Predicting disease-causing variant combinations

Sofia Papadimitriou[a,b,c], Andrea Gazzo[a,b,d], Nassim Versbraegen[a,b], Charlotte Nachtegael[a,b], Jan Aerts[e,f], Yves Moreau[e,g], Sonia Van Dooren[a,d,h], Ann Nowé[a,c], Guillaume Smits[a,i,j,1], and Tom Lenaerts[a,b,c,1]

[a]Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles-Vrije Universiteit Brussel, 1050 Brussels, Belgium; [b]Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium; [c]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, 1050 Brussels, Belgium; [d]Center for Medical Genetics, Reproduction and Genetics, Reproduction Genetics and Regenerative Medicine, Vrije Universiteit Brussel, UZ Brussel, 1090 Brussels, Belgium; [e]Center for Statistics, Universiteit Hasselt, 3590 Diepenbeek, Belgium; [f]Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Katholieke Universiteit Leuven, 3001 Leuven, Belgium; [g]Interuniversitair Micro-Electronica Centrum (IMEC), 3001 Leuven, Belgium; [h]Brussels Interuniversity Genomics High-Throughput Core, Université Libre de Bruxelles-Vrije Universiteit Brussel, 1090 Brussels, Belgium; [i]Hôpital Universitaire des Enfants Reine Fabiola, Université Libre de Bruxelles, 1020 Brussels, Belgium; and [j]Center of Human Genetics, Hôpital Erasme, Université Libre de Bruxelles, 1070 Brussels, Belgium

Edited by Aravinda Chakravarti, New York University School of Medicine, New York, NY, and approved April 26, 2019 (received for review September 20, 2018)

Notwithstanding important advances in the context of single-variant pathogenicity identification, novel breakthroughs in discerning the origins of many rare diseases require methods able to identify more complex genetic models. We present here the Variant Combinations Pathogenicity Predictor (VarCoPP), a machine-learning approach that identifies pathogenic variant combinations in gene pairs (called digenic or bilocus variant combinations). We show that the results produced by this method are highly accurate and precise, an efficacy that is endorsed when validating the method on recently published independent disease-causing data. Confidence labels of 95% and 99% are identified, representing the probability of a bilocus combination being a true pathogenic result, providing geneticists with rational markers to evaluate the most relevant pathogenic combinations and limit the search space and time. Finally, the VarCoPP has been designed to act as an interpretable method that can provide explanations on why a bilocus combination is predicted as pathogenic and which biological information is important for that prediction. This work provides an important step toward the genetic understanding of rare diseases, paving the way to clinical knowledge and improved patient care.

pathogenicity | bilocus combination | variants | prediction | oligogenic

Advances in high-throughput sequencing technologies and the application of massive parallel sequencing have revolutionized the field of human genetics, providing a huge amount of information on human genetic variation (1–5). Interpreting this variation has provided important insights into the genetic architecture of many rare diseases, notably those inherited in a Mendelian pattern (6–8), and has opened the path to promising preventive, diagnostic, and therapeutic strategies (9). The amount of genetic data available has also allowed for the development of successful predictive tools that integrate genetic, molecular, evolutionary, and/or structural information (10–13). Such tools are routinely applied in clinics to identify pathogenic variants potentially associated with a specific disease phenotype. Notwithstanding these advancements, the analysis of a growing number of rare human disorders has highlighted the difficulties in establishing a genotype–phenotype relationship due to non-Mendelian patterns of inheritance, incomplete penetrance, phenotypic variability, or locus heterogeneity (14–18). The classic concept of one gene leading to a particular phenotype appears to be an oversimplification, since to better explain the situation of an affected individual, one often needs to consider more complex genetic models where mutations in multiple genes cause or modulate the development of one or several simultaneous disease phenotypes (15, 19–21).

Oligogenic or multilocus genetic patterns have already been discovered for diseases initially considered to be monogenic, for instance, phenylketonuria (22) or hereditary nonsyndromic deafness (23). These types of diseases may have a central primary causative gene and a network of modifier genes, as in Hirschsprung disease (24) and cystic fibrosis (25), or they may present a spectrum of genetic models from monogenic to polygenic, as in the case of neurodevelopmental disorders (26, 27). Gene-disease network analysis studies further support the notion that a disease phenotype is hardly the result of a mutation in one gene alone, showing that the vast majority of Mendelian diseases may actually be modulated by multiple genes that are usually involved in similar pathways or cellular and biological processes (28, 29). Along with the cases where a phenotype or syndromic phenotypes can be modulated by several genes, a multilocus genetic pattern can also be observed in an affected individual where disease-causing monogenic mutations in different genes segregate independently, leading to multiple independent molecular clinical diagnoses (21, 30–33). Some cases of multiple diagnoses can affect different tissues (distinct), but others can share phenotypes (overlapping), indicating a possible relationship between the involved multilocus variations at the protein or cellular level. It is evident that in order for the clinical predictive tools to remain valuable for diagnostic purposes, they need an update toward these more elaborate biological and inheritance scenarios. For instance, such tools will need to consider that the nature or frequency of variants observed in oligogenic diseases

Significance

Directly assessing the pathogenicity of variant combinations in multiple genes was until now difficult. Nonetheless, this type of assessment can provide important benefits in identifying the genetic causes of rare diseases. The work presented in this paper aims to resolve this problem by presenting a machine-learning method able to predict the pathogenicity of variant combinations in gene pairs, based on pathogenic data. We demonstrate the high accuracy of this method and its effective capacity to identify novel instances. The method's decision-making process is also made explicit, a contribution that is useful for clinical interpretation. This pioneering work will lead to toolboxes for geneticists and clinicians that can aid them in counselling their patients more effectively.

Author contributions: S.P., A.G., J.A., Y.M., S.V.D., A.N., G.S., and T.L. designed research; S.P., A.G., N.V., C.N., G.S., and T.L. performed research; S.P., A.G., N.V., C.N., G.S., and T.L. analyzed data; and S.P., A.G., N.V., C.N., J.A., Y.M., S.V.D., A.N., G.S., and T.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The source files and code for VarCoPP, as well as the training data that were used, have been deposited at Github, and are available at https://github.com/sofiapapad90/VarCoPP/.

[1]To whom correspondence may be addressed. Email: guillaume.smits@erasme.ulb.ac.be or tlenaert@ulb.ac.be.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1815601116/-/DCSupplemental.

Published online May 24, 2019.

Downloaded at KU Leuven University Library on December 13, 2019

will be different from those observed in monogenic ones (20, 34). The current work makes this leap, introducing and validating a computational approach that predicts the pathogenicity of variant combinations as opposed to single variants, and this within the context of gene pairs.

This leap is made possible by the steady increase in literature reports on disease-causing variant combinations in gene pairs (bilocus variant combinations) in the last decades, which have been grouped and made publicly available via an online resource, the Digenic Diseases Database (DIDA) (35). This novel resource collects, organizes, and annotates cases where a bilocus genetic model helped to explain a patient's phenotypic variability and reduced penetrance, including, for example, the well-known cases of Bardet–Biedl syndrome (BBS) (36, 37) and retinitis pigmentosa (38). The first version of the database (which will be referred to henceforth as the DIDAv1) contained 213 manually curated bilocus variant combinations obtained from independent scientific papers involving 136 different genes and leading to 44 diseases (a detailed explanation of the curation process in the DIDA is provided in *SI Appendix*, Text S1). These variant combinations are divided into three classes based on their effect (Fig. 1). The first class, referred to as the "true digenic class," requires the presence of variants in two independent genes to trigger the disease, with carriers of the variants found in one gene being unaffected. The second class covers Mendelizing variants with modifiers, which is referred to as the "composite class." In this scenario, the individual carrying the Mendelizing variant can present symptoms of the disease, with the extra variant at the second gene modifying the severity of the symptoms or the age of onset. The DIDAv1 also contained a few cases of a third class, which is referred to as the "dual molecular diagnoses" class. This class consists of those cases wherein two disease-causing Mendelizing variants in different genes lead to two independent clinical diagnoses. Given their limited number in the DIDAv1, they were added to the composite class. An initial study on the DIDAv1 revealed that biological features defined at the variant, gene, and combination levels are sufficient to differentiate composite from true bilocus variant combinations, providing novel insights into the properties of disease-causing bilocus variant combinations (39). Although the terms bilocus and digenic are both used for pairs of variant combinations, we will use here the term bilocus so as to avoid confusion with those combinations referred to as "true digenic" (as discussed above).

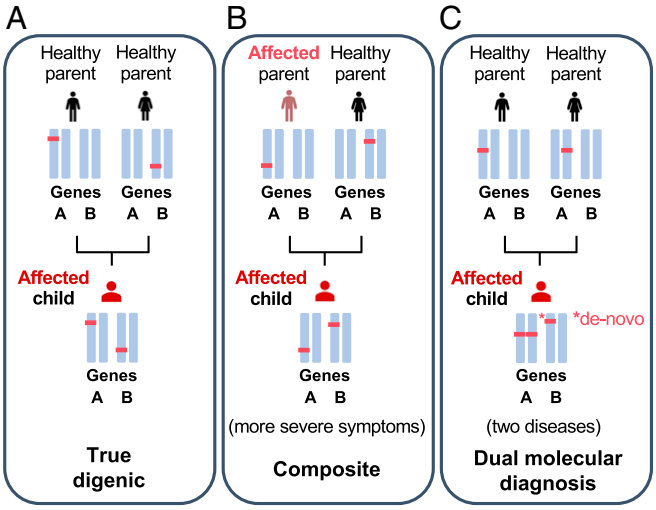


Fig. 1. Examples of different cases of disease-causing bilocus variant combinations present in an individual, and which can be detected by the VarCoPP. (*A*) "True digenic" case, where mutations on both genes should be present to trigger any symptoms of the disease. Individuals with the mutation in either one of the two genes remain unaffected. (*B*) One example of a "composite" case, where one mutation at the most deleterious gene can be sufficient to show disease symptoms (affected parent), but the second mutation affects the severity of symptoms or the age of onset. (*C*) One example of a dual molecular diagnosis case, which concerns the simultaneous aggregation of variants that cause two independent Mendelian diseases, with or without overlapping phenotypes. It should be noted that dual molecular diagnosis cases can include different inheritance models (e.g., segregation of two recessive diseases).

Based on the presence of these fully annotated bilocus disease data in the DIDAv1 and the variety of cases they cover, one can hypothesize that the transition from single to variant combination pathogenicity predictors is now possible, starting from variant combinations within gene pairs. Such a predictor should exclude the nonrelevant variant combinations [true negative (TN)], which will be abundantly present in a patient's exome, and accurately identify the scarce disease-causing ones [true positive (TP)]. To meet this challenge, we developed the Variant Combination Pathogenicity Predictor (VarCoPP), a pathogenicity predictor for combinations of variants in gene pairs, which is able to accurately identify disease-causing variant combinations using variant, gene, and gene pair information. The accuracy and sensitivity of the predictor are also validated on an independent dataset consisting of new bilocus disease data from novel publications that appeared after the construction of the DIDAv1. Moreover, by visualizing how each feature guides the pathogenicity prediction, the VarCoPP provides an explanation as to why a given bilocus variant combination is classified as disease-causing or not. To further support clinical geneticists in their analysis, statistical scores for each prediction, as well as 95% and 99% confidence labels for each evaluated combination, are provided. These labels capture the most relevant variant combinations that should be further analyzed clinically and provide potential for patient counseling. VarCoPP is available online at http://varcopp.ibsquare.be/ (40).

Results

Curation of the 1000 Genomes Project Data Reveals the Presence of Known Disease-Causing Bilocus Variant Combinations. In total, 46% of the individuals in the 1000 Genomes Project (1KGP) carry at least one variant found in the DIDAv1 (Fig. 2). Twenty-four individuals with a diverse, mostly African, ancestry carried four or more DIDA variants, with the vast majority of them being rare [minor allele frequency (MAF) < 0.01]. Information on these individuals and the corresponding variants can be found in *SI Appendix*, Tables S1 and S2.

In general, the majority of all overlapping variants (86%) are involved in disease-causing variant combinations belonging to the true digenic class, thus explaining their monogenic presence in a control population. Nevertheless, more than 10% of overlapping variants are involved in bilocus combinations with a monogenic + modifier effect. Most of the variants found in the 1KGP (69%) are located in the secondary (modifier) gene of the pair, possibly explaining why the control individuals carrying them could be asymptomatic. However, the rest of the overlapping variants are probably located in the primary (Mendelizing) gene, and some of them have been shown to cause disease symptoms in individuals in a dominant monogenic fashion, like the variants c.511C > T and c.637G > A in the WNT10A gene, which are involved in tooth agenesis (41); the variant c.670G > A in the PDX1 gene, which is involved in the development of maturity-onset diabetes of the young 4 (MODY 4) (42, 43); and the c.313G > A variant in the SLC7A9 gene, which is involved in nontype I cystinuria (44, 45). It should be noted that MODY could be overlooked, as the c.670G > A variant can present incomplete penetrance (42), as also suggested by its frequency in the Exome Aggregation Consortium database (0.002113), while incomplete penetrance is also well known for nontype I cystinuria. Tooth agenesis could also be easily clinically overlooked.

Downloaded at KU Leuven University Library on December 13, 2019

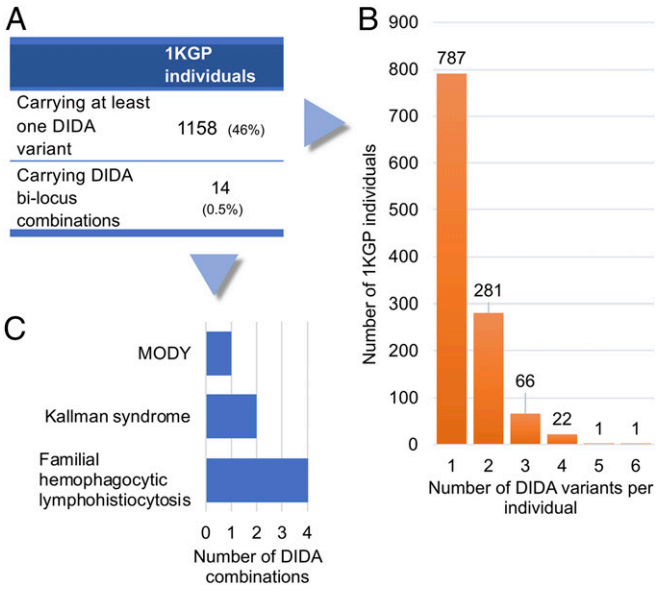


Fig. 2. Overlapping variants and bilocus combinations between the DIDA and 1KGP. (*A*) Statistics on 1KGP individuals carrying at least one DIDA independent variant or a disease-causing bilocus combination. (*B*) Histogram of 1KGP individuals carrying one or more DIDA variants (including those that carry DIDA combinations). (*C*) Histogram of the DIDA bilocus combinations found in the 1KGP and the diseases they are leading to.

Intriguingly, we discovered seven disease-causing bilocus combinations present in the DIDAv1, leading to MODY (43), Kallman syndrome (46), or familial hemophagocytic lymphohistiocytosis (47) in 14 individuals of the 1KGP (Fig. 2 and *SI Appendix*, Text S2 and Table S3). These combinations were not supported by functional evidence in the original studies and had not been compared with a large control cohort to further statistically ensure their relevance. However, some of the involved pairs were supported by familial evidence in their original papers (detailed information is provided in *SI Appendix*, Text S2). From a clinical point of view, the individuals could also be undiagnosed: A mild Kallman syndrome could be easily clinically overlooked, and, as stated beforehand, the c.670G > A variant for MODY can present incomplete penetrance. Furthermore, the bilocus combinations could be incompletely penetrant.

To ensure that the data used for the construction of the VarCoPP does not contain contradicting instances, we removed from our analysis these 14 individuals from the 1KGP neutral set, as well as the seven incriminated bilocus combinations from the DIDAv1 as a precaution.

The VarCoPP Identifies Accurately Pathogenic Variant Combinations. Using bilocus variant combinations randomly selected from individuals of the 1KGP (5) as the neutral set and the bilocus variant combinations from the DIDAv1 (35) as the disease-causing set, we successfully trained the VarCoPP (a summary of the procedure is provided in Fig. 3). We limited the search space to 1KGP variants with up to 3% MAF to match the frequency range observed in the DIDAv1, located in or close to exons (Fig. 3*A*). Each gene and variant inside a bilocus combination were ordered in the same way for both datasets, a process necessary for reliability (Fig. 3*B* and *Materials and Methods*). We then annotated our data with information at the variant, gene, and gene-pair level, leading to 21 characteristics (computationally called "features") in total per bilocus combination. This set was reduced to 11 after a feature selection procedure (Fig. 3*C* and *SI Appendix*, Tables S4 and S5). This annotated information for each bilocus combination of the DIDA and 1KGP was then used as training input for the machine learning method, allowing it to learn how to differentiate between pathogenic and neutral bilocus combinations.

Without going into the technical details (*Materials and Methods*), it is important to mention that the VarCoPP is an ensemble predictor (48), meaning that it is composed of a large number (500) of individual predictors that each try to solve the same task. The individual decisions of the predictors are combined via a majority vote to define the final class: If 50% or more of the predictors agree that a bilocus combination is disease-causing, then the "pathogenic" class label will be assigned to that combination (Fig. 3*D*). Our results show that the VarCoPP performs very well, achieving a TP rate of 0.88 and a false positive (FP) rate of 0.11 (*SI Appendix*, Fig. S1), meaning that 88% of the disease-causing combinations of the DIDAv1 are correctly identified, with 11% wrongful assignments of the disease-causing label in nonrelevant combinations. The Matthews correlation coefficient, a more robust measure for the predictive quality of binary classifications that takes into account the correlation between observed and predicted results, achieves a score of 0.74, confirming that the method is highly accurate (*SI Appendix*, Table S6). It is important to also note that these results were obtained using a stratified form of cross-validation on the training data (*Materials and Methods*), meaning that considerable efforts were made to avoid bias and overfitting in the construction and evaluation of the predictor.

For each variant combination given as input, the VarCoPP generates a final majority class label ("pathogenic" or "neutral") and two prediction scores: (*i*) a classification score (CS) (i.e., the median probability that the variant combination is pathogenic) calculated over all pathogenic probabilities provided by each individual predictor of the ensemble, and (*ii*) a support score (SS) (i.e., the percentage of individual predictors in the ensemble agreeing on the pathogenic label) (a detailed explanation of these scores is provided in Fig. 3*D* and *Materials and Methods*). The higher the CS and SS, the more confident the predictor is about the classification of a bilocus combination as pathogenic. To better split the neutral and disease-causing combinations, the CS threshold for pathogenic combinations was optimized to 0.489 (*Materials and Methods*). Consequently, as the predictor is based on a majority vote, a bilocus variant combination is predicted to be pathogenic when it has SS > 50 and CS > 0.489 (Fig. 4*A*). If we plot the predictions of the bilocus combinations of the DIDAv1 during cross-validation based on these two evaluation scores (CS on the *x* axis and SS on the *y* axis), we see that they are distributed in an S-shaped curve (Fig. 4*B*). The vast majority of the DIDAv1 data (88%) cluster with high confidence in the right part of the S-shaped curve.

Validation on Independent Disease-Causing Data Confirms the VarCoPP's Predictive Success. As the evaluation on independent data provides the best insight into the quality of a predictive method, we validated the VarCoPP on a set of 23 new bilocus disease-causing variant combinations, which were gathered from research articles published after the creation of the DIDAv1 (Fig. 3*E*, *SI Appendix*, Table S7, and Dataset S1). These independent bilocus variant combinations contained unexplored gene pairs associated with 10 diseases not previously reported in the DIDA. This independent set includes diseases such as Alport syndrome (49) [Online Mendelian Inheritance in Man (OMIM): catalog nos. 301050, 203780, and 104200], holoprosencephaly (50) (OMIM catalog no. 236100), and Leber congenital amaurosis (51) (OMIM catalog no. 204000).

The VarCoPP remains very successful when used with these new data (Fig. 4*C*): The vast majority of the new bilocus combinations (20 of 23) are correctly labeled as pathogenic, with a high confidence (SS > 80). Three bilocus combinations, one leading to chronic atypical neutrophilic dermatosis with lipodystrophy and elevated temperature syndrome (52) and two leading to Alport

Downloaded at KU Leuven University Library on December 13, 2019

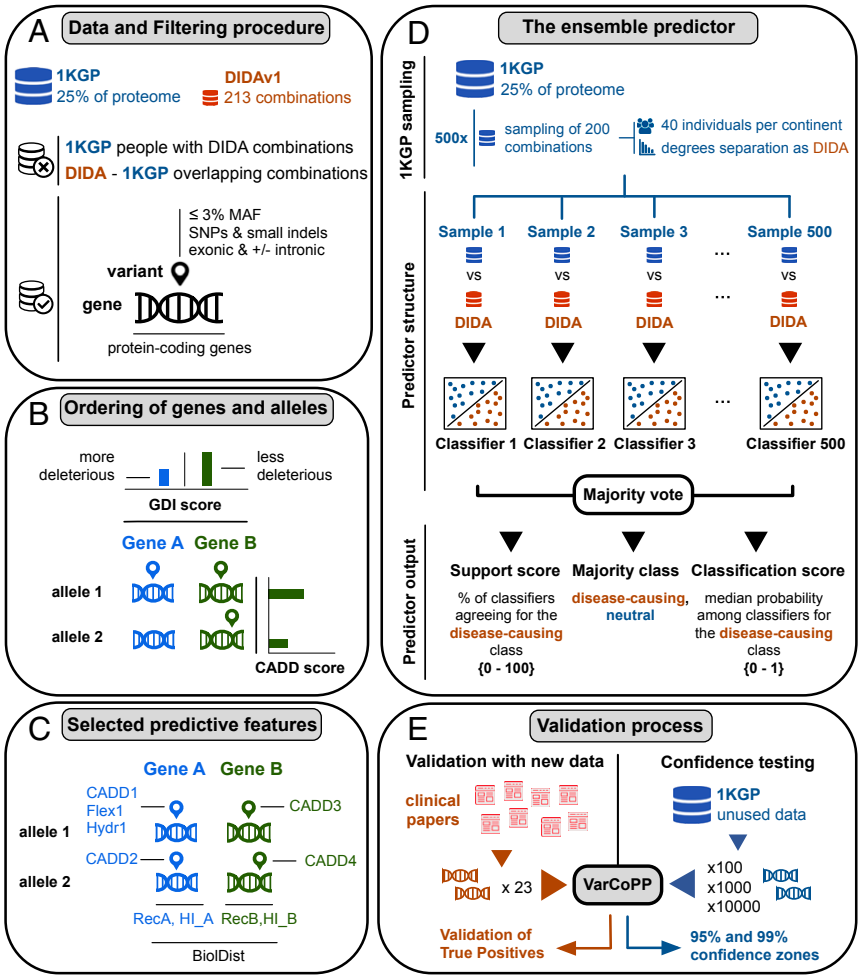


Fig. 3. Summary of the methodology procedure for the construction of the VarCoPP and the validation process. (*A*) Genes and variants were filtered in the same way for both the 1KGP and DIDAv1. Individuals of the 1KGP carrying DIDAv1 combinations, as well as the overlapping combinations, were filtered out. Exonic variants [single-nucleotide polymorphism (SNPs) and indels] were used with a MAF frequency of ≤3%, including intronic and synonymous variants close to the exon edges (±13 nucleotides). The genes involved in the procedure were only confirmed protein-coding genes, following the gene types present in the DIDAv1. (*B*) Bilocus variant combination is represented always using four alleles (two alleles for gene A and two alleles for gene B), including wild-type alleles. This was done in accordance with the information present in the DIDA, where each bilocus combination contained, at maximum, two mutated alleles inside each gene. With this representation, the variant zygosity is also being considered (e.g., for a homozygous variant, both available alleles of the gene contain the same variant information). In this specific panel, we show a bilocus combination with a heterozygous variant in gene A (the second allele is wild-type) and two different heterozygous variants in gene B. Gene A is always the gene with the lowest Gene Damage Index (GDI) score, thus with the higher probability of being a deleterious gene. Different variant alleles inside the same gene were ordered based on their CADD pathogenicity score, with the variant present in the first allele of that gene always having the highest CADD score. (*C*) Initial number of biological features used for classification was 21, but the final selected and more relevant features were filtered to 11. These included information at the variant level [Flex1 and Hydr1 (i.e., flexibility and hydrophobicity amino acid differences of the first variant allele of gene A), as well as CADD1, CADD2, CADD3, and CADD4, (i.e., the CADD scores of the four different alleles of a bilocus combination)], gene level [RecA, RecB, HI_A, HI_B (i.e., recessiveness and haploinsufficiency probabilities for gene A and gene B)], and gene-pair level [BiolDist (i.e., biological distance, a metric of biological relatedness between two genes of a pair based on protein–protein interaction information)]. A more detailed explanation of the features is provided in *SI Appendix*, Table S4. (*D*) After the filtering process, the 1KGP dataset contained billions of bilocus combinations compared with the DIDAv1 set, which contained 200 bilocus combinations. To solve this class imbalance problem, 500 random 1KGP samples, each containing 200 bilocus combinations, were extracted using two types of stratification: Each sample contained an equal amount (41) of bilocus combinations from individuals of each continent as well as an equal distribution of degrees of separation (i.e., a metric of protein–protein interaction distance) between the genes of each pair, following the degrees of separation distribution of the DIDAv1. Each 1KGP sample was used against the complete DIDAv1 set to train an individual classifier that gives a class probability for each bilocus combination. Based on a majority vote among the individual classifiers, the output of the VarCoPP for each tested bilocus combination is the final class ("neutral" or "disease-causing"), the SS (i.e., the percentage of the classifiers agreeing about the pathogenic class), and the CS (i.e., the median probability among the individual predictors that the bilocus combination is pathogenic). (*E*) To validate the VarCoPP on new disease-causing data, we collected 23 bilocus combinations from independent scientific papers, which included gene pairs not used during the training phase. To perform confidence testing, we extracted three different random sets of 100, 1,000 and 10,000 bilocus combinations from the 1KGP set, which included gene pairs not used during the training phase of the VarCoPP. By exploring the number of FPs predicted with these neutral sets, we defined 95% and 99% confidence zones that provide the minimum SS and CS boundaries above, of which a bilocus combination has a 5% or 1% probability, respectively, of being a FP.

syndrome (49), were wrongfully predicted as neutral, with support of SS = 46.2, SS = 24, and SS = 1.6, respectively. The gene pairs involved seem to be relevant for the studied disease, and the genes of the pairs were closely biologically related, indicating that their protein products are most likely directly interacting. However, low Combined Annotation Dependent Depletion (CADD) variant scores, a single-variant pathogenicity metric (12), and some missing gene recessiveness and haploinsufficiency values are most likely the reasons why those combinations were misclassified (*SI Appendix*, Text S3). When these missing data become

Downloaded at KU Leuven University Library on December 13, 2019

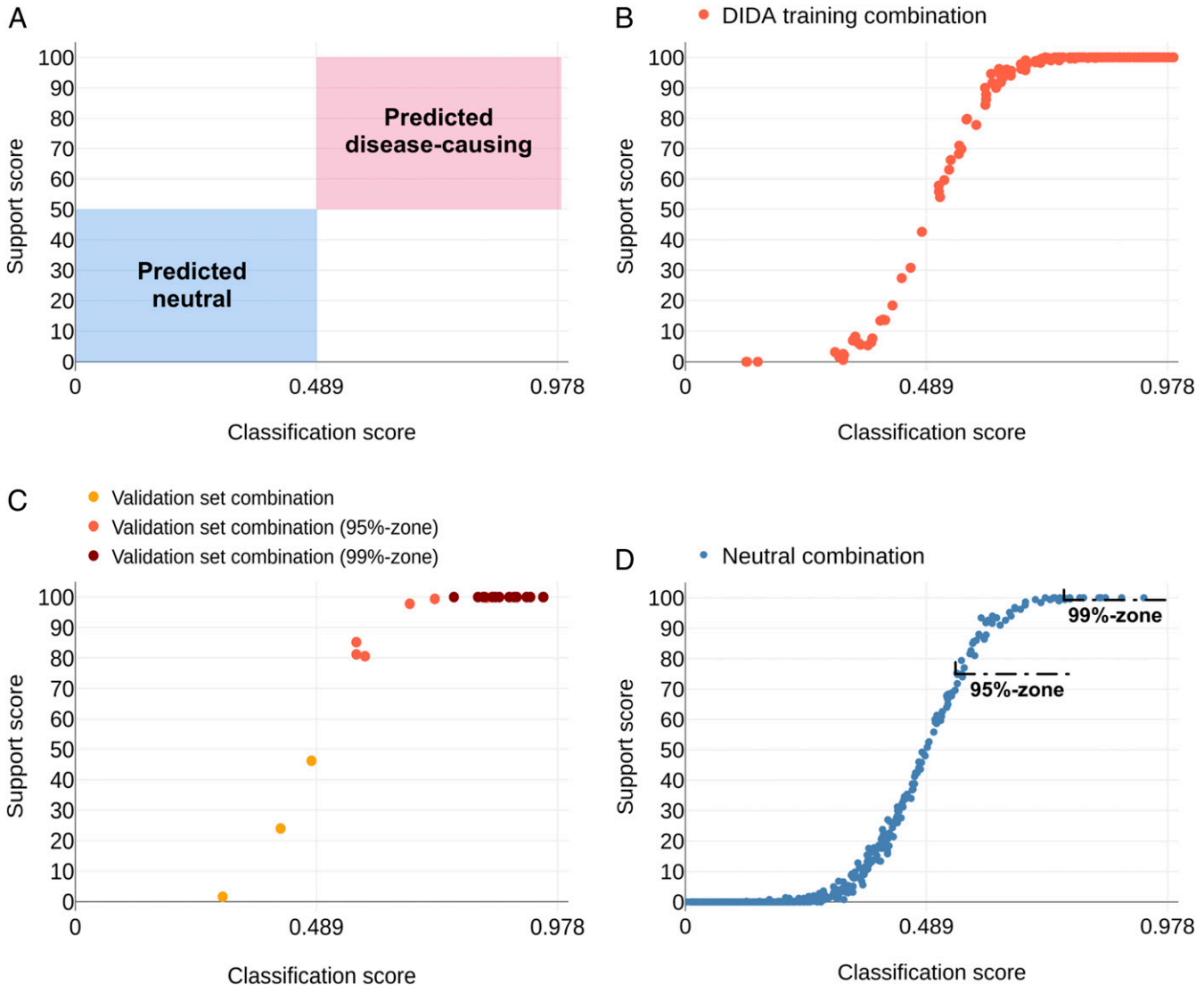


Fig. 4. Distribution of the predictions of the DIDAv1 and of the independent test bilocus combinations, based on the CS on the *x* axis and the SS on the *y* axis. (*A*) SS > 50 and CS > 0.489 were required to label a bilocus combination as disease-causing. The red box represents the area where a bilocus combination is predicted as disease-causing, while the blue box represents the area where a bilocus combination is predicted as neutral. (*B*) Distribution of disease-causing bilocus combinations of the DIDAv1 during a cross-validation procedure. (*C*) Distribution of the 23 disease-causing bilocus combinations of the validation set. (*D*) Distribution of the 1,000 neutral test set combinations. The 95% confidence zone has a minimal boundary of CS = 0.55 and SS = 75, and contains combinations with a 5% probability of being FPs, while the 99% confidence zone has a minimal boundary of CS = 0.74 and SS = 100, and contains combinations with a 1% probability of being FPs.

available or annotations are improved, the VarCoPP might also classify these three cases correctly.

Statistical Confidence Zones Make It Easy to Detect the Most Relevant Combinations. It can be expected that even after a standard variant filtering procedure, the number of neutral variant combinations (i.e., TNs) in an individual's exome will vastly outnumber the number of the real disease-causing ones (i.e., TPs). It is therefore highly relevant to estimate how likely it is that a variant combination predicted as pathogenic by the VarCoPP is actually a FP.

To examine this FP probability, we randomly collected neutral variant combinations from 1KGP individuals, consisting exclusively of gene pairs unknown to the VarCoPP, and calculated their prediction scores (i.e., their CS and SS) (Fig. 3*E*). We analyzed three different sets of such random combinations [sets of 100, 1,000 (Fig. 4*D*), and 10,000 combinations] to also examine whether the percentage of FPs changes relative to the sample size (Datasets S2, S3, and S4, respectively).

We observed that, on average, 93% of the combinations are correctly identified as neutral, of which 72% have a confirmative SS equal to zero, meaning that no predictor in the ensemble classified them as disease-causing (*SI Appendix*, Table S8). The overall fraction of FP combinations predicted as disease-causing fluctuates at around 7–8%. This percentage remains stable even if the sample size changes. Therefore, in general, there is only a 7% chance that a bilocus variant combination is wrongfully predicted to be disease-causing.

Using this insight, it is possible to define stringent confidence zones for the predictions, delimited by specific CS and SS scores, which denote the probability that a bilocus variant combination is a TP. We define in this manner a 95% confidence zone containing all predicted variant combinations that have at least CS ≥ 0.55 and SS ≥ 75. Combinations belonging to this zone have at least 95% probability to be TP disease-causing variant combinations. Similarly, we define a 99% confidence zone, which requires at least CS ≥ 0.74 and SS = 100, containing all predicted

Downloaded at KU Leuven University Library on December 13, 2019

combinations that have a 99% or higher probability of being a TP (*SI Appendix*, Table S8). These confidence zones are useful as the focus can fall directly on the bilocus variant combinations belonging to one of these two zones, and therefore have higher confidence of being relevant. Underlining again the quality of the VarCoPP, one can observe that all 20 correctly classified elements in the independent validation set discussed in the previous section belong to at least the 95% confidence zone, with 15 of those even present in the 99% confidence zone (Fig. 4*C*).

Although these confidence zones provide a guarantee on the probability of a variant combination being a TP, the absolute number of combinations falling in those zones increases with the number of variant combinations to be tested. This is also the case when testing single variants with monogenic pathogenicity predictors. A consequence of this observation is that the precision [i.e., the fraction of real disease-causing combinations (TPs) detected among those that were predicted to be disease-causing (TPs and FPs)] and recall (i.e., the fraction of the real disease-causing combinations predicted correctly as pathogenic over all real disease-causing combinations present in the dataset) will be affected: The smaller the fraction of real disease-causing combinations among all tested combinations, the smaller is the precision and the larger is the difficulty to recall them all (*SI Appendix*, Text S4, Fig. S2, and Table S9). As a consequence, it is best to filter down the number of variants and genes as much as possible before testing them for pathogenicity with the VarCoPP. Another possibility would be to apply post-VarCoPP FP reducing strategies, such as, for example, using trio data, to avoid considering further irrelevant combinations already present in an unaffected parent.

Confidence Zones Are Relevant for the Clinical Analysis of Disease-Specific Gene Panels. With the previously defined 95% and 99% confidence zones and additional filtering steps, we can restrict our analysis to the most relevant pathogenic bilocus variant combinations within full exomes. However, as a large absolute number of combinations to consider may still exist, one can further reduce the number of combinations by zooming in on those combinations that occur in a subset of genes related to the disease of interest (i.e., to restrict the analysis to well-defined gene panels). However, even by shifting to a gene panel, the current predictive quality of the VarCoPP might be altered due to the specific properties of the genes included in that panel.

First, we assessed the expected absolute number of FP combinations in the 95% and 99% confidence zones for different sizes of randomly generated gene panels (ranging from 10 to 300 genes). This analysis provides insight into the number of FP combinations present in each confidence zone that we can expect for a random gene panel of a given size, which consists only of neutral variants. This insight is essential as geneticists do not want to be confronted with a large amount of FPs in these zones, given the time and costs associated with analyzing and/or testing them. On the other hand, knowing how many FPs to expect in the confidence zones relative to the size of the gene panel provides a baseline that could be used to quantify differences between healthy patients and those having a specific disease phenotype: If the number of predicted variant combinations present in the confidence zones for a gene panel of a particular size exceeds significantly what is expected for random neutral combinations, then there may be important genetic information in the predicted results that merits future exploration.

The results for random gene panels of different sizes (10, 30, 100, and 300 genes) that contain neutral variant combinations from 1KGP individuals (details are provided in *Materials and Methods*) are shown in Table 1. One can first observe that the percentage of FPs does not fluctuate significantly among the random gene panels, similar to the random neutral validation data results described before (*SI Appendix*, Table S8). There is only a slight increase in the percentage of FPs in the 95% confidence zone for the 100 and 300 random gene panels. The strict 99% confidence zone appears to be more consistent, as for all random gene panels, it contains, on average, less than 1% of neutral bilocus combinations per individual. The absolute number of FP combinations increases, as expected, with the size of the gene panel; of the 1,312 variant combinations per individual that are generated, on average, for a panel of 300 genes, ∼12 (0.9%) may end up in the 99% confidence zone. Additional evaluations of those cases using knowledge about the disease phenotype or molecular functionalities will most likely further reduce these numbers to acceptable sets of combinations to evaluate clinically or test experimentally.

Second, as known disease gene panels can have more detrimental properties than randomly selected ones, given that they are known to be associated with a disease, it is important to see how these statistics change in such circumstances. We decided here to evaluate, on one hand, a gene panel for a disease known to be caused by bilocus variants (i.e., BBS) and, on the other hand, gene panels for a mono-to-polygenic disease [i.e., autism/intellectual disability (ID)], using SFARI Gene top categories, applied again on neutral combinations of 1KGP individuals (*Materials and Methods*). Whereas the first BBS set is expected to generate higher percentages of FPs as most of the genes are present in the DIDAv1, we expected to see a reduction in FPs in the latter panels.

As can be observed in Table 1, the VarCoPP appears to predict more FPs for the BBS gene panel compared with a gene panel of random genes with similar size. The BBS panel contains highly recessive genes with low haploinsufficiency probabilities (0.19 on average) and whose neutral 1KGP variants have relatively higher CADD scores compared with random genes. However, the defined confidence zones are still clinically relevant as the VarCoPP guarantees that, on average, less than one variant combination will be predicted as pathogenic and will be present in the strict 99% confidence zone. As a consequence, almost any bilocus combination present in the 99% confidence zone should be clinically relevant. Such an assertion could be tested in the future on new cohorts of BBS-demonstrated bilocus patients.

The gene panels of autism/ID, although larger in size, reveal lower FP fractions, as expected. This result is most likely due to the observation that genes in those panels have high haploinsufficiency probabilities (0.40–0.49 on average among the different panels), while the 1KGP variants present in those genes generally have lower CADD scores than average. Hence, the 95% and 99% confidence zones stay quite devoid of false predictions. Together, these results show that the VarCoPP can be very precise, making it a relevant tool for discovery and diagnosis.

The Synergy of Different Biological Features Determines the Pathogenicity. The VarCoPP combines a number of molecular features at the variant, gene, and gene-pair level to identify which variant combinations are potentially disease-causing. By analyzing how each feature influences the predictions independently, we can gain an idea about their relative importance for the full predictor. Through a feature selection procedure, we determined that a subset of 11 biological features of the original 21 (Fig. 3*C* and *SI Appendix*, Text S5, Fig. S3, and Table S4) is sufficient for making high-quality predictions while, at the same time, reducing the chance of overfitting.

For each of these 11 features, we calculated a Gini importance score (53), which quantifies the importance of a feature proportionally to the number of samples it can successfully differentiate. Fig. 5 shows that the CADD score of the first variant allele of gene A (CADD1) and that of the first variant allele of gene B (CADD3), along with the gene recessiveness probabilities (RecA, RecB), are the most important features for separating the two

Downloaded at KU Leuven University Library on December 13, 2019

Table 1. Performance of the VarCoPP on independent 10, 30, 100, and 300 random gene panels and on disease gene panels for the BBS and autism/ID genes, iterated 100 times on 100 random 1KGP individuals

	10 Random genes		30 Random genes		100 Random genes		300 Random genes		21 BBS genes		24 SFARI 1 genes		79 SFARI 1 + 2 genes		237 SFARI 1 + 2 + 3 genes	
Gene panels	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Combinations	3.03	1.9	14.12	12.9	143.63	81.8	1,312	463.8	8.83	12.9	11.99	15.9	146.51	161.4	1,672.06	1,548.2
% TNs, SS = 0	74.65	19.5	74.87	15.5	72.87	10.9	73.02	6.5	58.24	35.5	90.04	17.9	86.02	12.9	79.88	6.94
% FPs	7.23	11.6	6.54	8.9	7.93	6.8	7.39	3.4	12.66	20.4	1.99	7.4	2.81	5.0	4.22	3.2
% 95-FPs	4.62	8.9	4.48	7.8	5.53	5.4	5.13	2.7	8.45	16.2	1.39	4.9	2.02	4.1	2.75	2.4
95-FPs	0.16	0.4	0.73	2.1	7.15	7.2	67.27	50.4	1.04	2.4	0.19	0.6	3.18	6.6	48.71	58.6
% 99-FPs	0.81	2.7	0.78	2.4	0.88	1.2	0.88	0.7	2.44	7.3	0.44	2.6	0.46	1.3	0.48	0.76
99-FPs	0.03	0.1	0.11	0.4	1.16	1.5	11.86	11.5	0.35	0.9	0.03	0.2	0.67	1.9	7.80	11.7

95-FPs, FPs falling in the 95% confidence zone; 99-FPs, FPs falling in the 99% confidence zone; SFARI 1, high-confidence category; SFARI 2, strong candidate category; SFARI 3, suggestive evidence category.

bilocus combination classes. Their capacity to differentiate between pathogenic and neutral combinations becomes clear by comparing their value distributions between the two sets (*SI Appendix*, Fig. S4).

Although the CADD pathogenicity of variants is important for the VarCoPP to classify a variant combination, using CADD1 and CADD3 alone (the CADD scores of the most pathogenic variant alleles of each gene inside a combination) is not sufficient to achieve satisfactory results (run with two features in *SI Appendix*, Fig. S3). By adding information about the genes' recessiveness (run with four features in *SI Appendix*, Fig. S3), we see an improvement in classification, but it is the addition of the complete biological information (i.e., all 11 selected features) that provides the best performance. Therefore, it is the synergy of all features that contributes to the correct classification of a bilocus combination, underlining the necessity of developing a tool like the VarCoPP compared with solely using combinations based on single-variant pathogenicity information.

From Black-Box to White-Box Predictions That also Explain Classification Decisions. Since it is the synergy between the features that determines whether a particular variant combination in a pair of genes is pathogenic or not, their joint impact on the prediction process should provide an even better understanding of how the VarCoPP makes its decisions. Understanding this decision process transforms the VarCoPP from a black-box predictor into a white-box predictor, an issue that is becoming more and more important as these artificial decision makers may have a crucial impact on patients and people in general.

Using a method that follows the decision steps for each new bilocus combination in each individual predictor inside the VarCoPP (*Materials and Methods* and *SI Appendix*, Text S5), we can show the preference of each feature for either the neutral or disease-causing class. That preference or decision gradient can be either positive or negative, depending on whether the feature pushes the decision to the pathogenic or neutral class, respectively. For example, in the DIDAv1, most disease-causing combinations are between genes that correspond to proteins that are directly or indirectly (i.e., separated by one intermediate protein) interacting. Thus, if the biological distance feature between the two genes of a variant combination is rather low, meaning that the genes are very close in the protein–protein interaction network, the decision gradient for the biological distance feature will be positive, driving the prediction toward the pathogenic class. Performing that preference analysis for each feature and individual predictor inside the VarCoPP when predicting a bilocus combination, a distribution of decision gradient values for that feature is produced. The simplest way to visualize these values per feature is by using boxplots that reveal both the median and variance among the individual predictors in the VarCoPP (as can be seen in Fig. 6, where positive values that vote for the pathogenic class are depicted in red, while negative values voting for the neutral class are depicted in blue).

The higher the confidence of a prediction, the more clearly it is that features show preference for a particular class. As can be seen in Fig. 6, there is a clear positive or negative preference among the features for cases where there is full support for either the disease-causing (Fig. 6*A*, SS = 100) or the neutral (Fig. 6*B*, SS = 0) class for a bilocus combination. However, in cases where the prediction is ambiguous, such as, for example, in cases where the average support from the individual predictors in the VarCoPP is close to the threshold (SS ~ 50), we observe that such a clear consensus among the features is missing (Fig. 6*C*).

These visualizations provide a good indication as to why we reach a disease-causing or neutral prediction for a bilocus variant combination. By examining the actual values of the features that most strongly influence the decision process for a particular combination [i.e., are furthest away from zero (Fig. 6)], we can obtain insight into why this combination is assigned to a specific class and people can assess whether they agree with this assignment. For example, if we see that the CADD1 feature of a variant combination drives significantly the prediction toward the pathogenic class, we can most probably expect that the CADD feature value of the corresponding variant is relatively high.

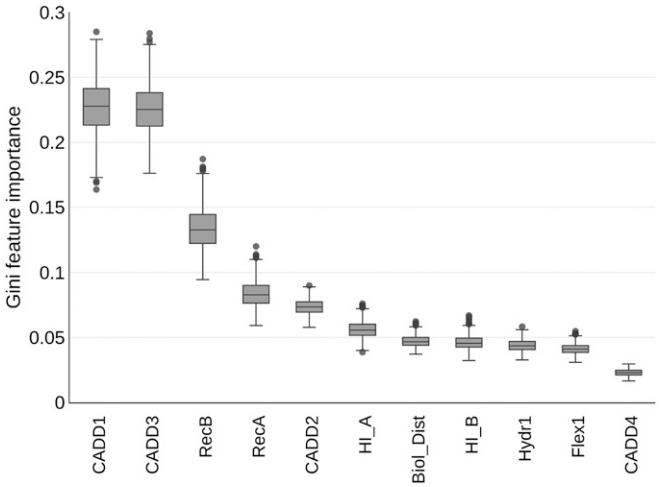


Fig. 5. Boxplot of the Gini importance for each feature among all 500 individual predictors of the VarCoPP using the training DIDA and 1KGP data.

Downloaded at KU Leuven University Library on December 13, 2019

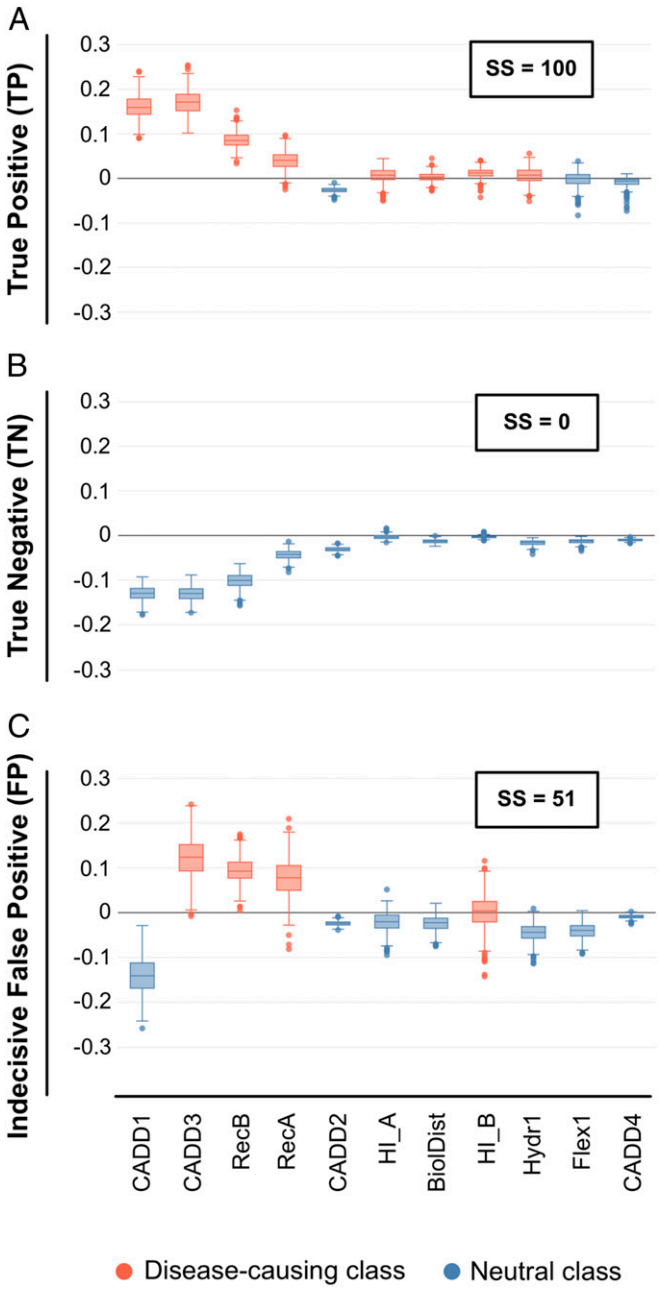


Fig. 6. Decision profile (DP) boxplots that show the class preference (or decision) gradients of each feature used for the classification of test bilocus combinations. Features whose median decision gradient values, among all classifiers of the VarCoPP, fall above zero on the *y* axis are in favor of the disease-causing class (red color), whereas features whose median decision gradient values fall below zero on the *y* axis are in favor of the neutral class (blue color). (*A*) DP boxplot for a TP bilocus combination with SS = 100 (Dataset S1, testpos_21), where the vast majority of features have a median decision value above zero. (*B*) DP boxplot for a TN bilocus combination with SS = 0 (Dataset S3, testneg_769), where all features have a median decision value below zero agreeing for the neutral class. (*C*) Example of an indecisive DP boxplot for a neutral bilocus combination of the set of 1,000 test neutral combinations, which was predicted as disease-causing with SS = 51 (Dataset S3, testneg_358).

Discussion

This work demonstrates that sufficient genetic knowledge is available to produce pathogenicity predictors capable of differentiating between pathogenic and neutral bilocus variant combinations. We presented here the VarCoPP, a clinically competent predictive tool, which is precise and sensitive both in cross-validation settings (87% correct predictions) and also when tested on new independent data. Its performance will further increase by improving the quality of the genetic annotations and by using more training data.

The VarCoPP provides robust 95% and 99% confidence labels, which constitute an objective assessment of the relevance of newly identified pathogenic bilocus variant combinations. These zones are important as a form of primary filtering and evaluation of the predictions, while further statistical and biological verification can be performed for those 95% confidence- and 99% confidence-labeled variant combinations. Such an approach boosts the clinical relevance of the VarCoPP as it limits the search space produced by all variant combinations of a gene panel or exome to the most relevant ones and, as a consequence, reduces the required time needed to further explore these relevant results.

Moreover, our method has been designed to produce "white-box" predictions by providing insights into the importance of the biological features in distinguishing disease-causing combinations from neutral ones (Fig. 5). Furthermore, it can provide objective explanations on the class decision made by the predictor for each new bilocus combination that is being tested (Fig. 6). While the former provides a way to assess the relevance of novel features in further developments of the VarCoPP, the latter allows users to assess the relevance of the prediction using their genetic and biological expertise and to capture reasoning differences for different bilocus instances. Providing such decision transparency for automated systems is highly important, given the effect that predictions may have on individuals and society.

Although we can now start to analyze combinations in patient exomes, it is important to keep in mind that the magnitude of the search space increases dramatically when moving to a full-exome analysis. Although there is only a 1% chance of observing a FP in the 99% confidence zone, the absolute number of FP combinations will exponentially increase, a classic problem that is unfortunately encountered in most types of bioinformatics predictors when tested at the exome level. Additional pre- or postfiltering steps to reduce these absolute numbers are thus required, which can be done, for instance, by adding knowledge about the disease or comparing the predictions with genetic information obtained for the parents in trio studies. In line with the former, the study can be limited to gene panels known to be associated with the disease or belonging to the relevant pathways. We demonstrated that such a focus will indeed help in limiting the number of nonrelevant bilocus combinations: Using a panel of 150 random genes produces potentially one nonrelevant combination in the 99% confidence zone, confirming the clinical relevance of our method. Furthermore, rare diseases' recessive gene panels (like BBS) may produce a bit more FPs, in contrast to known haploinsufficient gene panels (like those of neurodevelopmental disorders). Clinical users of the VarCoPP should be aware of this issue in the analysis of their target disease.

The results furthermore show that especially the CADD scores of the first variant allele of each gene, an expected observation as we order the variant alleles inside each gene based on pathogenicity, but also the gene recessiveness probabilities seem to be the main drivers of predictions. Although these features independently show great importance, it is the combination of all 11 selected features, including those with a lower effect, that leads to the highest classification accuracy. These results make the VarCoPP a clinically important tool that is more informative and accurate than simply selecting potentially relevant variant combinations based solely on monogenic variant pathogenicity scores, such as the CADD scores.

Further expansions of the VarCoPP into the oligogenic realm should consider that the variant filtering criteria that were shown to be important for the method differ from the "strict" criteria that are commonly used to identify pathogenic variants in rare

Downloaded at KU Leuven University Library on December 13, 2019

Mendelian diseases (i.e., rare exonic variants with a strong monogenic effect). Although the majority of positive cases in the DIDA have a MAF of less than or equal to 3%, we observed that some variants involved in rare oligogenic diseases can reach, for instance, a MAF of up to 18% (54). As these are present but constitute exceptions in the current DIDA dataset, we restrict ourselves for now to a MAF of 3% for the creation of the neutral dataset as well. Nonetheless, this threshold can be further relaxed in the future as more data on pathogenic bilocus combinations become available.

Similarly, while it is widely presumed that genes involved in the same disease can belong to the same molecular pathway or biological process, this does not necessarily apply to all cases. It is shown in the DIDAv1 that for some gene pairs, such as the ANOS1-PROKR2 pair found in many studies associated with Kallman syndrome (46, 55–57), no interaction or coexpression information is known yet, indicating that potentially more complex pathways and cellular mechanisms may be involved to cause disease. Nonetheless, the gene pairs in the neutral data used by the VarCoPP were filtered in such a way that they contained genes with a similar distance distribution from a protein–protein interaction network (i.e., degrees of separation) as the one observed in the DIDAv1. As a consequence, the importance of the biological distance feature, which is strongly related to this degrees of separation metric, is reduced. It remains to be seen whether this stratification should not be relaxed when moving into the realm of oligogenic disease cases, as subsets of genes involved in different pathways may be responsible for the observed phenotype. However, relaxing this biological distance normalization will lead to a less "clever" predictor with a slightly higher FP rate, as it would provide an obvious way to learn separating known pathogenic from random neutral bilocus combinations (*SI Appendix*, Fig. S5).

The VarCoPP is a bilocus variant combination pathogenicity predictor that is trained using combinations involved in known oligogenic diseases. As our predictor is not phenotypically driven, it could also be used to predict bilocus combinations involved in cases of dual molecular diagnosis (i.e., cases where several independent monogenic diseases are present in an individual due to the segregation of monogenic variants in two unrelated loci). The recent work of Posey et al. (21) provides a collection of such dual molecular diagnosis cases. An analysis of 76 cases in that paper revealed that the VarCoPP predicted 67 (88%) correctly (*SI Appendix*, Fig. S6 and Dataset S5). These results are again very promising, especially since dual diagnosis cases are almost completely missing from the DIDAv1. Nonetheless, such cases appear to consist of strong monogenic variants and genes whose nature and properties are different compared with those causing or modulating the diseases contained in the DIDAv1. Within the context of another study, expanding on Gazzo et al. (39), it is observed that dual diagnosis instances are indeed separated from the other types of bilocus diseases. Although further developments of the VarCoPP should incorporate these cases for training, a distinction should be further made between dual diagnosis instances with distinct and overlapping phenotypes. Especially the latter appear to be relevant for a predictor that aims to find synergies between variants, which is the long-term ambition of the VarCoPP.

In conclusion, the VarCoPP reveals that the first steps to multivariant pathogenicity predictions can be taken. Our method shows great predictive ability during cross-validation and using independent validation sets, which may be further improved with the advent of new data and the inclusion of additional biological information. The provision of statistical evaluations, as well as white-box explanations on the obtained results, establishes the VarCoPP as a pioneering clinical tool for the detection of disease-causing variants implicated in more complex genetic patterns. By scoring bilocus combinations and gene pairs, gene triplets or quadruplets may be identified in exome or gene panel data as causative genetic models for a particular disease, paving the path for the detection of multilocus signatures derived with machine learning approaches. The VarCoPP therefore provides an important leap forward, allowing for more fine-grained pathogenic predictions.

Materials and Methods

An illustrated summary of the materials and methods used in this study is presented in Fig. 3. Additional details on each subsection in this section can be found in *SI Appendix*, Text S5.

Data Filtering and Annotation. We filtered the variants and genes between the DIDAv1 and the 1KGP so that both sets contained comparable information (Fig. 3*A*), using exonic and splicing SNPs, as well as indels of MAF equal to or less than 3%. Individuals in the 1KGP who carried disease-causing bilocus combinations, as well as the corresponding overlapping combinations in the DIDAv1, were removed (Fig. 2 and *SI Appendix*, Table S3). Variants and genes inside each bilocus combination were ordered in both datasets so that gene A and the first variant allele of each gene in a bilocus combination were the most pathogenic ones according to the Gene Damage Index score (58) and CADD score (12), respectively (Fig. 3*B*). We then annotated both sets based on information at the variant, gene, and gene-pair levels, leading initially to 21 features per entry. After a feature selection procedure, this set was reduced to 11 features (an overview and explanation of the features are provided in Fig. 3*C* and *SI Appendix*, Text S5 and Tables S4 and S5).

Stratification of the 1KGP Data and Training. To train the VarCoPP, we created 500 balanced sets (Fig. 3*D*), each consisting of 200 1KGP bilocus combinations of randomly chosen gene pairs and the 200 disease-causing combinations of the DIDAv1. For each 1KGP subset, we included 40 individuals per continent. However, we observed that there is no significant difference in performance when the predictor is trained using 1KGP combinations only from individuals of a particular continent against the DIDAv1, confirming no population bias (*SI Appendix*, Table S10). Each random control subset contained gene pairs following a degrees of separation distribution equal to that of the DIDAv1, based on information obtained from the Human Gene Connectome tool (59) (*SI Appendix*, Fig. S5). We used the scikit-learn version 0.18.1 implementation (60) of the Random Forest (RF) algorithm (53) as a classifier for each of the 500 balanced sets. Each RF consisted of 100 decision trees using bootstrapping with a maximum tree depth of 10, using the square root of the features for each split. We implemented a leave-one-pair-out stratified cross-validation procedure individually for each predictor (39).

Validation of the VarCoPP. We collected 23 new disease-causing bilocus combinations derived from independent scientific papers, which were published after the release of the DIDAv1 (Fig. 3*E*, *SI Appendix*, Table S7, and Dataset S1). For confidence testing, we collected different sets of random 100, 1,000, and 10,000 neutral bilocus combinations from the 1KGP that were unused during training (Fig. 3*E* and Datasets S2–S4). For the gene panel analysis, we created random panels consisting of 10, 30, 100, and 300 genes and tested each gene panel on 100 random 1KGP individuals, with 100 iterations. For the BBS analysis, we used the 21-gene list obtained from the Genome Diagnostics Nijmegen laboratory (http://gdnm.nl/en/), and for autism/ID, we used the SFARI Gene panels (https://gene.sfari.org/).

Feature Selection and Interpretation. We applied a recursive feature elimination procedure (61) on a balanced set with median performance among all sets leading to a performance peak with 10 features (*SI Appendix*, Fig. S3). As no variant features about the second variant allele of gene B remained, we included for interpretability reasons the CADD score of this allele (CADD4), finalizing the number of selected features to 11. To create the decision boxplots per bilocus combination, we used the "treeinterpreter" Python package (https://github.com/andosa/treeinterpreter).

Tool and Code Availability. The VarCoPP can be accessed online at http://varcopp.ibsquare.be/ (40). This online tool annotates a list of given variants (single-nucleotide polymorphisms and indels) and scores all possible bilocus variant combinations present in that list, including those with heterozygous compound variants. The source code to reproduce the performance results of the VarCoPP for the training dataset and the validation sets is present at https://github.com/sofiapapad90/VarCoPP (62).

Downloaded at KU Leuven University Library on December 13, 2019

ACKNOWLEDGMENTS. We thank all the members of the Interuniversity Institute for Bioinformatics in Brussels, especially the group of people interested in digenic and oligogenic diseases, for their comments and valuable suggestions. This work was supported by funding from the Actions de Recherche Concertées project Deciphering Oligo- and Polygenic Genetic Architecture in Brain Developmental Disorders (A.G., N.V., C.N., and T.L.); the European Regional Development Fund (ERDF) and the Brussels-Capital Region-Innoviris within the framework of the Operational Programme 2014–2020 through the ERDF-2020 project ICITY-RDI.BRU (27.002.53.01.4524; S.P., A.N., S.V.D., and T.L.); the Fonds de la Recherche Scientifique-Fonds National de la Recherche Scientifique Fund for Research Training in Industry and Agriculture (S.P.); Vrije Universiteit Brussel PhD funding (S.P.); and the Vrije Universiteit Brussel, Reproduction and Genetics and Regenerative Medicine Cluster, Reproduction, and Genetics Research Group (A.G. and S.V.D.).

1. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. W. Fu *et al*.; NHLBI Exome Sequencing Project, Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013). Erratum in: *Nature* **495**, 270 (2013).
3. M. Lek *et al*.; Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
4. NHLBI GO Exome Sequencing Project (ESP), Exome Variant Server. http://evs.gs.washington.edu/EVS/. Accessed 15 May 2019.
5. 1000 Genomes Project Consortium, A. Auton *et al*., A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
6. S. B. Ng, D. A. Nickerson, M. J. Bamshad, J. Shendure, Massively parallel sequencing and rare disease. *Hum. Mol. Genet.* **19**, R119–R124 (2010).
7. M. J. Bamshad *et al*., Exome sequencing as a tool for mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
8. J. X. Chong *et al*.; Centers for Mendelian Genomics, The genetic basis of mendelian phenotypes: Discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
9. M. N. Bainbridge *et al*., Whole-genome sequencing for optimized patient management. *Sci. Transl. Med.* **3**, 87re3 (2011).
10. I. A. Adzhubei *et al*., A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
11. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
12. M. Kircher *et al*., A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
13. D. Raimondi, A. M. Gazzo, M. Rooman, T. Lenaerts, W. F. Vranken, Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. *Bioinformatics* **32**, 1797–1804 (2016).
14. V. van Heyningen, P. L. Yeyati, Mechanisms of non-mendelian inheritance in genetic disease. *Hum. Mol. Genet.*, **13** (suppl. 2), R225–R233 (2004).
15. J. L. Badano, N. Katsanis, Beyond Mendel: An evolving view of human genetic disease transmission. *Nat. Rev. Genet.* **3**, 779–789 (2002).
16. A. A. Schäffer, Digenic inheritance in medical genetics. *J. Med. Genet.* **50**, 641–652 (2013).
17. J. R. Lupski, J. W. Belmont, E. Boerwinkle, R. A. Gibbs, Clan genomics and the complex architecture of human disease. *Cell* **147**, 32–43 (2011).
18. R. Chen *et al*., Analysis of 589,306 genomes identifies individuals resilient to severe mendelian childhood diseases. *Nat. Biotechnol.* **34**, 531–538 (2016).
19. R. L. Nussbaum, R. R. McInnes, H. F. Willard, A. Hamosh, "Genetics of common disorders with complex inheritance" in *Thompson & Thompson Genetics in Medicine* (Elsevier/Saunders, Philadelphia, PA, 2007), pp. 151–174.
20. J. F. Robinson, N. Katsanis, "Oligogenic disease." in *Vogel and Motulsky's Human Genetics*, M. R. Speicher, S. E. Antonarakis, A. G. Motulsky, Eds. (Springer-Verlag, Berlin, Germany, 2010), pp. 243–262.
21. J. E. Posey *et al*., Resolution of disease phenotypes resulting from multilocus genomic variation. *N. Engl. J. Med.* **376**, 21–31 (2017).
22. C. R. Scriver, P. J. Waters, Monogenic traits are not simple: Lessons from phenylketonuria. *Trends Genet.* **15**, 267–272 (1999).
23. T. Friedman *et al*., Modifier genes of hereditary hearing loss. *Curr. Opin. Neurobiol.* **10**, 487–493 (2000).
24. A. S. Brooks, B. A. Oostra, R. M. Hofstra, Studying the genetics of Hirschsprung's disease: Unraveling an oligogenic disorder. *Clin. Genet.* **67**, 6–14 (2005).
25. G. R. Cutting, Modifier genes in Mendelian disorders: The example of cystic fibrosis. *Ann. N. Y. Acad. Sci.* **1214**, 57–69 (2010).
26. A. S. Cristino *et al*., Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. *Mol. Psychiatry* **19**, 294–301 (2014).
27. L. E. L. M. Vissers, C. Gilissen, J. A. Veltman, Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17**, 9–18 (2016).
28. K.-I. Goh *et al*., The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8685–8690 (2007).
29. A. Bauer-Mehren *et al*., Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One* **6**, e20284 (2011).
30. Y. Yang *et al*., Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
31. Y. Yang *et al*., Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870–1879 (2014).
32. J. E. Posey *et al*., Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet. Med.* **18**, 678–685 (2016).
33. F. S. Jehee *et al*.; Baylor-Hopkins Center for Mendelian Genomics, Dual molecular diagnosis contributes to atypical Prader-Willi phenotype in monozygotic twins. *Am. J. Med. Genet. A.* **173**, 2451–2455 (2017).
34. N. Katsanis, The continuum of causality in human genetic disorders. *Genome Biol.* **17**, 233 (2016).
35. A. M. Gazzo *et al*., DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Res.* **44**, D900–D907 (2016).
36. N. Katsanis, The oligogenic properties of Bardet-Biedl syndrome. *Hum. Mol. Genet.* **13**, R65–R71 (2004).
37. O. M'hamdi, I. Ouertani, H. Chaabouni-Bouhamed, Update on the genetics of bardet-biedl syndrome. *Mol. Syndromol.* **5**, 51–56 (2014).
38. T. P. Dryja, L. B. Hahn, K. Kajiwara, E. L. Berson, Dominant and digenic mutations in the peripherin/RDS and ROM1 genes in retinitis pigmentosa. *Invest. Ophthalmol. Vis. Sci.* **38**, 1972–1982 (1997).
39. A. Gazzo *et al*., Understanding mutational effects in digenic diseases. *Nucleic Acids Res.* **45**, e140 (2017).
40. VarCoPP. http://varcopp.ibsquare.be/. Accessed 17 January 2018.
41. H. He *et al*., Involvement of and interaction between WNT10A and EDA mutations in tooth agenesis cases in the Chinese population. *PLoS One* **8**, e80393 (2013).
42. B. N. Cockburn *et al*., 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**, 971–978 (2004).
43. A. Chapla *et al*., Maturity onset diabetes of the young in India–A distinctive mutation pattern identified through targeted next-generation sequencing. *Clin. Endocrinol.* **82**, 533–542 (2015).
44. M. Font-Llitjós *et al*., New insights into cystinuria: 40 new mutations, genotype-phenotype correlation, and digenic inheritance causing partial phenotype. *J. Med. Genet.* **42**, 58–68 (2005).
45. Z. Gucev *et al*., Cystinuria AA (B): Digenic inheritance with three mutations in two cystinuria genes. *J. Genet.* **90**, 157–159 (2011).
46. J. Sarfati *et al*., 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**, 659–669 (2010).
47. K. Zhang *et al*., Synergistic defects of different molecules in the cytotoxic pathway lead to clinical familial hemophagocytic lymphohistiocytosis. *Blood* **124**, 1331–1334 (2014).
48. Z. Sun *et al*., A novel ensemble method for classifying imbalanced data. *Pattern Recognit.* **48**, 1623–1637 (2015).
49. M. A. Mencarelli *et al*., Evidence of digenic inheritance in Alport syndrome. *J. Med. Genet.* **52**, 163–174 (2015).
50. C. Mouden *et al*., Complex mode of inheritance in holoprosencephaly revealed by whole exome sequencing. *Clin. Genet.* **89**, 659–668 (2016).
51. F. Coppieters *et al*., Genetic screening of LCA in Belgium: Predominance of CEP290 and identification of potential modifier alleles in AHI1 of CEP290-related phenotypes. *Hum. Mutat.* **31**, E1709–E1766 (2010).
52. A. Brehm *et al*., Additive loss-of-function proteasome subunit mutations in CANDLE/PRAAS patients promote type I IFN production. *J. Clin. Invest.* 125, 4196–4211 (2015). Erratum in: *J. Clin. Invest.* **126**, 795 (2016).
53. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
54. M. Girardelli, J. Vuch, A. Tommasini, S. Crovella, A. M. Bianco, Novel missense mutation in the NOD2 gene in a patient with early onset ulcerative colitis: Causal or chance association? *Int. J. Mol. Sci.* **15**, 3834–3841 (2014).
55. C. Dodé *et al*., Kallmann syndrome: Mutations in the genes encoding prokineticin-2 and prokineticin receptor-2. *PLoS Genet.* **2**, e175 (2006).
56. P. Canto, P. Munguía, D. Söderlund, J. J. Castro, J. P. Méndez, Genetic analysis in patients with Kallmann syndrome: Coexistence of mutations in prokineticin receptor 2 and KAL1. *J. Androl.* **30**, 41–45 (2009).
57. N. D. Shaw *et al*., Expanding the phenotype and genotype of female GnRH deficiency. *J. Clin. Endocrinol. Metab.* **96**, E566–E576 (2011).
58. Y. Itan *et al*., The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13615–13620 (2015).
59. Y. Itan *et al*., HGCS: An online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics* **15**, 256 (2014).
60. F. Pedregosa *et al*., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
61. I. Guyon, J. Weston, S. Barnhill, Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
62. S. Papadimitriou *et al*., VarCoPP. Github. https://github.com/sofiapapad90/VarCoPP/. Deposited 31 March 2019.

Downloaded at KU Leuven University Library on December 13, 2019