PNAS
www.pnas.org

Supplementary Information for

Predicting disease-causing variant combinations

Sofia Papadimitriou, Andrea Gazzo, Nassim Versbraegen, Charlotte Nachtegael, Jan Aerts, Yves Moreau, Sonia Van Dooren, Ann Nowé, Guillaume Smits* and Tom Lenaerts*

Corresponding authors:
Tom Lenaerts: tlenaert@ulb.ac.be
Guillaume Smits: guillaume.smits@erasme.ulb.ac.be

This PDF file includes:

Supplementary Text S1 to S5
Figs. S1 to S6
Tables S1 to S10
References for SI reference citations

Other supplementary materials for this manuscript include the following:
Datasets S1 to S5

Supplementary Information Text

Text S1. Supplementary information for the curation process of the bi-locus combinations in DIDA

In order to include a bi-locus combination in DIDA as pathogenic, the authors responsible for creating DIDA applied criteria that are based on their relevance and the existence of different levels of evidence for their pathogenicity.

For the first version of DIDA, 54 out of 108 scientific articles that were published before January 2013 and were listed in the work of Schäffer *et al.*(1), were manually selected to be included in the database. From these articles, 125 bi-locus combinations were extracted. A Pubmed search with the keyword "*digenic*" was then conducted to include medical papers published until June 2015, leading to the inclusion of 88 another digenic combinations from 28 different publications. This first version of DIDA (DIDAv1) corresponds to the **training set of VarCoPP**. Additional bi-locus combinations for developing the second version of DIDA (DIDAv2) (where many of them were used as a **validation set for VarCoPP**) were obtained through a Pubmed search with the keyword "*digenic*" to retrieve publications between July 2015 and June 2017, adding 45 new digenic combinations to the database.

Only combinations derived from **clinical studies** were accepted, and not those derived from statistical or predictive methods. For these combinations, three different types of evidence or substantial pathogenicity proof were assessed, according to those described by Schäffer *et al*.:

- **Evidence of a protein-protein or a protein-DNA interaction** for the two genes or proteins involved in the bi-locus combination or whether there is a combined effect of the variants at the functional level (referenced in DIDA as "**Functional evidence**"). This type of evidence highlights experimentally the pathogenic effect of the variant combination as opposed to the single effect of the involved variants, and to assess whether this effect is related to the observed phenotype in the patient carrying the particular variant combination. It was also required that the study was done in human cells, so those conducted on animal models were not accepted.
- **Phenotypic difference in the studied family** according to the segregation of the bi-locus combination (referenced in DIDA as "**Familial evidence**"). An ideal evidence would be the involvement of families with extended pedigrees, where we can compare the phenotype of the affected individuals having the bi-locus combination, that of individuals carrying one of the involved variants, and of individuals carrying none of the variants. A bi-locus combination derived from a pedigree analysis involving only one patient and their parents needed to have one of the other types of evidence, in order to be accepted.
- **Indirect evidence** based on whether the products of the two genes are involved in the same pathway, are co-expressed or whether the two genes are implicated in the same disease (referenced in DIDA as different types of "**Gene relationship**").

As additional criteria, only bi-locus combinations involving a maximum of four mutated alleles were included in DIDA. These contain single nucleotide variants and small indels. Copy number variations (CNVs) and repeats are actually excluded, as well as genetic diseases where environmental factors are suspected to take part in the development of the disease phenotype, or where more than two genes are potentially involved in the phenotype. Several plots on statistics regarding the type of data included in DIDA are present in the "Statistics" page of the database (http://dida.ibsquare.be/statistics/).

Text S2. Supplementary information for overlapping bi-locus combinations between DIDA and the 1KGP

During an initial screening of bi-locus combinations in the 1000 Genomes Project (1KGP), we discovered at that time 7 bi-locus combinations of DIDA that were also present in 14 individuals of the 1KGP. None of the studies analysing these combinations presented a comparison with the 1KGP. The bi-locus combinations were present in parents or grandparents of 1KGP families and no relative carried extra DIDA bi-locus combinations.

We detected 2 bi-locus combinations leading to Kallman syndrome both derived from the same study(2), which were also present in 5 individuals of the 1KGP. These combinations involved the gene pairs *KISS1R - PROKR2* (dd202), which had already been reported to be relevant for digenicity(3), and *GNRHR - PROKR2* (dd203). Although *GNRHR* has been involved in digenic cases with other genes (*PROK2* and *FGFR1*) leading to congenital hypogonadotropic hypogonadism(4-6), no other indications of digenicity have been reported between this gene and *PROKR2*. The c.719G>A *GNRHR* heterozygous variant of dd203 has been described previously to be involved in mild symptoms of congenital hypogonadotropic hypogonadism(7). On the other hand the c.991G>A *PROKR2* heterozygous variant of dd203 has already been reported to be implicated in Kallman syndrome with variable phenotypes(8). In the current study, the authors compared cases with a small control cohort of 14 individuals carrying mutations of either *KAL1* and *PROKR2*, therefore did not explicitly check for bi-locus combinations in the control cohort. The bi-locus combinations presented are also not supported by familial or functional evidence.

We also detected in 2 individuals of the 1KGP a bi-locus combination (dd220) leading to maturity-onset diabetes of the young(9) that involved a novel digenic pair *NEUROD1-PDX1*. Both mutations in these genes are heterozygous, in protein-coding areas. In general, it is known that the clinical features of MODY can overlap with those of polygenic diabetes, and several genes have been associated so far with this disease. For the dd220 combination, this novel gene pair can be relevant for MODY, as both genes are required for β-cell development, growth and insulin gene expression participating in a transcription complex that is important for short-range DNA looping(10). Both heterozygous c.723C>G *NEUROD1* and c.670G>A *PDX1* mutations of that bi-locus combination had already known to be individually implicated in forms of obesity and diabetes, respectively(11,12). However, the patient that carried this bi-locus combination was not obese, but showed a serious decrease of insulin intake compared to controls. The c.670G>A *PDX1* variant can present incomplete oligogenic penetrance(12) - also suggested by its frequency in the ExAC database (0.002113) - and, thus, sometimes can be overlooked. At the same time, although studies have shown a contradictory importance of *NEUROD1* with the associated clinical phenotypes being unclear, microRNA analysis showed that silencing this gene led to loss of β-cell proliferation, through the overexpression of miR-24(13). Although the study that presented this combination included a familial analysis, they authors did not provide functional evidence. Furthermore, in the study the authors performed a genetic comparison with a small cohort of 60 control individuals with normal glucose intake, something that could maybe have led to loss of statistical power to show the presence of the bi-locus combination in the control cohort. In general, the relevance of this bi-locus combination may not be unimportant, but it should be noted that for MODY other less-known molecular mechanisms may be involved that make this bi-locus combination present in two individuals of the 1KGP.

Finally, we detected 4 novel bi-locus combinations (dd180, dd188, dd193 and dd196) leading to familial hemophagocytic lymphohistiocytosis (FHL) in 7 individuals, which were all derived from the same study(14). It is worth noticing that one 1KGP individual carried two of these combinations: dd188 and dd193. The variant combinations involve heterozygous

mutations in *PRF1*, *UNC13D* and *STXBP2* genes, with three of them (dd180, dd188, dd193) sharing the variant c.272C>T at *PRF1* gene. All of these genes are involved in the lymphocyte cytotoxicity activity for cytotoxic granules and *UNC13D* along with *STXBP2* belong also in the same cytotoxic pathway(15), while *PRF1* assists at a later stage to the delivery and penetration of granzymes(16).
More specifically:

- Combination dd196 involves an intronic c.*12G>A *STXBP2* variant and a protein-coding c.2896C>T *UNC13D* variant. These genes belong in the same degranulation pathway and the authors show that patients with variants in genes of that pathway had earlier age of onset and defective degranulation, compared to those who carried variants in *PRF1* and another gene. With these findings the authors suggest a synergistic deleterious effect of the involved mutations that supports the notion of a digenic inheritance for FHL. Digenic implications of this pair have also been suggested in further murine and human studies(17,18).
- Combination dd180 involves two protein coding c.272C>T *PRF1* and c.3160A>G *UNC13D* variants. On the other hand, combinations dd188 and dd193 involve the gene pair *PRF1 - STXBP2*. Combination dd188 involves two protein-coding variants c.272C>T in *PRF1* and c.1034C>T in *STXBP2*, while dd193 involves the *PRF1* c.272C>T variant and one intronic variant c.795-4C>T in *STXBP2*. The authors present that patients with one mutation in the *PRF1* gene and another mutation at one of the genes of the granulation pathway had later age of onset compared to those carrying mutations in genes of the same cytotoxic pathway and normal degranulation, although they showed decreased perforin expression. Nevertheless, digenic variants in the pair *PRF1 - UNC13D*, but not the exact dd180 combination, have been reported in another later study(18).

As the genes involved in that study are implicated in FHL and digenicity has been suggested in further studies, we cannot conclude that the specific variant combinations that have an overlap between DIDA and 1KGP are not relevant, although in the paper there is a stronger indication of digenicity for the dd196 combination, compared to the rest. It should be noted, however, that the study did not present a familial analysis, neither further functional evidence concerning the digenic cases. Moreover, the authors did not compare their cases with controls or other variant databases for further proof of relevance, but performed a clinical analysis.

Text S3. Interpretation of validation cases predicted as neutral

Among the 23 independent variant combinations of the independent validation set, we predicted 3 of them as neutral (Testpos_4, Testpos_10 and Testpos_15). We provide here our interpretation of why we fail to predict them as disease-causing. In general, this can be attributed to missing values of some features for the gene recessiveness or haploinsufficiency, low CADD score values and high gene haploinsufficiency values in some cases. The annotated features for this data set can be further studied at the *SI Appendix*, Dataset S1.

1. Testpos_4, which was predicted with SS = 46.2, carrying novel inherited variants in the *PSMB4-PSMB9* gene pair, is involved in the development of CANDLE syndrome, an autoinflammatory disease, and it was detected in two siblings of a familial study. The authors state that the heterozygous c.44_45insG mutation in *PSB4* causes reduced gene expression, while the second missense c.494G>A mutation in *PSMB9* affects a highly conserved protein residue. Both *PSMB4* and *PSMB9* were predicted to be haploinsufficient in our data. Recessiveness probability of *PSMB9* was not known and therefore we used a median value, something that may have affected the results, as it seems that this feature (RecB), along with the CADD score of the *PSMB9* variant allele (CADD3) that is low in this case, are important for classification. Therefore, we see that this unknown value along with the low value of CADD3 can have an impact on the predictions.
2. Testpos_10, which was predicted with SS = 24, contains variants in the *COL4A4-COL4A3* gene pair known in general to be implicated in Alport syndrome, the first one being a heterozygous in-frame c.1293_1310del mutation and the latter being a splicing c.1504+1G>A variant that is predicted to cause the loss of the 5’ splice site. The patient carrying this bi-locus combination showed symptoms for heamaturia and proteinuria, but no hearing loss or ocular lesions. In the paper, the authors show that none of the monogenic predictors (SIFT, MutationTaster, Polyphen2) was able to provide a pathogenicity score for these variants. Looking at the CADD values, we saw that these were actually low (2.36 and 2.84). These low scores most probably guide the prediction towards the neutral class.
3. Testpos_15, which was predicted with SS = 1.6, originated from a female patient who showed symptoms of intermittent heamaturia and proteinuria and not hearing loss or ocular lessions, therefore not severe symptoms of the disease. It contains three heterozygous variants, one in the *COL4A5* gene and two in the *COL4A4* gene, both known to be involved in Alport syndrome. Males carrying mutations in this pair show more severe symptoms of the disease, as the *COL4A5* gene lies in the X-chromosome. In this case, although the CADD score of the *COL4A5* variant is relatively high, the very low CADD scores for the *COL4A4* variants (1.70 and -0.22) and the high *COL4A4* haploinsufficiency probability value (0.84) seem to guide the prediction of this bi-locus combination towards the neutral class. We would like to note that in our method, we represent a hemizygous X-linked variant in males similarly to a homozygous variant, to depict the fact that there is no wild-type allele present to compensate in certain cases for the gene and protein function, and thus, this variant, and consequently the bi-locus combination, can be predicted to have a stronger impact in males than females.

Text S4. Interpretation of PR curves for confidence zones

As the amount of neutral bi-locus combinations tested increases, the precision drops (*SI Appendix*, Fig. S2*A*), as expected. Clearly, the more combinations need to be checked the more the absolute number of elements in the two confidence zones will increase. Figure S2*B* in *SI Appendix* shows the precision and recall of the 95%-zone and the 99%-zone when VarCoPP is tested on the elements of the positive validation set together with the elements of a collection of either 100, 1000 and 10000 neutral combinations that fall into these zones. As before, increasing the number of neutral cases decreases the precision: the precision of finding all 20 TPs in the 95%-zone drops from ~80% to ~30% when increasing the amount of bi-locus combinations from 100 to 1000. The more stringent 99%-zone improves precision and recall significantly (*SI Appendix*, Fig. S2*B*), yet at the cost of missing some of the real culprits (*SI Appendix*, Table S9).

Text S5. Supplementary Materials and Methods

- **Data collection, filtering and annotation**

We used the first version of the Digenic Diseases Database (DIDAv1) as the disease-causing data set and collected information on the bi-locus combinations, genes and variants. 213 bi-locus combinations were present in the database and included single nucleotide variations (SNVs) and small insertions and deletions (indels). These combinations contributed to 44 different diseases. To create the control bi-locus combinations, we used the variant data of the 1000 Genomes Project (1KGP) of Phase 3. The 1KGP contains a broad representation of the human genetic variation, including variants of 2,504 healthy individuals from 26 different populations. For computational reasons, we selected a random 25% subset of the human proteome from the 1KGP, using the list of the curated human proteome from Uniprot(19) and filtered the variants using Highlander (http://sites.uclouvain.be/highlander/). This filtering enabled us to limit the amount of gene pairs and bi-locus combinations present in the control set, in order to be similar to the type of disease-causing combinations composing DIDA.

The filtering process took place in such a way that both sets would contain comparable variants and genes. We used the Ensembl BioMart tool(20) for the Grch37.p13 version of the human genome to obtain information about exon positions and 1KGP Minor Allele Frequency (MAF). We included variants of MAF<=3%, as this variant frequency threshold represents the vast majority of variants in DIDAv1 and in this way we could limit potential noise of the abundance of more common variants from the control dataset. We included exonic SNVs and small indels, as well as intronic variants with a maximum distance of 13 nucleotides (nt) from the exon edge. We also collected synonymous variants close to splicing sites in a maximum distance of 6 nt from the exon boundary, according to the synonymous variants present in DIDAv1. At the gene level, we selected only protein-coding genes from 1KGP, as this type of genes was present in DIDAv1. To further ensure the presence of true protein-coding genes in the control set, we further filtered the genes based on the protein-coding genes of the consensus coding sequence (CCDS) project(21). Finally, we removed from the 1KGP data set 14 individuals who carried disease-causing bi-locus combinations (Fig. 2, and *SI Appendix*, Table S3), as well as the 7 corresponding bi-locus combinations. After the filtering process, 200 disease-causing bi-locus DIDAv1 combinations remained, whereas around 8,000,000 unique gene pairs were present in the 1KGP control set leading to billions of possible combinations to choose.

We annotated both sets based on information at the variant, gene and gene pair level. A summary of the features and a brief explanation is displayed in *SI Appendix*, Table S4 and shown schematically in Fig. 3C. We used the combined annotation dependent depletion (CADD) score(22) as a single-variant deleteriousness metric because it can predict the damaging effect of not only missense variants, but also splicing, nonsense variants, as well as small deletions and insertions and shows good overall performance. We implemented an in-house code to calculate the flexibility and hydrophobicity differences between the wild-type and mutated amino acids, using the flexibility variation scale of Bhaskaran & Ponnuswamy(23) and the Wimley & White whole residue hydrophobicity scale(24) respectively. We also extracted information from Pfam(25) using the Ensembl BioMart tool for the Grch37.p13 version of the human genome(20) to predict whether the variant alleles lie in a conserved protein family domain. For the gene features, we used the gene haploinsufficiency(26) and recessiveness probability(27) retrieved from dbNSFP2.8(28,29). Finally, as a gene pair feature we exploited the biological distance, a metric of biological relatedness in terms of protein-protein interactions between two genes, obtained with the script provided by the Human Gene Connectome tool(30).

- **Representation of a bi-locus variant combination**

We represented a bi-locus variant combination as a vector of categorical and numerical features (see Fig. 3B and *SI Appendix*, Table S4 and Table S5 for numerical representation). In general, a bi-locus combination always contained four different alleles (2 for gene A and 2 for gene B), including wild-type alleles. This was done in accordance with the type of information in DIDA, where for each bi-locus combination we had maximum two mutated alleles in each gene. Therefore, we encoded each variant-related feature four times (four dimensions), each one representing a different allele. With this representation, we also considered the zygosity of the variants, meaning that if a variant was homozygous in one gene, then both alleles corresponding to this gene would contain the same information. We encoded gene-related features using two dimensions, one for each gene involved in the bi-locus combination. In the end, a set of vectors was created where each vector represented a bi-locus combination and each element in the vector represented a unique feature of that combination.

To ensure a fair prediction process, we defined the order of variants and genes inside each bi-locus combination (Fig. 3B) in the same way for both data sets. We used the Gene Damage Index (GDI)(31) to determine the order of the two genes in a bi-locus combination, so that gene A was the gene with the lowest GDI index and, thus, has with a higher probability to be associated with a disease. We also ordered different variant alleles inside the same gene (in cases of heterozygous-compound variants in gene A or gene B) using the CADD raw score(22), so that the first variant allele in that gene would be the one with the highest score.

- **Stratification and sampling of the 1KGP neutral data for training**

The 1KGP neutral set contained a huge number of bi-locus combinations compared to DIDA, leading to a class imbalance problem (Fig. 3D). A recent work has shown that Random Forests (RFs)(32) can outperform other classifiers even when used as base-classifiers to create ensemble predictors, especially in cases where one of the two training sets, *i.e.* the 1KGP set in our study, is much bigger and has to be under sampled(33). Although partitioning of the entire 1KGP data set into smaller samples could offer a good solution(33,34) we were aware that this would pose other computational issues, as it would have resulted in millions of different predictors and would dramatically increase the running time of our method. Therefore, we decided to perform a bagging procedure instead and we created multiple balanced sets, each consisting of 200 1KGP bi-locus combinations of randomly chosen gene pairs and the 200 disease-causing combinations of DIDAv1. We first performed a preliminary analysis to assess the variance of the performance of our method relative to the number of neutral variants it is trained with, by creating different ensemble predictors trained on either 10, 100, 500, 1000, 1500, 2000 and 3500 balanced training sets. Although we did not observe significant performance differences, we finally decided to use 500 balanced sets as they showed small variance and they were the starting point of a performance plateau for bigger sets (*SI Appendix*, Table S6).

The bi-locus control combinations contained one to four variant alleles similarly to those present in DIDAv1. Bi-locus combinations with four non-wild type alleles in two different genes were not included in the 1KGP set, as this type of combinations was not yet present in DIDAv1. As variants from African populations are generally over-represented in the 1KGP set and consequently in our random subsets, we created an equal continent distribution among the individuals for each random 1KGP subset, in order to avoid bias towards variants specific to a particular population. However, preliminary analysis has shown that there is no significant difference in performance when the predictor is trained using 1KGP variant combinations only from individuals of a particular continent against DIDAv2 confirming that the method, being a qualitative machine-learning approach, is not subject to population bias (*SI Appendix*, Table S10). Finally, each random control subset contained gene pairs following a degrees of separation distribution (i.e. number of direct protein interacting connections between two genes) equal to that of DIDAv1, based on information obtained from the Human Gene Connectome tool(30) to again avoid obvious biological relatedness differences between the randomly chosen control pairs and the disease-causing gene pairs (*SI Appendix*, Fig. S5).

- **Training of the ensemble-based machine learning method**

We used the scikit-learn version 0.18.1 implementation(32) of the RF algorithm(32) as a base-classifier for each of the 500 balanced sets. Each RF predictor consisted of 100 decision trees using bootstrapping with a maximum tree depth of 10, while the number of features to consider when looking for the best split at each node was set to the square root of their total number. As most of the used features were numerical, except for the two-class Pfam categorical feature, we selected the default Gini importance option to assess feature importance.

To assess the overall performance between the independent RFs, we implemented a Leave-One-Pair-Out stratified cross-validation procedure individually for each predictor, similarly to the methodology presented in the work of Gazzo *et al.*(36). In this type of cross-validation, we iteratively removed all variant combinations of a particular gene pair and trained the RF predictor with the rest of the combinations. Then, we used the bi-locus combinations of the left-out pair to assess the performance. This procedure was repeated for all gene pairs inside each balanced set and to conclude statistics for the performance among the RF predictors, the results were averaged among all balanced sets.

- **Selection of the most relevant biological features**

The initial number of predictor features, based on the representation of each bi-locus combination, was 21 (*SI Appendix*, Table S4). To optimize the model and minimize overfitting, we used a recursive feature elimination (RFE) procedure(37) using a balanced set with median performance among all sets. In each iteration we removed the least important feature based on the Gini importance metric, and calculated statistics about the predictor performance (e.g. accuracy, sensitivity, MCC etc.).

The order of elimination of features (starting from all 21 features to 0) was: Flex4, Hydr4, Pfam2, Pfam4, Pfam1, Flex2, Pfam3, Hydr2, CADD4, Flex3, Hydr3, Hydr1, Flex1, Biol_Dist, HI_B, HI_A, CADD2, RecA, RecB, CADD1, CADD3. This means that for e.g. the turn with 2 features at the x-axis included only CADD1 and CADD3.

Based on this procedure, we observed that the use of the 10 lastly eliminated features led to the first optimal performance peak (*SI Appendix*, Fig. S3). We also observed that in general, information about the fourth allele of a bi-locus combination (gene B, allele B) was removed at the early stages of the feature elimination procedure, indicating that this allele does not contribute significantly to the classification. Although no variant features about the fourth allele remained among the 10 selected ones, we decided for interpretability reasons to include the CADD score of this allele (CADD4), being the feature of the fourth allele that was eliminated last, finalizing the number of selected features to 11 (Fig. 3C).

- **Definition of the Classification Score (CS) and Support Score (SS)**

VarCoPP implements a majority vote to predict the final class ("disease-causing" or "neutral") for a bi-locus combination ***x*****,** see Equation (1).

Equation (1):

$$maj.vote\ (\boldsymbol{x}) = \arg\max_{c=N,D} \sum_{t=1}^{T} d_{t,c}\ (\boldsymbol{x})$$

Each individual RF *t*, where $t \in \{1,\cdots,T\}$ and *T* is the total number of RFs, gives a class decision $d_{t,c}(x)$, with class *c* = {neutral (*N*), disease-causing (*D*)} for an instance *x,* see Equation (2). The majority vote function returns the class that obtained the most votes among the RFs. To make this class decision, each RF calculates a probability $p_t(x)$ for the disease-causing class and a probability 1-$p_t(x)$ for the neutral class. Based on the work of Fan & Lin(38), adjusting the probability decision thresholds can improve the performance of a multi-label classifier, a reasoning that could also be applied to our two-label classification. We identified that the best prediction performance was obtained for a probability threshold of 0.489, which has the best median True Positive / False Positive ratio (0.88/0.11) over all individual RF predictors in the ensemble (*SI Appendix*, Fig. S1). Therefore, the decision function $d_{t,c}(x)$ is defined as follows:

Equation (2):

$$d_{t,c}\ (\boldsymbol{x}) = \begin{cases} 0, & (\ c = N \text{ and } p_t(\boldsymbol{x})\ > 0.489) \text{ and } (c = D \text{ and } p_t(\boldsymbol{x})\ \leq 0.489) \\ 1, & (c = N \text{ and } p_t(\boldsymbol{x})\ \leq 0.489) \text{ and } (c = D \text{ and } p_t(\boldsymbol{x})\ > 0.489) \end{cases}$$

Each final class prediction for a bi-locus combination *x* is, therefore, supported by a classification score (CS), see Equation (3), that is defined as the median (med) of the disease-causing class probabilities over all 500 independent RFs for that bi-locus combination.

Equation (3):

$$CS\ (\boldsymbol{x}) = \underset{t=1,\ldots,T}{\text{med}} \{p_t\ (\boldsymbol{x})\}$$

The final prediction also provides a support score (SS), see Equation (4), that indicates the percentage of RFs that agree on the disease-causing label for a bi-locus combination ***x***.

Equation (4):

$$SS\ (\boldsymbol{x}) = \frac{\sum_{t=1}^{T} d_{t,D}(\boldsymbol{x})}{T} \times\ 100$$

Using these equations, we determined that bi-locus variant combinations predicted as disease-causing with VarCoPP require CS > 0.489 and SS > 50. Higher CS and SS provide a stronger indication of pathogenicity for a bi-locus combination.

- **Validation of VarCoPP using independent disease-causing and neutral data**

As a validation set, we collected 23 new disease-causing bi-locus combinations derived from independent scientific papers published after the release of DIDAv1 (Fig. 3E and *SI*

Appendix, Table S7 and Dataset S1). We also collected different sets of random 100, 1000 and 10000 neutral bi-locus combinations from the 1KGP that were unused during training to assess the amount of FPs that we can obtain (Fig. 3D and *SI Appendix*, Datasets S2, S3 and S4), without considering the population of individuals and degrees of separation between genes in a bi-locus combination. However, the genes and variants had similar properties to those used during the training of VarCoPP. The gene pairs that were involved in all these bi-locus combinations had not been exploited during the training phase of the predictor. To assess the predictive ability of VarCoPP on these new data, we created three Precision-Recall (PR) curves using the validation set with the 23 disease-causing combinations and either the 100, 1000 or 10000 random neutral test bi-locus test combinations (*SI Appendix*, Fig. S2). The variant filtering procedure, as well as the order of variants and genes inside each combination, was defined in the same way as for the training sets that were used in this work.

- **Gene panel analysis for assessment of False Positives**

We first assessed the performance of VarCoPP on random lists of genes on multiple 1KGP individuals. We created 100 iterations of random panels consisting of 10, 30, 100 and 300 genes and tested each gene panel on 100 random 1KGP individuals. To create the gene panel specific for Bardet-Biedl syndrome (BBS), we used the publicly available 21-gene list obtained from the Genome Diagnostics Nijmegen laboratory (http://www.genomediagnosticsnijmegen.nl/). We obtained the autism/intellectual disability (ID) gene panels from the SFARI Gene database (https://gene.sfari.org/) and more specifically the gene panel of scoring category 1 (high confidence, 24 genes), category 2 (strong candidate, 56 genes) and category 3 (suggestive evidence, 158 genes). The variant filtering procedure on these panels was similar with the one performed during the creation of the predictor.

- **Understanding the contribution of each feature to the predictions using *treeinterpreter***

To gain insight in how each feature contributes to the classification of a bi-locus combination as either disease-causing or neutral, we exploited the *treeinterpreter* Python package (https://github.com/andosa/treeinterpreter, Ando Saabas). This library explores, for each bi-locus combination, information on the order and the way features are used during a decision path from the root of each decision tree to the final leaf. This helps assess how much each feature contributes or "votes" for either the disease-causing or neutral class. The contribution or "vote" for a particular class is represented with a contribution value that is calculated from *treeinterpreter*. A positive feature contribution value means that this feature votes for the disease-causing class, while a negative feature contribution value means that the feature contributes to the neutral class. This also indicates that in the end, the sum of the decision values of all features for a particular bi-locus combination delivers the final classification; positive sum for the disease-causing class and negative sum for the neutral class. As we obtain a contribution value for each decision tree, in order to obtain the feature contribution values for each bi-locus combination inside an individual RF, the contribution values were averaged among all trees of that RF, providing us with a vector called the Decision Profile (DP), which contained average decision values for each feature. Feature contribution values should not be confused with the actual feature values of a bi-locus combination; the former values are derived using *treeinterpreter*, whereas the latter are the actual values of the features used for classification. Thus, if a bi-locus combination is represented with a feature vector x = (x1,x2,...,x10), then, using the *treeinterpreter* library, we get a DP vector y=(y1,y2,...,y10) for each individual RF, where y_i is the contribution value of the feature *i*.

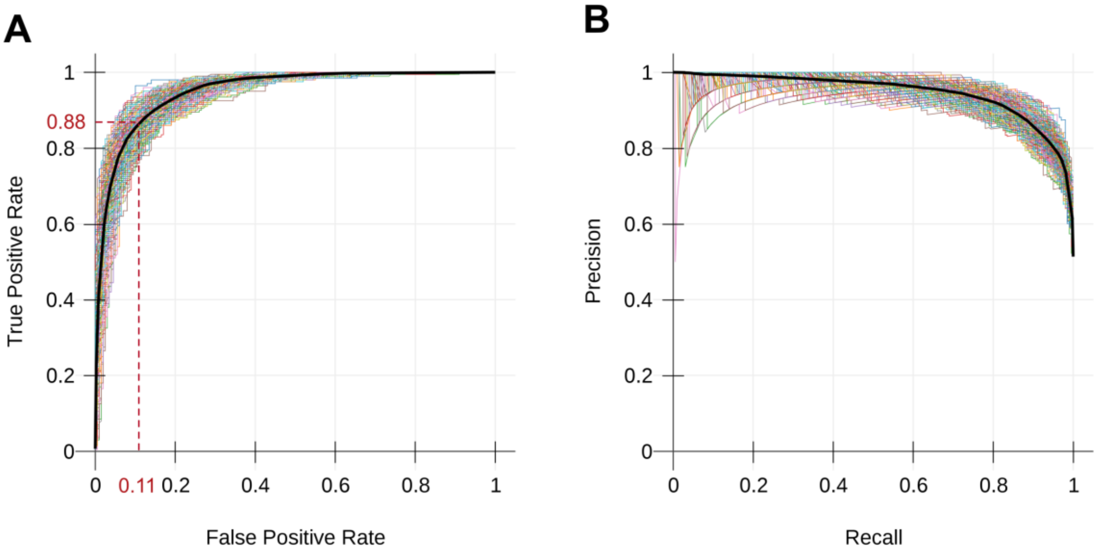

Fig. S1. Performance of VarCoPP during cross-validation. The different colored lines represent the curve from each individual predictor of the ensemble, while the black curve is the average curve among all predictors. (A) Receive Operator Characteristic (ROC) curve. The median optimal True Positive (TP) and False Positive (FP) ratio is 0.88 / 0.11 and is associated with a median probability threshold of 0.489 that differentiates disease-causing from neutral combinations. (B) Precision - Recall (PR) curve.

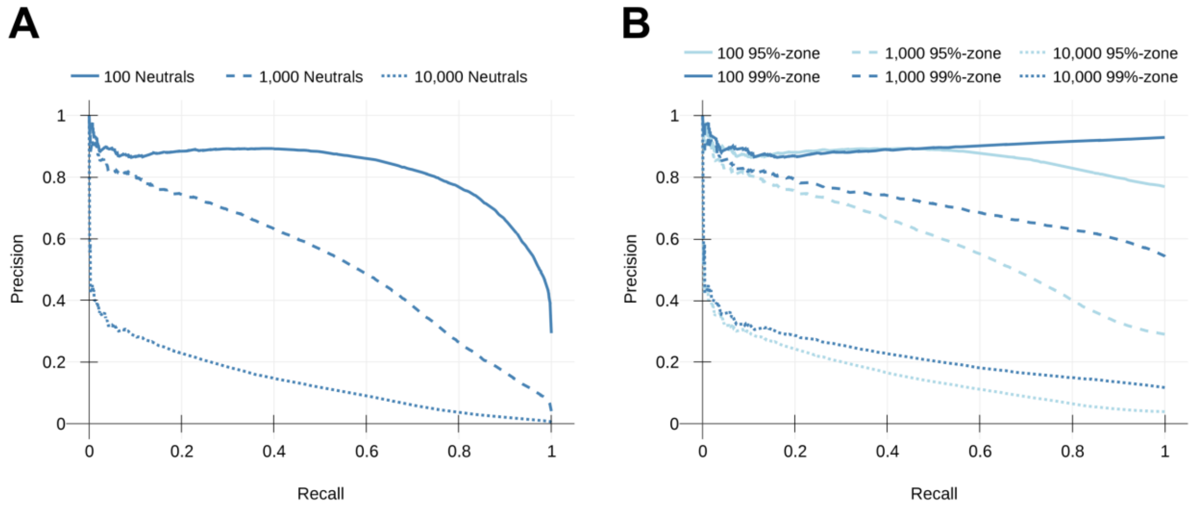

Fig. S2. Precision - Recall curve (PR) curves when VarCoPP is tested on the neutral test set and validation set bi-locus combinations. An interpretation of the PR curves is available in *SI Appendix*, Text S4. (A) PR using the instances of the validation set together with either the 100 (straight line), 1000 (dashed line) or 10000 test neutral combinations (dotted line). This PR curve contains the average performance among all independent 500 predictors of VarCoPP. (B) PR curve using the validation set instances together with those of either the 100 (straight line), 1000 (dashed line) or 10000 test neutral sets (dark blue), which fall into the 95%-zone (light blue) or 99%-zone (dark blue).

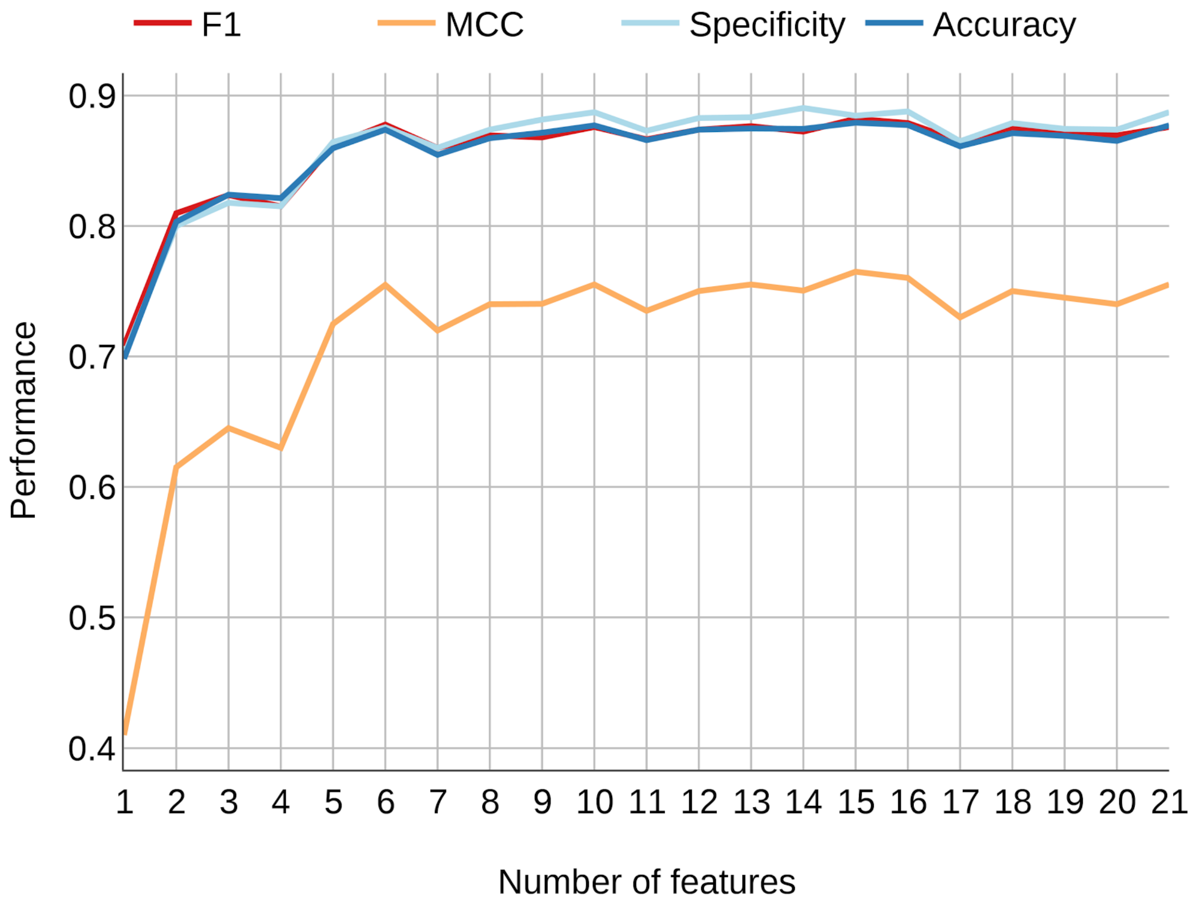

Fig. S3. Predictor performance during the recursive feature elimination procedure of a balanced training set with median performance among the balanced sets, after a cross validation procedure. The x-axis represents the **number** of features included for prediction during each elimination procedure, while the y axis represents the average performance during the cross-validation for that number of features. The order of elimination of features (starting from all 21 features to 0) was: Flex4, Hydr4, Pfam2, Pfam4, Pfam1, Flex2, Pfam3, Hydr2, CADD4, Flex3, Hydr3, Hydr1, Flex1, Biol_Dist, HI_B, HI_A, CADD2, RecA, RecB, CADD1, CADD3. This means that for e.g. the turn with 2 features at the x-axis included only CADD1 and CADD3. Performance was measured in terms of accuracy, specificity, F1 score (a balanced measure between precision and recall) and Matthews Correlation Coefficient (a balanced measure of the quality of a two-class classification, resembling a correlation coefficient between the observed and predicted binary classifications). It is shown that a set of 10 features inferred the first best peak among all performance measures (with slightly better scores than that of 6 features). For interpretation reasons, we also included CADD4, which was not part of the 10 first selected features, finalizing the number of features to 11.

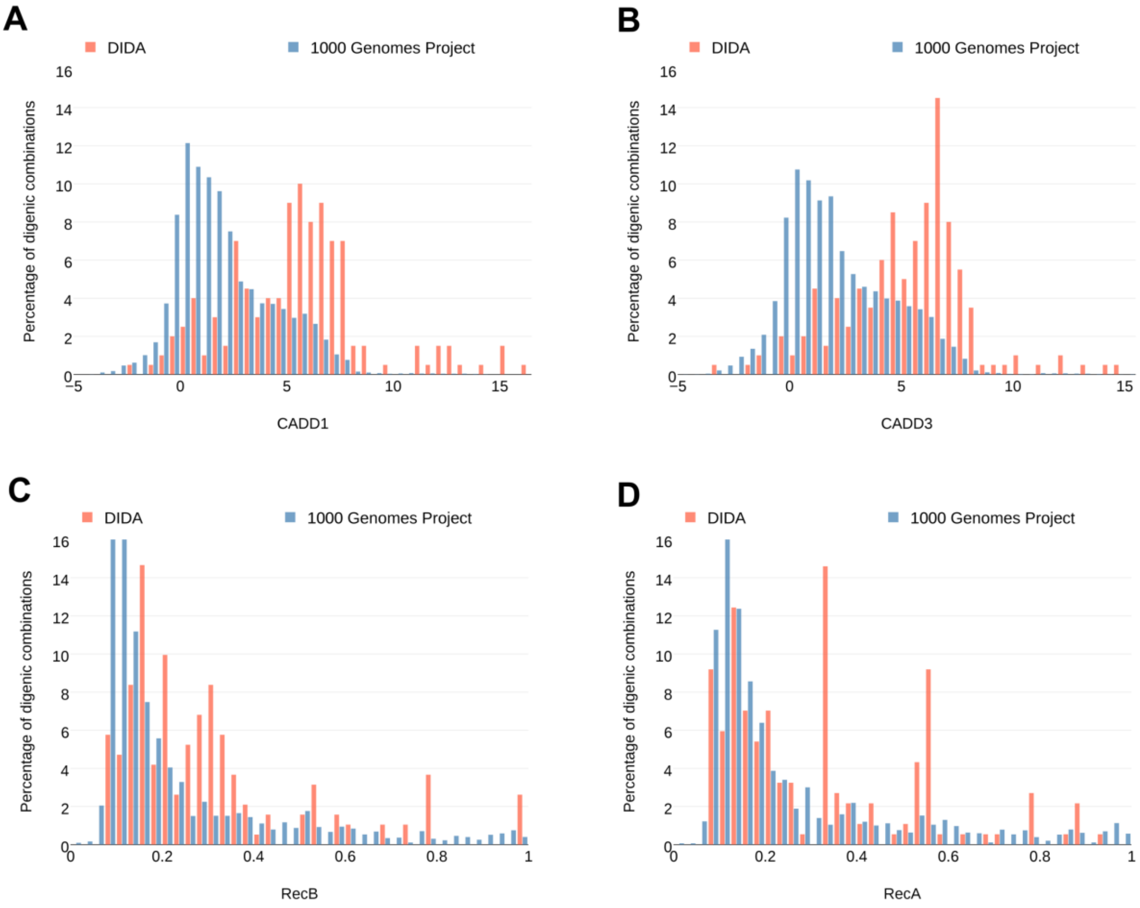

Fig. S4. Distribution of feature values for (A) CADD1, (B) CADD3, (C) RecB and (D) RecA between DIDAv1 and the 1000 Genomes Project training sets.

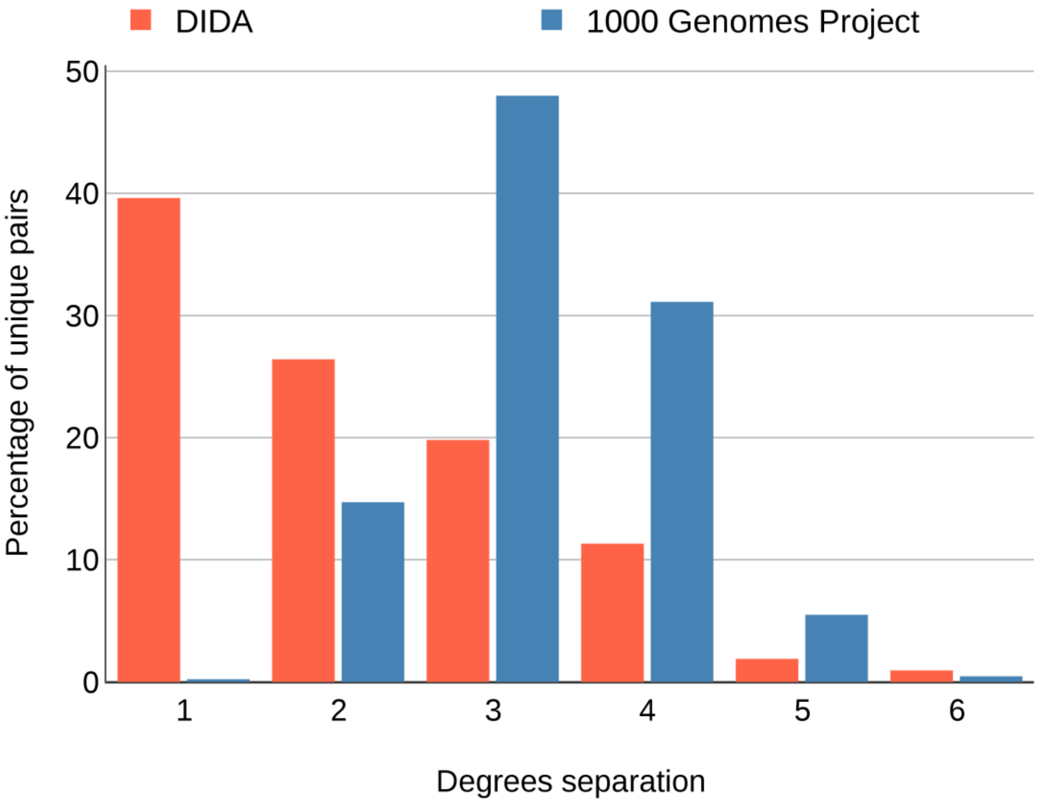

Fig. S5. Histogram of the degrees separation for the gene pairs present in the subset of the 1KGP (in blue) used in our analysis and DIDA (in red). Degrees separation is a metric that depicts how many proteins intervene in the pathway between a pair of genes. Value of 1 indicates that the proteins of two genes in a pair are directly interacting.

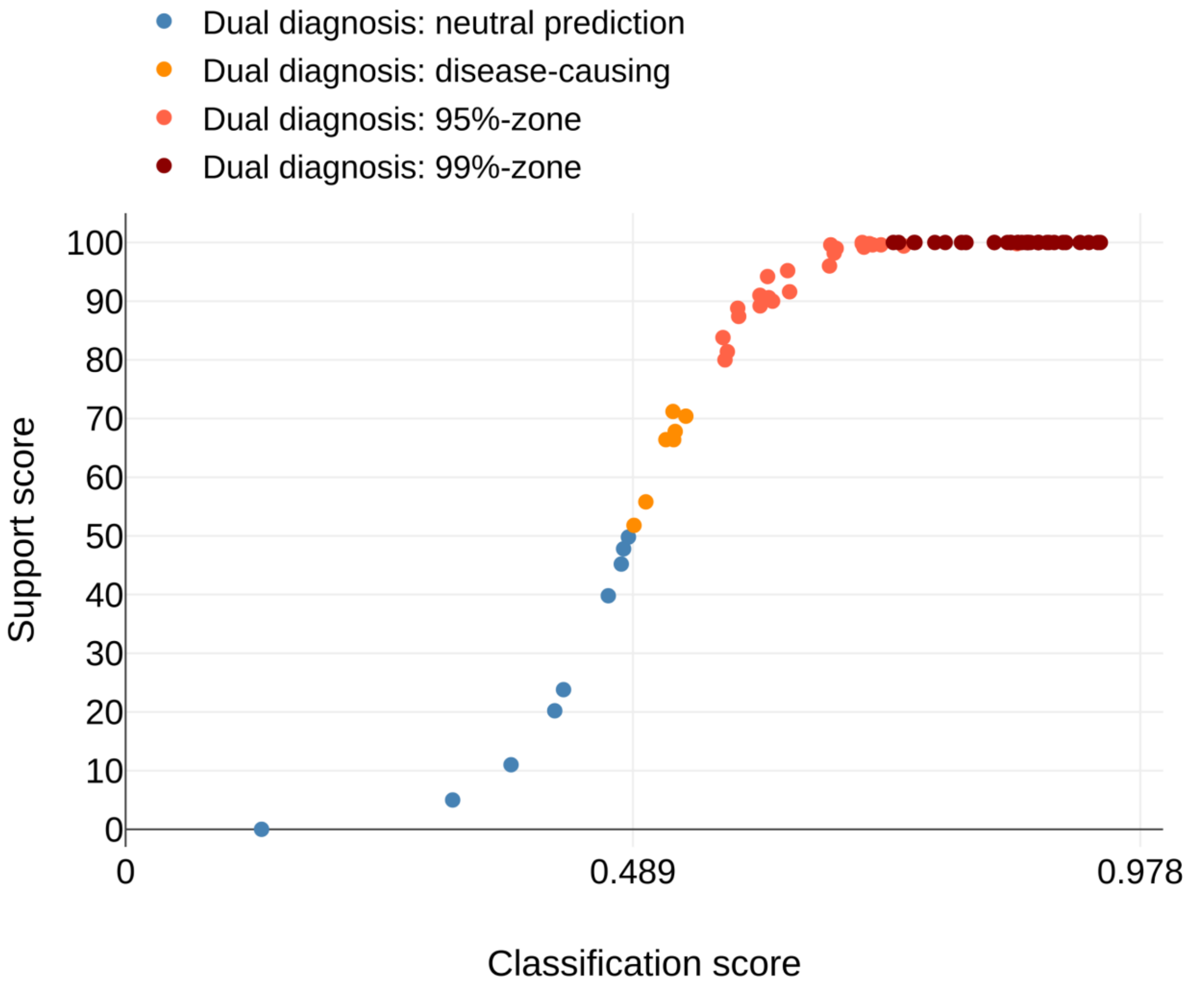

Fig. S6. Performance of VarCoPP on 76 Dual Molecular Diagnosis cases, extracted from Posey *et al.*(39)

Table S1. Information on the ancestry of the 1KGP individuals carrying 4 or more DIDA bi-locus combinations (including those that carry a DIDA combination).

Number of DIDA variants per 1KGP individual	**# 1KGP individuals**	**Ancestry**
4 DIDA variants	22	Europe: 3 Africa: 19
5 DIDA variants	1	South Asia
6 DIDA variants	1	Europe

Table S2. Information on the frequency of the DIDA variants found in the 1KGP individuals of Table S1, as well as the effect of the DIDA combinations they are involved in. Associated overlapping 1KGP-DIDA bi-locus variant combinations that were removed from the creation of VarCoPP (see Table S3) are depicted with *. Bi-locus combinations that were removed from the creation of VarCoPP during the variant filtering process, are depicted with ** (TD = True Digenic, MD = Monogenic + Modifier, N/A: Unknown).

Variant dbSNP ID	Gene	gnomAD MAF	1KGP MAF	Corresponding DIDA combination(s)	Associated disease name in DIDA	Oligogenic effect(s) of DIDA combinations
rs186471205	MYH7B	0.01062	0.004193	dd171	Left ventricular non-compaction	N/A
rs72546668	CAV3	0.002667	0.001997	dd070	Familial Long QT syndrome	N/A
rs61747728	NPHS2	0.03025	0.014577	dd131	Familial idiopathic steroid-resistant nephrotic syndrome	N/A
rs151257815	STXBP2	0.01171	0.004	dd193*	Familial hemophagocytic lymphohistiocytosis	TD
rs41281314	CDH23	0.01378	0.049321	dd009	Usher syndrome	MD
rs28464386	STXBP2	0.00763	0.027	dd195, dd196*	Familial hemophagocytic lymphohistiocytosis	TD
rs73507527	KISS1R	0.006764	0.019968	dd140, dd202*	Normosmic congenital hypogonadotropic hypogonadism, Kallman syndrome	N/A
rs78861628	PROKR2	0.004944	0.01258	dd202*	Kallman syndrome	N/A
rs75366116	UNC13D	0.005203	0.019169	dd184	Familial hemophagocytic lymphohistiocytosis	TD
rs6499838	BBS2	0.01198	0.038	dd105	Bardet-Biedl syndrome	N/A
rs34982899	NPHS1	0.01519	0.010783	dd115	Familial idiopathic steroid-resistant nephrotic syndrome	TD
rs78028658	UNC13D	0.0005123	9.98E-04	dd176	Familial hemophagocytic lymphohistiocytosis	TD

rs74315416	PROKR2	0.002196	9.98E-04	dd012, dd015, dd018, dd136	Kallman syndrome	N/A (dd012, dd015), MD (dd018, d136)
rs150880478	TTC8	0.005035	0.007188	dd112	Bardet-Biedl syndrome	N/A
rs35947132	PRF1	0.02916	0.013179	dd174, dd176, dd177, dd178, dd180*, dd182, dd187, dd188*, dd191, dd193*	Familial hemophagocytic lymphohistiocytosis	TD
rs118049905	UNC13D	0.00001335	0.001797	dd182, dd185, dd194, dd196*, dd197	Familial hemophagocytic lymphohistiocytosis	TD
rs138382758	MFN2	0.002177	0.001997	dd152	Charcot-Marie-Tooth disease	MD
rs117761837	STXBP2	0.01057	0.004593	dd188*, dd190	Familial hemophagocytic lymphohistiocytosis	TD
rs532361142	BBS2	0.00003893	0.00003893	dd083	Bardet-Biedl syndrome	TD
rs145125791	NPHS1	0.005974	0.003994	dd132	Familial idiopathic steroid-resistant nephrotic syndrome	N/A
rs201763096	NEXN	0.004068	0.003195	dd217	Familial isolated hypertrophic cardiomyopathy	N/A
rs540150447	STX11	0.0007816	0.003594	dd198	Familial hemophagocytic lymphohistiocytosis	TD
rs117106081	PROKR2	0.005727	0.012181	dd203*	Kallman syndrome	N/A
rs386598428	EDNRB	0.01011	0.005	dd022**	Hirschsprung disease	TD
rs41263993	CCDC28B	0.01112	0.006	dd084, dd085, dd204	Bardet-Biedl syndrome	MD (dd084, dd085), TD

Table S3. Information on individuals of the 1KGP carrying DIDA bi-locus combinations.

1KGP individual	Population (Continent)	DIDA combination ID	Associated disease	PMID
HG02375	CDX (East Asia)	dd203	Kallmann syndrome	20022991
HG04042	STU (South Asia)	dd220	MODY	25041077
HG00123	GBR (Europe)	dd193	Familial hemophagocytic lymphohistiocytosis	24916509
NA12842	CEU (Europe)	dd180	Familial hemophagocytic lymphohistiocytosis	24916509
NA12812	CEU (Europe)	dd196	Familial hemophagocytic lymphohistiocytosis	24916509
NA12342	CEU (Europe)	dd193	Familial hemophagocytic lymphohistiocytosis	24916509
		dd188	Familial hemophagocytic lymphohistiocytosis	24916509
HG03388	MSL (Africa)	dd202	Kallmann syndrome	20022991
NA19225	YRI (Africa)	dd202	Kallmann syndrome	20022991
NA20807	TSI (Europe)	dd188	Familial hemophagocytic lymphohistiocytosis	24916509
HG03410	MSL (Africa)	dd202	Kallmann syndrome	20022991
HG01670	IBS (Europe)	dd193	Familial hemophagocytic lymphohistiocytosis	24916509
HG03713	ITU (South Asia)	dd220	MODY	25041077
NA19031	LWK (Africa)	dd202	Kallmann syndrome	20022991
HG00112	GBR (Europe)	dd188	Familial hemophagocytic lymphohistiocytosis	24916509

Table S4. Features used to annotate the 1KGP and DIDA data sets. Those that remained after the feature selection procedure are marked with (*).

Feature	Feature Abbreviations	Description	PMID
CADD raw score	CADD1* CADD2* CADD3* CADD4*	First variant allele of gene A Second variant allele of gene A First variant allele of gene B Second variant allele of gene B	24487276
Pfam protein domain	Pfam1 Pfam2 Pfam3 Pfam4	First variant allele of gene A Second variant allele of gene A First variant allele of gene B Second variant allele of gene B	26673716
Amino acid hydrophobicity difference	Hydr1* Hydr2 Hydr3 Hydr4	First variant allele of gene A Second variant allele of gene A First variant allele of gene B Second variant allele of gene B	8836100
Amino acid flexibility difference	Flex1* Flex2 Flex3 Flex4	First variant allele of gene A Second variant allele of gene A First variant allele of gene B Second variant allele of gene B	- Bhaskaran & Ponnuswamy1988
Haploinsufficiency probability	HI_A* HI_B*	Haploinsufficiency probability for gene A Haploinsufficiency probability for gene B	20976243
Recessiveness probability	RecA* RecB*	Recessiveness probability for gene A Recessiveness probability for gene B	22344438
Biological distance	Biol_Dist*	Biological relatedness between gene A and gene B	24694260

Table S5. Computer readable representation of values for the features presented in Table S4.

Variant features	Type	Values	Representation in the vector	Explanation
CADD raw score (alleles 1 – 4)	Numerical	Raw score NaN Wild-type	[raw score] [CADD median]* [-3]	* Median value of the CADD score of the same type of variants for either DIDA or 1KGP.
Amino acid flexibility difference (alleles 1 – 4)	Numerical	Protein variant Intronic/splicing variant NaN (frameshift/loss of stop mutation) Wild-type	[difference value] [0.0] [median value]* [0.0]	* Median value of either DIDA (0.012) or 1KGP (0.0).
Amino acid hydrophobicity difference (alleles 1 – 4)	Numerical	Protein variant Intronic/splicing variant NaN (frameshift/loss of stop mutation) Wild-type	[difference value] [0.0] [median value]* [0.0]	*Median value of either DIDA (-0.1) or 1KGP (-0.01).
Pfam region (alleles 1 – 4)	Categorical	Yes No NaN Wild-type	[1] [0] [0] [same value as the first allele of the variant]	
Haploinsufficiency probability (geneA/B)	Numerical	Known probability NaN	[value] [median = 0.19898]	
Recessiveness probability (geneA/B)	Numerical	Known probability NaN	[value] [median = 0.12788]	
Biological distance	Numerical	Known distance NaN	[distance] [median]*	*Median for gene pairs with “NaN” pathway for DIDA (4.72) and 1KGP (18.16).

Table S6. Mean performance and standard deviation (sd) statistics among all 500 RFs of the ensemble predictor during cross-validation and using the final selection of 11 features.

Number of balanced sets	**Accuracy**		**Precision**		**Sensitivity**		**MCC**	
	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
10	0.884	0.014	0.892	0.017	0.876	0.014	0.77	0.022
50	0.874	0.013	0.884	0.014	0.863	0.015	0.750	0.024
100	0.875	0.016	0.883	0.019	0.865	0.016	0.751	0.031
500	0.878	0.014	0.886	0.016	0.868	0.015	0.757	0.026
500*	0.877	0.014	0.880	0.016	0.874	0.015	0.756	0.027
1000	0.878	0.014	0.886	0.016	0.867	0.015	0.756	0.026
1500	0.876	0.014	0.884	0.016	0.867	0.015	0.754	0.026
2000	0.877	0.014	0.885	0.016	0.867	0.015	0.755	0.026
3500	0.877	0.014	0.885	0.016	0.867	0.015	0.755	0.027

***** adjusted probability threshold of 0.489

Table S7. Validation set of 23 new disease-causing bi-locus combinations. For details about the variants and the feature annotation of these combinations, see *SI Appendix*, Dataset S1.

Combination ID	Gene pair*	Type	Associated disease	PMID
testpos_1	*TRIM54 - TRIM63*	triallelic	Protein aggregate myopathy	25801283
testpos_2	*PSMA3 - PSMB8*	diallelic	CANDLE syndrome	26524591
testpos_3	*PSMA3 - PSMB8*	diallelic	CANDLE syndrome	26524591
testpos_4	*PSMB4 - PSMB9*	diallelic	CANDLE syndrome	26524591
testpos_5	*PSMB4 - PSMB8*	diallelic	CANDLE syndrome	26524591
testpos_6	*CDK5RAP2 - CEP152*	diallelic	Seckel syndrome	26436113
testpos_7	*COL4A4 - COL4A3*	diallelic	Alport sydrome	25575550
testpos_8	*COL4A4 - COL4A3*	diallelic	Alport sydrome	25575550
testpos_9	*COL4A4 - COL4A3*	diallelic	Alport sydrome	25575550
testpos_10	*COL4A4 - COL4A3*	diallelic	Alport sydrome	25575550
testpos_11	*COL4A4 - COL4A3*	diallelic	Alport sydrome	25575550
testpos_12	*COL4A5 - COL4A4*	diallelic	Alport sydrome	25575550
testpos_13	*COL4A5 - COL4A4*	diallelic	Alport sydrome	25575550
testpos_14	*COL4A5 - COL4A4*	diallelic	Alport sydrome	25575550
testpos_15	*COL4A5 - COL4A4*	triallelic	Alport sydrome	25575550
testpos_16	*SEC23A - MAN1B1*	tetrallelic	Overgrowth syndrome	27148587
testpos_17	*RP1L1 - C2orf71*	triallelic	Syndromic retinitis pigmentosa	29295593
testpos_18	*SHH - DISP1*	diallelic	Holoprosencephaly	26748417
testpos_19	*MITF - GJB2*	diallelic	Deafness	27057829
testpos_20	*AHI1 - CEP290*	triallelic	Leber congenital amaurosis	20683928
testpos_21	*RPE65 - CEP290*	triallelic	Leber congenital amaurosis	20683928
testpos_22	*CRB1 - CEP290*	triallelic	Leber congenital amaurosis	20683928
testpos_23	*AHI1 - CEP290*	triallelic	Joubert syndrome	20683928

*Gene order for each combination is based on the Gene Damage Index (GDI).

Table S8. Statistics on the performance of VarCoPP using random test sets of 100, 1000 and 10000 neutral bi-locus combinations derived from the 1KGP (TNs=True Negatives, FPs=False Positives, SS = Support Score, CS = Classification Score).

	number of 1KGP test combinations		
	100	**1000**	**10000**
definitive TNs (SS=0)	67.0%	72.1%	72%
FPs (SS>50)	7.0%	7.7%	7.2%
min. 95%-zone SS	80.6	74.8	74.8
min. 95%-zone CS	0.57	0.55	0.55
min. 99%-zone SS	100.0	100.0	100.0
min. 99%-zone CS	0.74	0.749	0.72

Table S9. Statistics on recall on the training data and on the validation set of 23 independent disease-causing bi-locus variant combinations, when applying the 95% and 99% confidence zones.

	Overall recall	95% confidence zone recall	99% confidence zone recall
Training set (cross-validation)	0.87	0.84	0.60
Validation set	0.87	0.87	0.60

Table S10. Average performance and standard deviation (sd) statistics among all 500 RFs of VarCoPP during cross-validation, using each time 1KGP individuals from one particular continent as the control training set against DIDA.

1KGP continent	**Accuracy**		**Precision**		**Sensitivity**		**MCC**	
	Average	Sd	Average	Sd	Average	Sd	Average	Sd
Africa	0.89	0.01	0.89	0.01	0.89	0.01	0.78	0.02
America	0.87	0.01	0.84	0.02	0.91	0.01	0.74	0.04
East Asia	0.86	0.01	0.83	0.02	0.91	0.01	0.73	0.04
Europe	0.86	0.01	0.83	0.02	0.91	0.01	0.73	0.03
South Asia	0.87	0.01	0.84	0.02	0.90	0.01	0.74	0.04

Additional Dataset S1 (separate Excel file)
Variant information, annotated features and VarCoPP prediction scores for the 23 bi-locus combinations of the disease-causing validation set.

Additional Dataset S2 (separate Excel file)
Variant information, annotated features and VarCoPP prediction scores for the 100 random bi-locus combinations extracted from 1KGP.

Additional Dataset S3 (separate Excel file)
Variant information, annotated features and VarCoPP prediction scores for the 1000 random bi-locus combinations extracted from 1KGP.

Additional Dataset S4 (separate Excel file)
Variant information, annotated features and VarCoPP prediction scores for the 10000 random bi-locus combinations extracted from 1KGP.

Additional Dataset S5 (separate Excel file)
Variant information, annotated features and VarCoPP prediction scores for the 76 Dual Molecular Diagnosis bi-locus combinations extracted from the paper of Posey *et al.*(39)

Supplementary Information References

1. Schäffer AA (2013) Digenic inheritance in medical genetics. *J Med Genet* 50(10):641-652

2. Sarfati J, et al. (2010) A comparative phenotypic study of kallmann syndrome patients carrying monoallelic and biallelic mutations in the prokineticin 2 or prokineticin receptor 2 genes. J Clin Endocrinol Metab 95(2):659–669.

3. Chan YM, et al. (2009) GNRH1 mutations in patients with idiopathic hypogonadotropic hypogonadism. PNAS 106(28):11703–11708.

4. Pitteloud N, et al. (2007) Digenic mutations account for variable phenotypes in idiopathic hypogonadotropic hypogonadism. J Clin Invest 117(2):457-463.

5. Shaw ND, et al. (2011) Expanding the phenotype and genotype of female GnRH deficiency. J Clin Endocrinol Metab 96(3):E566-576.

6. Mendez JP, et al. (2015) Triallelic digenic mutation in the prokineticin 2 and GNRH receptor genes in two brothers with normosmic congenital hypogonadotropic hypogonadism. Endocr Res 40(3):166-171.

7. de Roux et al. (1997) A Family with Hypogonadotropic Hypogonadism and Mutations in the Gonadotropin-Releasing Hormone Receptor. N Engl J Med 337:1597-1603.

8. Monnier C, et al. (2009) PROKR2 missense mutations associated with Kallmann syndrome impair receptor signalling activity. Hum Molec Genet 18:75-81.

9. Chapla A, et al. (2015) Maturity onset diabetes of the young in India - a distinctive mutation pattern identified through targeted next-generation sequencing. Clin Endocrinol 82(4):533–542.

10. Babu DA, Chakrabarti SK, Garmey JC, Mirmira RG (2008) Pdx1 and BETA2/NeuroD1 Participate in a Transcriptional Complex That Mediates Short-range DNA Looping at the Insulin Gene. J Biol Chem 283(13):8164-8172.

11. Gonsorcíková L, et al. (2008) Autosomal inheritance of diabetes in two families characterized by obesity and a novel H241Q mutation in NEUROD1. Pedriatr Diabetes 9(4 pt2):367-372

12. Cockburn, et al. (2004) Insulin promoter factor-1 mutations and diabetes in Trinidad: identification of a novel diabetes-associated mutation (E224K) in an Indo-Trinidadian family. J Clin Endocrinol Metab 89(2): 971–978.

13. Zhu Y, et al. (2013) MicroRNA-24/MODY Gene Regulatory Pathway Mediates Pancreatic β-Cell Dysfunction. Diabetes 62(9):3194-3206.

14. Zhang K, et al. (2014) Synergistic defects of different molecules in the cytotoxic pathway lead to clinical familial hemophagocytic lymphohistiocytosis. Blood 124(8):1331–1334.

15. De Saint BG, Menasche G, Fischer A. (2010) Molecular mechanisms of biogenesis and exocytosis of cytotoxic granules. Nat Rev Immunol 10(8):568-579.

16. Voskoboinik I, Smyth MJ, Trapani JA (2006). Perforin-mediated target-cell death and immune homeostasis. Nat Rev Immunol. 6(12):940-952.

17. Sepulveda FE, Garrigue A, Maschalidi S et al. (2016) Polygenic mutations in the cytotoxicity pathway increase susceptibility to develop HLH immunopathology in mice. Blood 127: 2113-2121.

18. Chen X, et al. (2018) Genetic variant spectrum in 265 Chinese patients with hemophagocytic lymphohistiocytosis: molecular analyses of PRF, UNC13D, STX11, STXBP2, SH2D1A, and XIAP. Clin Genet doi: 10.1111/cge.13261 (Epub ahead of print).

19. UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43(Database issue):D204–12.
20. Kinsella RJ, et al. (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database 2011:bar030.
21. Pruitt KD, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19(7):1316–1323.
22. Kircher M, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315.
23. Bhaskaran R, Ponnuswamy PK (2009) Dynamics of amino acid residues in globular proteins. Int J Pept Protein Res 24(2):180–191.
24. Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3(10):842–848.
25. Finn RD, et al. (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):D279–85.
26. Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 6(10):e1001154.
27. MacArthur DG, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823–828.
28. Liu X, Jian X, Boerwinkle E (2011) dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32(8):894–899.
29. Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Hum Mutat 34(9):E2393–E2402.
30. Itan Y, et al. (2014) HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. BMC Genomics 15:256.
31. Itan Y, et al. (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci U S A 112(44):13615–13620.
32. Breiman L (2001) Random Forests. J Mach Learn Res 45(1):5–32.
33. Sun Z, et al. (2015) A novel ensemble method for classifying imbalanced data. Pattern Recognit 48(5):1623–1637.
34. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Trans Syst Man Cybern C Appl Rev 42(4):463–484.
35. Pedregosa F, et al. (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12(October):2825−2830.
36. Gazzo A, et al. (2017) Understanding mutational effects in digenic diseases. Nucleic Acids Res 45(15):e140.
37. Guyon I, Weston J, Barnhill S (2002) Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn 46:389-422.
38. Fan RE, Lin CJ (2007) A Study on Threshold Selection for Multi-label Classification (National Taiwan University) Available at: https://www.csie.ntu.edu.tw/~cjlin/papers/threshold.pdf.
39. Posey JE, et al. (2017) Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N Engl J Med 376(1):21–31.