictions, but GO ontology relationships between the genes would most likely help prioritize potential digenic interactions. Because (PPIs were an indicative feature for DiGePred, protein family and domain similarity, derived from the Pfam[97] database, could be considered as a relevant feature as well. We used a random forest model as it suited our ensemble approach based on many features on disparate scales and limited training data. Alternatively, a support vector machine (SVM) or linear regression approaches could be used with feature normalization. As discussed in the next paragraph, we also believe that approaches that incorporate genetic variants into the prediction are promising; however, the small amount of available training data poses challenges. As more digenic disease pairs are identified, we anticipate that better predictive models will be developed and that these models will yield insight into the genes, pathways, evolutionary histories, and phenotypes associated with digenic disease.

Our approach intentionally separates the prediction of variants' effects on gene function from the identification of gene pairs that could cause disease when their functions are disrupted simultaneously. The focus on gene pairs is reflected in our use of gene-level and gene-pair-level systems biology, biological network, and evolutionary features that represent genes as a whole. The question of whether a variant affects gene function has been studied extensively. There are many methods for interpreting variants of unknown significance,[68,71,98–102] but there is low concordance between them.[103,104] The decoupling of these tasks enables users to apply the approaches they believe to be most appropriate to identify gene pairs of interest before screening for digenic disease potential. For example, in our collaboration with the UDN, this includes application of computational variant effect predictors, study of inheritance patterns, and clinical expertise. Our classifiers perform similarly well whether trained against gene pairs that have predicted disruptive variants or on all variant pairs from individuals (Figures S10–S18), suggesting that they are not simply identifying pairs containing monogenic disease genes. In the future, it may be beneficial to incorporate variant-level and gene-level information into a single algorithm, in particular in cases where there is structural information about the proteins of interest. Indeed, we have had success incorporating 3D modeling of variants and their interactions with the UDN. However, as we describe in the next paragraph, improper incorporation of variant information has potential to cause high false positive rates.

We compared DiGePred to the recently published ORVAL/VarCoPP digenic disease prediction server. This method was also developed with DIDA as positive training data. Because of the challenge of running the web server on a large scale, we were unable to evaluate its performance in our training, validation, test framework. Thus, we applied it to variant gene pairs from the 24 UDN individuals and their 38 unaffected relatives not used in the training or initial evaluation of DiGePred. At its strictest (99%) prediction threshold, we found an average of 855 predicted digenic disease pairs per individual without disease. This false positive rate is too high for clinical use. In contrast, DiGePred predicts two or fewer digenic pairs for

47% of these individuals and an average of under four digenic pairs per individual overall. We also observed that ORVAL predicted a similar fraction of digenic pairs in the unaffected and affected groups at increasingly strict classification score thresholds (Figure S20). Our analysis of the ORVAL method suggests that if one of the genes in a pair carries a variant that is predicted to be pathogenic by ORVAL's variant effect prediction component, then the gene pair is very likely to be predicted to be digenic. This suggests that strong variant-level effects may obscure signals specific to digenic disease.

Going forward, we will continue to refine our approach in collaboration with the UDN and other rare disease cohorts. The approach used to design DiGePred could be expanded to consider oligogenic combinations of greater than two genes. Trigenic and oligogenic cases are beginning to be identified,[105,106] and previous work has identified exclusive gene hubs that cause disease in combination.[107,108] In fact, many previously considered monogenic diseases are now being classified as oligogenic or multigenic with a range of phenotypes depending upon which genes and how many carry variants.[109,110] We also believe that there is the potential to integrate information from large-scale screens of genetic and synthetic lethal interactions in human cell lines and model organisms.[111–116]

In summary, we have developed DiGePred, a method for identifying gene pairs with digenic disease potential, and generated predictions for all pairs of human genes. Our use of this tool on rare-disease-affected individuals illustrates its potential to provide insight in real-world settings, and we anticipate that it will have broad utility in clinical genome interpretation.

Data and code availability

The data and code we used to train and evaluate DiGePred and other models considered are available at https://github.com/CapraLab/DiGePred. The trained DiGePred models are also available in the repository. In addition, digenic pairs from recent literature are provided as Dataset S2. The gene pairs predicted to be digenic above our most confident $F_{0.5}$ threshold are listed in Dataset S3, and the predictions using all models of DiGePred on all human gene pairs are in Datasets S4A–S4D. A website that enables the user to access all DiGePred predictions is available at http://www.meilerlab.org/index.php/servers/show?s_id=28.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.08.010.

Consortia

The members of the Undiagnosed Diseases Network are Maria T. Acosta, David R. Adams, Pankaj Agrawal, Mercedes E. Alejandro, Patrick Allard, Justin Alvey, Ashley Andrews, Euan A. Ashley, Mahshid S. Azamian, Carlos A. Bacino, Guney Bademci, Eva Baker, Ashok Balasubramanyam, Dustin Baldridge, Jim Bale, Deborah Barbouth, Gabriel F. Batzli, Pinar Bayrak-Toydemir, Alan H. Beggs, Gill Bejerano, Hugo J. Bellen, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, David P. Bick, Camille L. Birch, Stephanie Bivona, John Bohnsack, Carsten Bonnenmann, Devon Bonner, Braden E. Boone, Bret L. Bostwick, Lorenzo Botto, Lauren C. Briere, Elly Brokamp, Donna M. Brown, Matthew Brush, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, John Carey, Olveen Carrasquillo, Ta Chen Peter Chang, Hsiao-Tuan Chao, Gary D. Clark, Terra R. Coakley, Laurel A. Cobban, F. Sessions Cole, Heather A. Colley, Cynthia M. Cooper, Heidi Cope, William J. Craigen, Precilla D'Souza, Surendra Dasari, Mariska Davids, Jyoti G. Dayal, Esteban C. Dell'Angelica, Shweta U. Dhar, Naghmeh Dorrani, Daniel C. Dorset, Emilie D. Douine, David D. Draper, Laura Duncan, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Cecilia Esteves, Tyra Estwick, Liliana Fernandez, Carlos Ferreira, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Irman Forghani, Laure Fresard, William A. Gahl, Rena A. Godfrey, Alica M. Goldman, David B. Goldstein, Jean-Philippe F. Gourdine, Alana Grajewski, Catherine A. Groden, Andrea L. Gropman, Melissa Haendel, Neil A. Hanchard, Nichole Hayes, Frances High, Ingrid A. Holm, Jason Hom, Yong Huang, Alden Huang, Rosario Isasi, Fariha Jamal, Yong-Hui Jiang, Jean M. Johnston, Angela L. Jones, Lefkothea Karaviti, Emily G. Kelley, Dana Kiley, David M. Koeller, Isaac S. Kohane, Jennefer N. Kohler, Susan Korrick, Mary E. Koziura, Deborah Krakow, Donna M. Krasnewich, Joel B. Krier, Jennifer E. Kyle, Seema R. Lalani, Byron Lam, Brendan C. Lanpher, Ian R. Lanza, C. Christopher Lau, Pace Laura, Jozef Lazar, Kimberly LeBlanc, Brendan H. Lee, Hane Lee, Roy Levitt, Shawn E. Levy, Richard A. Lewis, Sharyn A. Lincoln, Pengfei Liu, Xue Zhong Liu, Nicola Longo, Sandra K. Loo, Joseph Loscalzo, Richard L. Maas, Calum A. MacRae, Ellen F. Macnamara, Valerie V. Maduro, Marta M. Majcherska, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Rong Mao, Thomas C. Markello, Ronit Marom, Gabor Marth, Beth A. Martin, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Thomas May, Jacob McCauley, Allyn McConkie-Rosell, Colleen E. McCormack, Alexa T. McCray, Thomas O. Metz, Matthew Might, Eva Morava-Kozicz, Paolo M. Moretti, Marie Morimoto, John J. Mulvihill, David R. Murdock, Avi Nath, Stanley F. Nelson, J. Scott Newberry, Sarah K. Nicholas, Donna Novacic, Devin Oglesbee, James P. Orengo, Stephen Pak, J. Carl Pallais, Christina G.S. Palmer, Moretti Paolo, Jeanette C. Papp, Neil H. Parker, Jennifer E. Posey, John H. Postlethwait, Lorraine Potocki, Barbara N. Pusey, Aaron Quinlan, Archana N. Raja, Genecee Renteria, Chloe M. Reuter, Lynette C. Rives, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Robb K. Rowley, Maura Ruzhnikov, Ralph Sacco, Jacinda B. Sampson, Susan L. Samson, Mario Saporta, Judy Schaechter, Timothy Schedl, Kelly Schoch, Daryl A. Scott, Lisa Shakachite, Prashant Sharma, Vandana Shashi, Kathleen Shields, Jimann Shin, Rebecca H. Signer, Catherine H. Sillari, Edwin K. Silverman, Janet S. Sinsheimer, Kathy Sisco, Kevin S. Smith, Lilianna Solnica-Krezel, Rebecca C. Spillmann, Joan M. Stoler, Nicholas Stong, Jennifer A. Sullivan, Shirley Sutton, David A. Sweetser, Holly K. Tabor, Cecelia P. Tamburro, Queenie K.-G. Tan, Mustafa Tekin, Fred Telischi, Willa Thorson, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Tiina K. Urv, Matt Velinder, Dave Viskochil, Tiphanie P. Vogel, Colleen E. Wahl, Melissa Walker, Nicole M. Walley, Chris A. Walsh, Jennifer Wambach, Jijun Wan, Lee-Kai Wang, Michael F. Wangler, Patricia A. Ward, Katrina M. Waters, Bobbie-Jo M. Webb-Robertson, Daniel Wegner, Monte Westerfield, Matthew T. Wheeler, Anastasia L. Wise, Lynne A. Wolfe, Jeremy D. Woods, Elizabeth A. Worthey, Shinya Yamamoto, John Yang, Amanda J. Yoon, Guoyun Yu, Diane B. Zastrow, Chunli Zhao, and Stephan Zuchner.

Acknowledgments

This work was supported by an award from The National Institutes of Health (NIH) Common Fund, through the Office of Strategic Coordination and the Office of the NIH Director, to the clinical sites (U01HG007674 to Vanderbilt University Medical Center). It was also supported by NIH award R35GM127087 (to J.A.C.). This work was conducted in part with the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. We thank members of the Capra, Meiler, and UDN Labs for helpful comments. Finally, we thank all affected individuals and their families.

Declaration of interests

The authors declare no competing interests.

Received: December 23, 2020
Accepted: August 25, 2021
Published: September 15, 2021

References

1. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.
2. Ionita-Laza, I., Makarov, V., Yoon, S., Raby, B., Buxbaum, J., Nicolae, D.L., and Lin, X. (2011). Finding disease variants in Mendelian disorders by using sequence data: methods and applications. Am. J. Hum. Genet. *89*, 701–712.
3. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.
4. Boycott, K.M., Rath, A., Chong, J.X., Hartley, T., Alkuraya, F.S., Baynam, G., Brookes, A.J., Brudno, M., Carracedo, A., Den Dunnen, J.T., et al. (2017). International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am. J. Hum. Genet. *100*, 695–705.
5. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., Mcmillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries (Challenges, and Opportunities).
6. Boycott, K.M., Hartley, T., Biesecker, L.G., Gibbs, R.A., Innes, A.M., Riess, O., Belmont, J., Dunwoodie, S.L., Jojic, N., Lassmann, T., et al. (2019). A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. Cell *177*, 32–37.
7. Gahl, W.A., Mulvihill, J.J., Toro, C., Markello, T.C., Wise, A.L., Ramoni, R.B., Adams, D.R., Tifft, C.J.; and UDN (2016). The NIH Undiagnosed Diseases Program and Network: Applications to modern medicine. Mol. Genet. Metab. *117*, 393–400.
8. Gahl, W.A., Wise, A.L., and Ashley, E.A. (2015). The Undiagnosed Diseases Network of the National Institutes of Health: A National Extension. JAMA *314*, 1797–1798.
9. Ramoni, R.B., Mulvihill, J.J., Adams, D.R., Allard, P., Ashley, E.A., Bernstein, J.A., Gahl, W.A., Hamid, R., Loscalzo, J., McCray, A.T., et al. (2017). The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. Am. J. Hum. Genet. *100*, 185–192.
10. Liu, N., Schoch, K., Luo, X., Pena, L.D.M., Bhavana, V.H., Kukolich, M.K., Stringer, S., Powis, Z., Radtke, K., Mroske, C., et al. (2018). Functional variants in TBX2 are associated with a syndromic cardiovascular and skeletal developmental disorder. Hum. Mol. Genet. *27*, 2454–2465.
11. Chao, H.-T., Davids, M., Burke, E., Pappas, J.G., Rosenfeld, J.A., McCarty, A.J., Davis, T., Wolfe, L., Toro, C., Tifft, C., et al. (2017). A Syndromic Neurodevelopmental Disorder Caused by De Novo Variants in EBF3. Am. J. Hum. Genet. *100*, 128–137.
12. Tokita, M.J., Chen, C.-A., Chitayat, D., Macnamara, E., Rosenfeld, J.A., Hanchard, N., Lewis, A.M., Brown, C.W., Marom, R., Shao, Y., et al. (2018). De Novo Missense Variants in TRAF7 Cause Developmental Delay, Congenital Anomalies, and Dysmorphic Features. Am. J. Hum. Genet. *103*, 154–162.
13. Machol, K., Jankovic, J., Vijayakumar, D., Burrage, L.C., Jain, M., Lewis, R.A., Fuller, G.N., Xu, M., Penas-Prado, M., Gule-Monroe, M.K., et al. (2018). Atypical Alexander disease with dystonia, retinopathy, and a brain mass mimicking astrocytoma. Neurol. Genet. *4*, e248.
14. Marcogliese, P.C., Shashi, V., Spillmann, R.C., Stong, N., Rosenfeld, J.A., Koenig, M.K., Martínez-Agosto, J.A., Herzog, M., Chen, A.H., Dickson, P.I., et al. (2018). IRF2BPL Is Associated with Neurological Phenotypes. Am. J. Hum. Genet. *103*, 245–260.
15. Schoch, K., Meng, L., Szelinger, S., Bearden, D.R., Stray-Pedersen, A., Busk, O.L., Stong, N., Liston, E., Cohn, R.D., Scaglia, F., et al. (2017). A Recurrent De Novo Variant in NACC1 Causes a Syndrome Characterized by Infantile Epilepsy, Cataracts, and Profound Developmental Delay. Am. J. Hum. Genet. *100*, 343–351.
16. Bostwick, B.L., McLean, S., Posey, J.E., Streff, H.E., Gripp, K.W., Blesson, A., Powell-Hamilton, N., Tusi, J., Stevenson, D.A., Farrelly, E., et al. (2017). Phenotypic and molecular characterisation of CDK13-related congenital heart defects, dysmorphic facial features and intellectual developmental disorders. Genome Med. *9*, 73.
17. Küry, S., van Woerden, G.M., Besnard, T., Proietti Onori, M., Latypova, X., Towne, M.C., Cho, M.T., Prescott, T.E., Ploeg, M.A., Sanders, S., et al. (2017). De Novo Mutations in Protein Kinase Genes CAMK2A and CAMK2B Cause Intellectual Disability. Am. J. Hum. Genet. *101*, 768–788.
18. Pomerantz, D.J., Ferdinandusse, S., Cogan, J., Cooper, D.N., Reimschisel, T., Robertson, A., Bican, A., McGregor, T., Gauthier, J., Millington, D.S., et al. (2018). Clinical heterogeneity of mitochondrial NAD kinase deficiency caused by a *NADK2* start loss variant. Am. J. Med. Genet. A. *176*, 692–698.
19. Oláhová, M., Yoon, W.H., Thompson, K., Jangam, S., Fernandez, L., Davidson, J.M., Kyle, J.E., Grove, M.E., Fisk, D.G., Kohler, J.N., et al. (2018). Biallelic Mutations in ATP5F1D, which Encodes a Subunit of ATP Synthase, Cause a Metabolic Disorder. Am. J. Hum. Genet. *102*, 494–504.
20. Johnston, J.J., van der Smagt, J.J., Rosenfeld, J.A., Pagnamenta, A.T., Alswaid, A., Baker, E.H., Blair, E., Borck, G., Brinkmann, J., Craigen, W., et al. (2018). Autosomal recessive Noonan syndrome associated with biallelic LZTR1 variants. Genet. Med. *20*, 1175–1185.
21. Poli, M.C., Ebstein, F., Nicholas, S.K., de Guzman, M.M., Forbes, L.R., Chinn, I.K., Mace, E.M., Vogel, T.P., Carisey, A.F., Benavides, F., et al. (2018). Heterozygous Truncating Variants in POMP Escape Nonsense-Mediated Decay and Cause a

Unique Immune Dysregulatory Syndrome. Am. J. Hum. Genet. *102*, 1126–1142.
22. Auer, F., Lin, M., Nebral, K., Gertzen, C.G.W., Haas, O.A., Kuhlen, M., Gohlke, H., Izraeli, S., Trka, J., Hu, J., et al. (2018). Novel Recurrent Germline JAK2 G571S Variant in Childhood Acute B-Lymphoblastic Leukemia: A Double Hit One Pathway Scenario. Blood *132*, 387.
23. Pehlivan, D., Bayram, Y., Gunes, N., Coban Akdemir, Z., Shukla, A., Bierhals, T., Tabakci, B., Sahin, Y., Gezdirici, A., Fatih, J.M., et al. (2019). The Genomics of Arthrogryposis, a Complex Trait: Candidate Genes and Further Evidence for Oligogenic Inheritance. Am. J. Hum. Genet. *105*, 132–150.
24. Badano, J.L., and Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. Nat. Rev. Genet. *3*, 779–789.
25. Van Heyningen, V., and Yeyati, P.L. (2004). Mechanisms of non-Mendelian inheritance in genetic disease. Hum. Mol. Genet. *13*, R225–R233.
26. Kajiwara, K., Berson, E.L., and Dryja, T.P. (1994). Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. Science *264*, 1604–1608.
27. Schäffer, A.A. (2013). Digenic inheritance in medical genetics. J. Med. Genet. *50*, 641–652.
28. Gazzo, A.M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., Smits, G., and Lenaerts, T. (2016). DIDA: A curated and annotated digenic diseases database. Nucleic Acids Res. *44* (D1), D900–D907.
29. Lupski, J.R. (2012). Digenic inheritance and Mendelian disease. Nat. Genet. *44*, 1291–1292.
30. Deltas, C. (2018). Digenic inheritance and genetic modifiers. Clin. Genet. *93*, 429–438.
31. Gazzo, A., Raimondi, D., Daneels, D., Moreau, Y., Smits, G., Van Dooren, S., and Lenaerts, T. (2017). Understanding mutational effects in digenic diseases. Nucleic Acids Res. *45*, e140.
32. Papadimitriou, S., Gazzo, A., Versbraegen, N., Nachtegael, C., Aerts, J., Moreau, Y., Van Dooren, S., Nowé, A., Smits, G., and Lenaerts, T. (2019). Predicting disease-causing variant combinations. Proc. Natl. Acad. Sci. USA *116*, 11878–11887.
33. Boudellioua, I., Kulmanov, M., Schofield, P.N., Gkoutos, G.V., and Hoehndorf, R. (2018). OligoPVP: Phenotype-driven analysis of individual genomic information to prioritize oligogenic disease variants. Sci. Rep. *8*, 14681.
34. Renaux, A., Papadimitriou, S., Versbraegen, N., Nachtegael, C., Boutry, S., Nowé, A., Smits, G., and Lenaerts, T. (2019). ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. Nucleic Acids Res. *47* (W1), W93–W98.
35. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45* (D1), D353–D361.
36. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46* (D1), D649–D655.
37. Jaccard, P. (1912). THE Dros. Inf. Serv.TRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytol. *11*, 37–50.
38. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Res. *45* (D1), D865–D876.
39. Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Res. *43*, D82–D86.
40. Poon, H., Quirk, C., DeZiel, C., and Heckerman, D. (2014). Literome: PubMed-scale genomic knowledge base in the cloud. Bioinformatics *30*, 2840–2842.
41. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, D497–D501.
42. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., and Mewes, H.W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. *36*, D646–D650.
43. Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) *2010*, baq023.
44. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. *45* (D1), D362–D368.
45. Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell *167*, 657–669.e21.
46. Capra, J.A., Williams, A.G., and Pollard, K.S. (2012). ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput. Biol. *8*, e1002567.
47. Chen, W.-H., Minguez, P., Lercher, M.J., and Bork, P. (2012). OGEE: an online gene essentiality database. Nucleic Acids Res. *40*, D901–D906.
48. Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M., and Bork, P. (2017). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res. *45* (D1), D940–D944.
49. Fadista, J., Oskolkov, N., Hansson, O., and Groop, L. (2017). LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics *33*, 471–474.
50. Matsuya, A., Sakate, R., Kawahara, Y., Koyanagi, K.O., Sato, Y., Fujii, Y., Yamasaki, C., Habara, T., Nakaoka, H., Todokoro, F., et al. (2008). Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. Nucleic Acids Res. *36*, D787–D792.
51. Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. *6*, e1001154.
52. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
53. Schrauwen, I., Chakchouk, I., Acharya, A., Liaqat, K., Irfanullah, Nickerson, D.A., Bamshad, M.J., Shah, K., Ahmad, W., Leal, S.M.; and University of Washington Center for

Mendelian Genomics (2018). Novel digenic inheritance of PCDH15 and USH1G underlies profound non-syndromic hearing impairment. BMC Med. Genet. *19*, 122.
54. Ameratunga, R., Koopmans, W., Woon, S.-T., Leung, E., Lehnert, K., Slade, C.A., Tempany, J.C., Enders, A., Steele, R., Browett, P., et al. (2017). Epistatic interactions between mutations of TACI (*TNFRSF13B*) and *TCF3* result in a severe primary immunodeficiency disorder and systemic lupus erythematosus. Clin. Transl. Immunology *6*, e159.
55. Hoyos-Bachiloglu, R., Chou, J., Sodroski, C.N., Beano, A., Bainter, W., Angelova, M., Al Idrissi, E., Habazi, M.K., Alghamdi, H.A., Almanjomi, F., et al. (2017). A digenic human immunodeficiency characterized by IFNAR1 and IFNGR2 mutations The Journal of Clinical Investigation. J. Clin. Invest. *127*, 4415–4420.
56. Abdallah, A.M., Carlus, S.J., Al-Mazroea, A.H., Alluqmani, M., Almohammadi, Y., Bhuiyan, Z.A., and Al-Harbi, K.M. (2019). Digenic inheritance of LAMA4 and MYH7 mutations in patient with infantile dilated cardiomyopathy. Medicina (Kaunas) *55*, 1–10.
57. Heida, A., van der Does, L.J.M.E., Ragab, A.A.Y., and de Groot, N.M.S. (2019). A Rare Case of the Digenic Inheritance of Long QT Syndrome Type 2 and Type 6. Case Rep. Med. *2019*, 1384139.
58. Kong, Y., Xu, K., Yuan, K., Zhu, J., Gu, W., Liang, L., and Wang, C. (2019). Digenetic inheritance of SLC12A3 and CLCNKB genes in a Chinese girl with Gitelman syndrome. BMC Pediatr. *19*, 114.
59. Nieto-Marín, P., Jiménez-Jáimez, J., Tinaquero, D., Alfayate, S., Utrilla, R.G., Rodríguez Vázquez del Rey, M.D.M., Perin, F., Sarquella-Brugada, G., Monserrat, L., Brugada, J., et al. (2019). Digenic Heterozigosity in SCN5A and CACNA1C Explains the Variable Expressivity of the Long QT Phenotype in a Spanish Family. Rev. Española Cardiol. *72*, 324–332.
60. Stone, S.I., Wegner, D.J., Wambach, J.A., Cole, F.S., Ornitz, D.M., and Urano, F. (2019). 26-OR: Digenic FGFR1/KLB Variants Associated with Endocrine Specific FGF-21 Signaling Defects and Extreme Insulin Resistance. Diabetes *68*, 26.
61. Nozu, K., Inagaki, T., Fu, X.J., Nozu, Y., Kaito, H., Kanda, K., Sekine, T., Igarashi, T., Nakanishi, K., Yoshikawa, N., et al. (2008). Molecular analysis of digenic inheritance in Bartter syndrome with sensorineural deafness. J. Med. Genet. *45*, 182–186.
62. Yang, Y., Ye, W., Guo, J., Zhao, L., Tu, M., Zheng, Y., and Li, L. (2019). CLCN7 and TCIRG1 mutations in a single family: Evidence for digenic inheritance of osteopetrosis. Mol. Med. Rep. *19*, 595–600.
63. Zastrow, D.B., Zornio, P.A., Dries, A., Kohler, J., Fernandez, L., Waggott, D., Walkiewicz, M., Eng, C.M., Manning, M.A., Farrelly, E., et al. (2017). Exome sequencing identifies de novo pathogenic variants in *FBN1* and *TRPS1* in a patient with a complex connective tissue phenotype. Cold Spring Harb. Mol. Case Stud. *3*, a001388.
64. Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics *8*, 25.
65. Breiman, L. (2004). Consistency for a simple model of random forests. Tech. Rep. 670. (Stat. Dep. Univ. Calif. Berkeley).
66. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
67. (2016). ExAC project pins down rare gene variants. Nature *536*, 249.
68. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
69. Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452-7.
70. Vaser, R., Adusumalli, S., Ngak Leng, S., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nat. Protoc. *11*, 1–9.
71. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47* (D1), D886–D894.
72. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.
73. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. *47* (W1), W199–W205.
74. Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D'Eustachio, P., and Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers (Basel) *4*, 1180–1211.
75. Park, Y., and Marcotte, E.M. (2012). Flaws in evaluation schemes for pair-input computational predictions. Nat. Methods *9*, 1134–1136.
76. Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*, aaf1420.
77. Vucic, S., Rothstein, J.D., and Kiernan, M.C. (2014). Advances in treating amyotrophic lateral sclerosis: insights from pathophysiological studies. Trends Neurosci. *37*, 433–442.
78. Farg, M.A., Soo, K.Y., Warraich, S.T., Sundaramoorthy, V., Blair, I.P., and Atkin, J.D. (2020). Erratum: Ataxin-2 interacts with FUS and intermediate-length polyglutamine expansions enhance FUS-related pathology in amyotrophic lateral sclerosis. Hum. Mol. Genet. *29*, 703–704.
79. Ostrowski, L.A., Hall, A.C., and Mekhail, K. (2017). Ataxin-2: From RNA Control to Human Health and Disease. Genes (Basel) *8*, 157.
80. Mikhael, S., Dugar, S., Morton, M., Chorich, L.P., Tam, K.B., Lossie, A.C., Kim, H.-G., Knight, J., Taylor, H.S., Mukherjee, S., et al. (2021). Genetics of agenesis/hypoplasia of the uterus and vagina: narrowing down the number of candidate genes for Mayer-Rokitansky-Küster-Hauser Syndrome. Hum. Genet. *140*, 667–680.
81. Morcel, K., Camborieux, L., Guerrier, D.; and Programme de Recherches sur les Aplasies Müllériennes (2007). Mayer-Rokitansky-Küster-Hauser (MRKH) syndrome. Orphanet J. Rare Dis. *2*, 13.
82. Patnaik, S.S., Brazile, B., Dandolu, V., Ryan, P.L., and Liao, J. (2015). Mayer-Rokitansky-Küster-Hauser (MRKH) syndrome: a historical perspective. Gene *555*, 33–40.
83. Jenkins, M.H., Alrowaished, S.S., Goody, M.F., Crawford, B.D., and Henry, C.A. (2016). Laminin and Matrix

metalloproteinase 11 regulate Fibronectin levels in the zebrafish myotendinous junction. Skelet. Muscle *6*, 18.
84. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798.
85. McKusick, V.A. (2007). Mendelian Inheritance in Man and its online version, OMIM. Am. J. Hum. Genet. *80*, 588–604.
86. Amberger, J., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. *37*, D793–D796.
87. Yujian, L., and Bo, L. (2007). A normalized Levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. *29*, 1091–1095.
88. Soboloff, J., Spassova, M.A., Tang, X.D., Hewavitharana, T., Xu, W., and Gill, D.L. (2006). Orai1 and STIM reconstitute store-operated calcium channel function. J. Biol. Chem. *281*, 20661–20665.
89. Lewis, R.S. (2001). CALCIUM SIGNALING MECHANISMS IN T LYMPHOCYTES. Annu. Rev. Immunol. *19*, 497–521.
90. Partisetis, M., Le Deist, F., Hivroz, C., Fischer, M., Korns, H., and Choquets, D. (1994). The Calcium Current Activated by T Cell Receptor and Store Depletion in Human Lymphocytes Is Absent in a Primary Immunodeficiency. J. Biol. Chem. *269*, 32327–32335.
91. Lioudyno, M.I., Kozak, J.A., Penna, A., Safrina, O., Zhang, S.L., Sen, D., Roos, J., Stauderman, K.A., Cahalan, M.D., and Tsien, R.Y. (2008). Orai1 and STIM1 move to the immunological synapse and are up-regulated during T cell activation. Proc. Natl. Acad. Sci. *105*, 2011–2016.
92. Lacruz, R.S., and Feske, S. (2015). Diseases caused by mutations in *ORAI1* and *STIM1*. Ann. N Y Acad. Sci. *1356*, 45–79.
93. McCarl, C.A., Picard, C., Khalil, S., Kawasaki, T., Röther, J., Papolos, A., Kutok, J., Hivroz, C., Ledeist, F., Plogmann, K., et al. (2009). ORAI1 deficiency and lack of store-operated Ca2+ entry cause immunodeficiency, myopathy, and ectodermal dysplasia. J. Allergy Clin. Immunol. *124*, 1311–1318.e7.
94. Parry, D.A., Holmes, T.D., Gamper, N., El-Sayed, W., Hettiarachchi, N.T., Ahmed, M., Cook, G.P., Logan, C.V., Johnson, C.A., Joss, S., et al. (2015). A homozygous STIM1 mutation impairs store-operated calcium entry and natural killer cell effector function without clinical immunodeficiency. J. Allergy Clin. Immunol. *137*, 955–957.e8.
95. Picard, C., Mccarl, C.-A., Papolos, A., Khalil, S., Lüthy, K., Hivroz, C., Ledeist, F., Rieux-Laucat, F., Rechavi, G., Rao, A., et al. (2009). STIM1 Mutation Associated with a Syndrome of Immunodeficiency and Autoimmunity. N. Engl. J. Med. *360*, 1971–1980.
96. Feske, S., Gwack, Y., Prakriya, M., Srikanth, S., Puppel, S.-H., Tanasa, B., Hogan, P.G., Lewis, R.S., Daly, M., and Rao, A. (2006). A mutation in Orai1 causes immune deficiency by abrogating CRAC channel function. Nature *441*, 179–185.
97. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49* (D1), D412–D419.
98. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.
99. Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information (Valdar and Thornton).
100. Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. (2013). ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. Israel Journal of Chemistry. https://doi.org/10.1002/ijch.201200096.
101. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. *38*, W529–W533.
102. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
103. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. *24*, 2125–2137.
104. Castellana, S., and Mazza, T. (2013). Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. Brief Bioinform. *14*, 448–459.
105. Gifford, C.A., Ranade, S.S., Samarakoon, R., Salunga, H.T., Yvanka De Soysa, T., Huang, Y., Zhou, P., Elfenbein, A., Wyman, S.K., Bui, Y.K., et al. (2019). Oligogenic inheritance of a human heart disease involving a genetic modifier. Science *364*, 865–870.
106. Chen, Y., Barajas-Martinez, H., Zhu, D., Wang, X., Chen, C., Zhuang, R., Shi, J., Wu, X., Tao, Y., Jin, W., et al. (2017). Novel trigenic CACNA1C/DES/MYPN mutations in a family of hypertrophic cardiomyopathy with early repolarization and short QT syndrome. J. Transl. Med. *15*, 78.
107. Duerinckx, S., Jacquemin, V., Drunat, S., Vial, Y., Passemard, S., Perazzolo, C., Massart, A., Soblet, J., Racapé, J., Desmyter, L., et al. (2020). Digenic inheritance of human primary microcephaly delineates centrosomal and non-centrosomal pathways. Hum. Mutat. *41*, 512–524.
108. Yao, Q., Li, E., and Shen, B. (2019). Autoinflammatory disease with focus on NOD2-associated disease in the era of genomic medicine. Autoimmunity *52*, 48–56.
109. Wallace, M.J., El Refaey, M., Mesirca, P., Hund, T.J., Mangoni, M.E., and Mohler, P.J. (2021). Genetic Complexity of Sinoatrial Node Dysfunction. Front. Genet. *12*, 654925.
110. Monasky, M.M., Micaglio, E., Ciconte, G., and Pappone, C. (2020). Brugada Syndrome: Oligogenic or Mendelian Disease? Int. J. Mol. Sci. *21*, 1687.
111. Nijman, S.M.B. (2011). Synthetic lethality: general principles, utility and detection using genetic screens in human cells. FEBS Lett. *585*, 1–6.
112. Srivas, R., Shen, J.P., Yang, C.C., Sun, S.M., Li, J., Gross, A.M., Jensen, J., Licon, K., Bojorquez-Gomez, A., Klepper, K., et al. (2016). A Network of Conserved Synthetic Lethal Interactions for Exploration of Precision Cancer Therapy. Mol. Cell *63*, 514–525.

113. O'Neil, N.J., Bailey, M.L., and Hieter, P. (2017). Synthetic lethality and cancer. Nat. Rev. Genet. *18*, 613–623.
114. Guo, J., Liu, H., and Zheng, J. (2016). SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. Nucleic Acids Res. *44* (D1), D1011–D1017.
115. Gong, X., Du, J., Parsons, S.H., Merzoug, F.F., Webster, Y., Iversen, P.W., Chio, L.-C., Van Horn, R.D., Lin, X., Blosser, W., et al. (2018). Aurora A Kinase Inhibition Is Synthetic Lethal with Loss of the RB1 Tumor Suppressor Gene. Cancer Discov. *9*, 248–263.
116. Li, X., O'neil, N.J., Moshgabadi, N., and Hieter, P. (2014). Synthetic Cytotoxicity: Digenic Interactions with TEL1/ATM Mutations Reveal Sensitivity to Low Doses of Camptothecin. Genetics *197*, 611–623.