frontiers in Immunology

METHODS
published: 27 June 2018
doi: 10.3389/fimmu.2018.01340

CDG: An Online Server for Detecting Biologically Closest Disease-Causing Genes and its Application to Primary Immunodeficiency

*David Requena[1]†, Patrick Maffucci[1,2,3]†, Benedetta Bigio[1], Lei Shang[1], Avinash Abhyankar[4], Bertrand Boisson[1,5,6], Peter D. Stenson[7], David N. Cooper[7], Charlotte Cunningham-Rundles[2,3], Jean-Laurent Casanova[1,5,6,7,8,9], Laurent Abel[1,5,6] and Yuval Itan[10,11]**

[1] St. Giles Laboratory of Human Genetics of Infectious Diseases (Rockefeller Branch), The Rockefeller University, New York, NY, United States, [2] Graduate School, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [3] Department of Medicine, Division of Clinical Immunology, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [4] New York Genome Center, New York, NY, United States, [5] Laboratory of Human Genetics of Infectious Diseases (Necker Branch), INSERM U1163, Paris, France, [6] Paris Descartes University, Imagine Institute, Paris, France, [7] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom, [8] Howard Hughes Medical Institute, New York, NY, United States, [9] Pediatric Immunology-Hematology Unit, Necker Hospital for Sick Children, Paris, France, [10] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [11] Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, United States

OPEN ACCESS

Edited by:
*Carrie L. Lucas,
Yale University, United States*

Reviewed by:
*Yu Zhang,
National Institutes of Health (NIH), United States
Francisco Javier Espinosa-Rosales,
Instituto Nacional de Pediatria, Mexico
Maja Tarailo-Graovac,
University of Calgary, Canada*

***Correspondence:**
*Yuval Itan
yuval.itan@mssm.edu*

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Primary Immunodeficiencies, a section of the journal Frontiers in Immunology

Received: 16 February 2018
Accepted: 29 May 2018
Published: 27 June 2018

Citation:
Requena D, Maffucci P, Bigio B, Shang L, Abhyankar A, Boisson B, Stenson PD, Cooper DN, Cunningham-Rundles C, Casanova J-L, Abel L and Itan Y (2018) CDG: An Online Server for Detecting Biologically Closest Disease-Causing Genes and its Application to Primary Immunodeficiency. Front. Immunol. 9:1340. doi: 10.3389/fimmu.2018.01340

High-throughput genomic technologies yield about 20,000 variants in the protein-coding exome of each individual. A commonly used approach to select candidate disease-causing variants is to test whether the associated gene has been previously reported to be disease-causing. In the absence of known disease-causing genes, it can be challenging to associate candidate genes with specific genetic diseases. To facilitate the discovery of novel gene-disease associations, we determined the putative biologically closest known genes and their associated diseases for 13,005 human genes not currently reported to be disease-associated. We used these data to construct the closest disease-causing genes (CDG) server, which can be used to infer the closest genes with an associated disease for a user-defined list of genes or diseases. We demonstrate the utility of the CDG server in five immunodeficiency patient exomes across different diseases and modes of inheritance, where CDG dramatically reduced the number of candidate genes to be evaluated. This resource will be a considerable asset for ascertaining the potential relevance of genetic variants found in patient exomes to specific diseases of interest. The CDG database and online server are freely available to non-commercial users at: http://lab.rockefeller.edu/casanova/CDG.

Keywords: disease-causing gene, gene filtering, next-generation sequencing, genomics, human gene connectome

INTRODUCTION

Genetic mutations have been found to underlie a large number of inherited human diseases. In the past decade, refinements in next-generation sequencing techniques (NGS) have made it possible to detect the full set of gene variants in patients. The average human genome contains about 20,000 coding variants and hundreds of thousands of non-coding variants (1). A common approach to identify candidate variants for further investigation from NGS data involves screening for those in known disease-causing genes (2–4). However, variants in novel disease-associated genes should be estimated by computational predictions (5).

Databases such as the Human Gene Mutation Database [HGMD (6)], and ClinVar (7, 8) provide manually curated information about mutations in known disease-causing genes, also known as the Clinome (9). Several methods including the Search Tool for the Retrieval of Interacting Genes/Proteins [STRING (10)], Exomiser that prioritizes genetic variants from a vcf file (11), the Probabilistic functional gene network of Homo Sapiens [HumanNet (12)], and Functional Coupling [FunCoup (13, 14)] can be used to assess human genes directly connected to candidate genes. The human gene connectome [HGC (15)] extends these approaches by prioritizing candidate genes according to their computed biological distances from known disease-causing genes.

We generated a complementary resource, the closest disease-causing genes (CDG) database and server to identify novel gene-disease associations. CDG computes the biologically closest known disease-causing genes and corresponding diseases for 13,005 human candidate genes not currently observed to be disease-causing, allowing investigators to associate these candidate genes with known disease phenotypes. We demonstrate the efficiency of this method in five patients with various primary immunodeficiencies and modes of inheritance, significantly reducing the number of candidate genes in these examples by using CDG (see Supplementary Material, Section 2 for details). CDG also identifies novel gene candidates for lists of diseases defined by an investigator. Thus, this resource provides a reference for the potential relevance of novel candidate genes to specific disease phenotypes, simplifying the analysis of NGS data.

MATERIALS AND METHODS

CDG Generation

Human Gene Mutation Database is a manually curated database of variants that may be associated with or predisposing to human genetic conditions (16, 17). From the HGMD March 2015 public full version (updated through December 2014), we selected 5,430 HGMD genes classified as high-quality disease-causing or disease-associated mutations (mostly linked to monogenic diseases). We next identified 13,005 protein-coding genes present in the HGC that are not currently reported to be disease-causing in the HGMD database. Briefly, the HGC (15) is a network of all human genes (represented as nodes), where each edge represents the direct biological distance between two human genes. Direct biological distance is defined as the inverse confidence score for binding connectivity provided by STRING (10). The HGC biological distance between any two genes is defined as the weighted sum of direct distances in the shortest path connecting two given genes (calculated using the Dijkstra algorithm), on the network containing most protein-coding human genes.

For each of these 13,005 genes, we calculated their biologically CDG and associated diseases by first retrieving the corresponding connectome for each gene from the HGC database (15, 18). A gene-specific connectome contains, for any given human gene, the set of all other human genes ranked by their biological distance to that specific gene. Then, following the HGC criterion for biological relatedness, we selected only the HGMD known disease-causing genes in the connectome within $p < 0.01$. Additionally, we assigned the corresponding human phenotype ontology codes [HPO (19)] to each gene-phenotype association (Figure S1 in Supplementary Material). A summary of the CDG, diseases, and routes associated with each of the 13,005 genes not currently known to be disease-causing is provided in Table S1 in Supplementary Material.

Validation

We validated CDG and compared the performance of CDG with FunCoup and HumanNet using a validation set of genes not used during the construction of the original CDG database. As validation set, we used two external datasets (1) a new HGMD dataset, containing 339 disease-causing genes added between January and September 2015 (i.e., not used to construct CDG); and (2) the pathogenic genes from ClinVar not present in HGMD, comprising 84 genes. We calculated the CDG for each of these genes as described above and compared the performance of CDG versus FunCoup and HumanNet in terms of number of predicted genes and how many predicted diseases coincided with the reported disease. As FunCoup and HumanNet do not associate diseases, we retrieved the disease names related to each predicted gene from HGMD. To compare the predicted and expected disease names, we implemented in CDG the following phrase-comparison procedure (1) first, the disease names were compared by exact coincidence. Then (2) using the "starts-with" comparison: if one phrase exactly starts with the other phrase, or *vice versa*. If at this point no matches were found, we used (3) the Levenshtein distance algorithm (20). All comparisons between disease names for the validation dataset were verified manually.

Data Storage and web Access

To make CDG easily accessible, we created a webserver that allows to consult the CDG database using either genes or diseases as input.

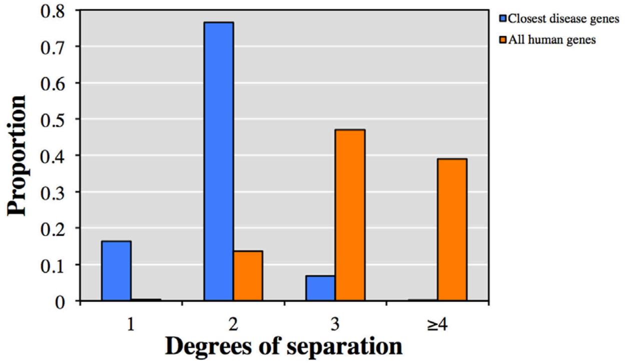


FIGURE 1 | Predicted degrees of separation between (blue) the 13,005 genes from human gene mutation database (HGMD) not known to be disease-causing and their closest predicted HGMD disease-causing genes, and (orange) between all pairs of human genes.

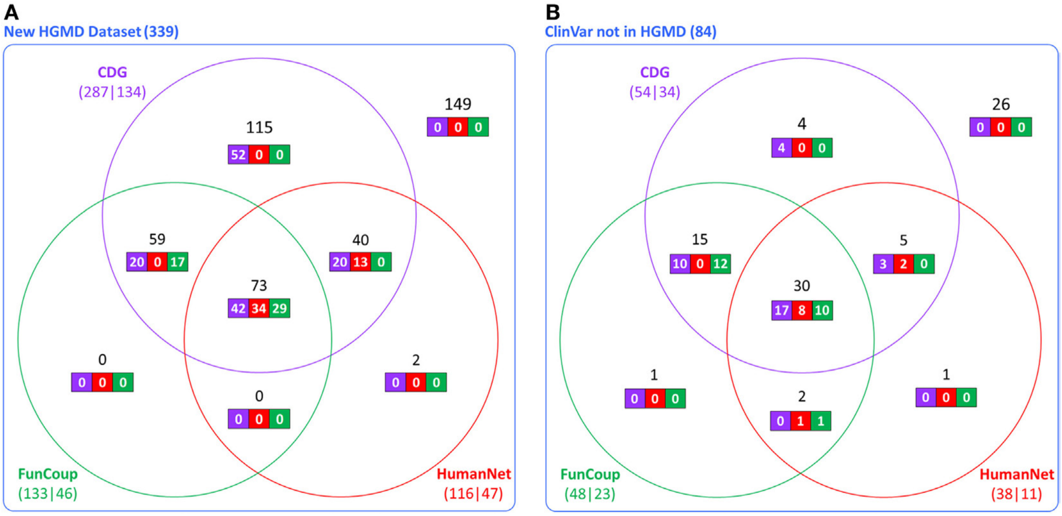


FIGURE 2 | Comparative performance of CDG, FunCoup, and HumanNet using **(A)** 339 new genes in human gene mutation database (HGMD) and **(B)** using 84 genes in ClinVar that are not in HGMD. The numbers below each method show the number of genes with at least one predicted gene (left) and how many were associated with the expected disease (right). Black numbers show the gene distribution across the three servers and white numbers show how many were associated with the expected disease in each server.

If the input gene is known to be disease-causing, the server provides the known associated diseases. And if the gene is unknown to be disease-causing, predicted data is displayed. The server also allows using disease names as input, returning the list of both known and predicted causative genes. The disease names in the CDG database are as reported in HGMD. If the user input is not a HGMD disease name, the procedure to compare disease names described above is used to estimate the closest HGMD disease name.

For the CDG server, MySQL was used to structure and store the multi-dimensional profile of the results of this study, and to process queries to allow efficient access. JSP and servlets were used to parse inputs and generate queries. The web interface is stored on a Rockefeller University Linux-based server in solid state drives. The CDG resource is platform-independent and is freely available to all non-commercial users. The CDG database and server will be periodically updated with new public versions of HGMD, STRING, and HGC.

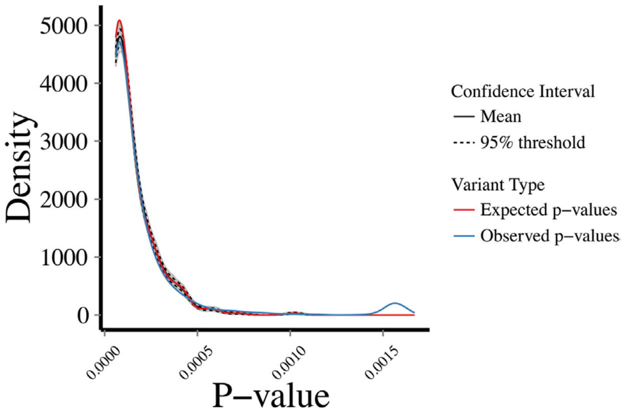


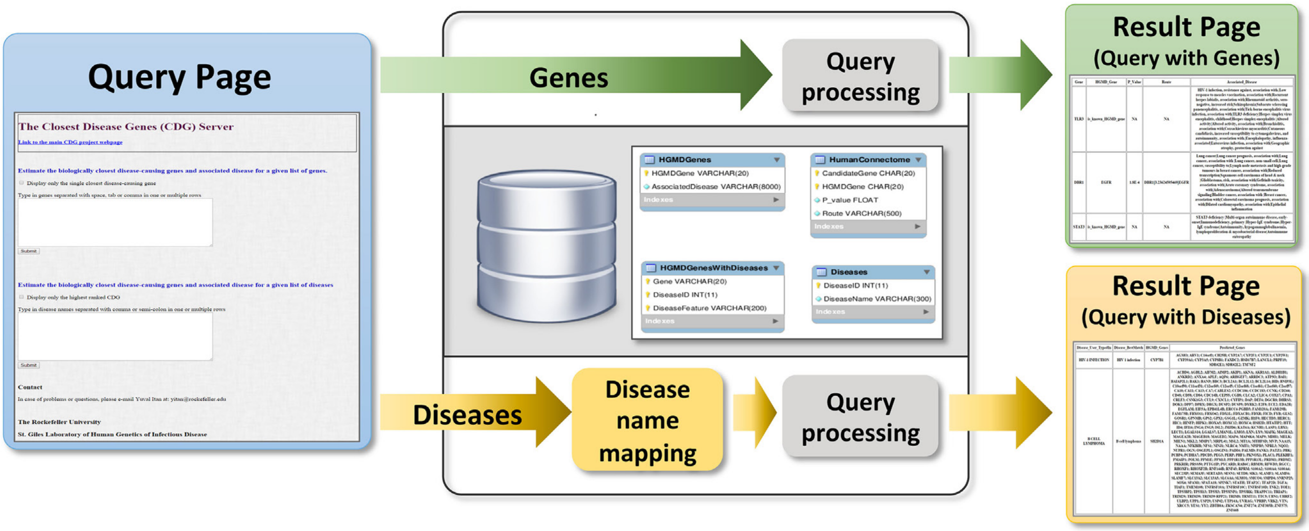
FIGURE 3 | Bootstrapping simulations between a set of (1) expected: *p*-values between 13,005 genes not reported to cause disease and their predicted CDGs; (2) observed: *p*-values between new human gene mutation database genes (i.e., not used to generate the CDGs presented in this study) and their predicted CDGs. Test performed by random sampling using a Gaussian distribution.

RESULTS

CDG Validation and Comparative Analysis

We first explored the relationship between the 13,005 genes not currently described to cause clinical phenotypes with HGMD known disease-causing genes. Each of the 13,005 genes was associated on average with 48 HGMD disease-causing genes and 7 diseases by HGC biological proximity (see Table S1 in Supplementary Material for the top-ranked associations). Notably, 92.9% of the associated disease-causing genes were within one or two degrees of separation from the corresponding query gene (**Figure 1**). Conversely, only 13.9% of all human gene pairs were within one or two degrees of separation ($p < 10^{-300}$, two-tailed equal variance t-test).

The accuracy and utility of these associations was then assessed using new disease-causing genes not known during the construction of the CDG database. Using the first dataset (339 new genes from HGMD), we found that 287 had at least one predicted gene by CDG, compared to 133 using FunCoup and 116 using HumanNet. From these predicted genes, 134 of 287 were associated with the expected disease by CDG, compared to 46 genes of 133 by FunCoup and 47 of 116 by HumanNet (**Figure 2A**). We repeated the comparison using the second dataset (84 genes from ClinVar not present in HGMD) and observed that CDG similarly outperformed the other two software both in number of genes with at least one predicted disease-causing gene and also in correct association with the expected disease (**Figure 2B**).



The Closest Disease Genes (CDG) Server

Link to the main CDG project webpage

Estimate the biologically closest disease-causing genes and associated disease for a given list of genes.

□ Display only the single closest disease-causing gene

Type in genes separated with space, tab or comma in one or multiple rows

Estimate the biologically closest disease-causing genes and associated disease for a given list of diseases

□ Display only the highest ranked CDG

Type in disease names separated with comma or semi-colon in one or multiple rows

Contact

In case of problems or questions, please e-mail Yuval Itan at: yitan@rockefeller.edu

The Rockefeller University

St. Giles Laboratory of Human Genetics of Infectious Disease

FIGURE 4 | Schematic of the closest disease-causing genes (CDG) server pipeline, where CDG can be estimated by queries of gene or disease lists provided by the user.

To address the robustness of the predictions, we randomly sampled 1,000 sets of 287 genes from the 5,430 known disease-causing genes and estimated their CDGs and associated diseases. CDG identified the expected disease in 86.33% of cases by exact disease name match. Then, we examined the profiles of biological proximity for CDG predictions and known disease-causing genes. Assuming a Gaussian distribution, we performed 10,000 bootstrapping simulations for HGC p-values of CDG predictions between the observed 287 new HGMD genes with at least one CDG and the expected set of 13,005 genes not currently known to cause disease. The observed and expected CDG predictions yielded similar p-value profiles for biological relatedness between the observed and expected gene sets and their CDG (**Figure 3**). Therefore, CDG associations are expected to be more robust and relevant for the putative diseases associated with candidate genes than previous methods. Due to the lack of flat files from FunCoup and HumanNet, it was not possible to repeat this analysis with these methods. Thus, we expect that CDG predictions are of significant utility to researchers exploring genes without published phenotypes.

Examples of CDG Usage

Finally, we demonstrated the utility of CDG in WES data in five patients with various primary immunodeficiencies, modes of inheritance, and known mutated genes that were not in the HGMD public database during CDG generation (extended description and flowchart in Supplementary Material, Section 2). Phenotypes and associated genotypes in these examples include (1) severe autoinflammation, a homozygous mutation in *RNF31* (21); (2) Epidermodysplasia verruciformis, a homozygous mutation in *STK4* (*MST1*) (22); (3) herpes simplex encephalitis, a homozygous mutation in *UNC93B1* (23); (4) common variable immunodeficiency, a heterozygous mutation in *IKZF1* (24); and (5) natural killer cell deficiency, compound heterozygous mutations in *GINS1* (25). The range of initial number of genes per patient was 14,800–18,862. We then applied standard QC (DP > 4, MQ > 40, and QD > 2), minor allele frequency (<1%) (26), and gene-level filtering using GDI (27) and MSC (28), reducing the number of genes in each patient to the range from 18 to 322 candidate genes (numbers mostly dependent on mode of inheritance). Finally, applying the CDG server to the number of genes to investigate reduced this range from 1 to 11, a reduction in candidate genes of 92.1–96.6%, without losing any of the pathogenic genes.

CONCLUSION

We provide the first resource by estimating the closest known disease-causing genes and their associated diseases for 13,005 human genes not currently known to be disease-causing. From the comparisons performed, we conclude that CDG predictions capture meaningful candidate disease-causing genes and diseases. We propose to use CDG with lists of genes from NGS studies or similar sources to (1) explore the likelihood of candidate genes being associated with a disease of interest by investigation of its CDGs and associated diseases; (2) rapidly identify known diseases associated with HGMD disease-causing genes; and (3) assign CDGs and associated diseases in variant annotation software. We are also providing an option for users to perform CDG queries based on OMIM (29), although this resource contains less pathogenic mutations compared to HGMD. See Supplementary Material, Section 3 for further details regarding the webserver's construction.

Users can submit genes to the webserver (**Figure 4**) to obtain two outputs (1) all CDGs and associated diseases, including their routes to the input genes (i.e., HGC-predicted genes on the shortest path) and (2) only the most significant CDG for each input gene (by p-value). If the input is a known disease-causing gene, the output will be all known associated diseases. Disease names can also be input to obtain known and predicted disease-causing genes for the phenotype concerned. When CDG does not provide desirable results, we propose to rerun it with different diseases that are phenotypically close to the disease that is investigated. We expect that the use of CDG with gene-level filtering methods, such as the gene damage index (27), will facilitate the discovery of new disease-causing genes. The CDG server will be updated when new public versions of HGMD become available, and new features will be added, including filtering by degrees of separation and phenotype matching. The CDG resource and database are available for download from the main CDG webpage: http://lab.rockefeller.edu/casanova/CDG.

AUTHOR CONTRIBUTIONS

YI initiated the study. PS, DC, CC-R provided data and expertise. DR, PM, BB, PS, and YI analyzed the data. DR and LS generated the webserver. DR performed the comparison. DR, PM, AA, J-LC, LA, and YI wrote the manuscript. J-LC, LA, and YI supervised the study. All the authors revised and approved the final version of the manuscript.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health [AI-101093, AI-086037, AI-048693, AI-088364, and T32-GM007280], the Jeffrey Modell Foundation, and the Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai. We thank Shen-Ying Zhang and Emmanuelle Jouanguy for their help with identifying pathogenic mutations in patients' exomes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/articles/10.3389/fimmu.2018.01340/full#supplementary-material.

TABLE S1 | List of all CDGs, their associated HPO IDs and disease names. Also including the p-values and predicted routes between CDGs and candidate genes.

REFERENCES

1. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* (2013) 14(7):460–70. doi:10.1038/nrg3455
2. Casanova JL, Conley ME, Seligman SJ, Abel L, Notarangelo LD. Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. *J Exp Med* (2014) 211(11):2137–49. doi:10.1084/jem.20140520
3. Maffucci P, Filion CA, Boisson B, Itan Y, Shang L, Casanova JL, et al. Genetic diagnosis using whole exome sequencing in common variable immunodeficiency. *Front Immunol* (2016) 7:220. doi:10.3389/fimmu.2016.00220
4. Meyts I, Bosch B, Bolze A, Boisson B, Itan Y, Belkadi A, et al. Exome and genome sequencing for inborn errors of immunity. *J Allergy Clin Immunol* (2016) 138(4):957–69. doi:10.1016/j.jaci.2016.08.003
5. Itan Y, Casanova JL. Novel primary immunodeficiency candidate genes predicted by the human gene connectome. *Front Immunol* (2015) 6:142. doi:10.3389/fimmu.2015.00142
6. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* (2017) 136(6):665–77. doi:10.1007/s00439-017-1779-6
7. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* (2014) 42:D980–5. doi:10.1093/nar/gkt1113
8. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* (2016) 44(D1):D862–8. doi:10.1093/nar/gkv1222
9. Altman RB, Balling R, Brinkley JF, Coiera E, Consorti F, Dhansay MA, et al. Commentaries on "Informatics and medicine: from molecules to populations". *Methods Inf Med* (2008) 47(4):296–317.
10. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* (2015) 43:D447–52. doi:10.1093/nar/gku1003
11. Smedley D, Jacobsen JO, Jager M, Kohler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc* (2015) 10(12):2004–15. doi:10.1038/nprot.2015.124
12. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* (2011) 21(7):1109–21. doi:10.1101/gr.118992.110
13. Alexeyenko A, Schmitt T, Tjärnberg A, Guala D, Frings O, Sonnhammer ELL. Comparative interactomics with Funcoup 2.0. *Nucleic Acids Res* (2012) 40:D821–8. doi:10.1093/nar/gkr1062
14. Schmitt T, Ogris C, Sonnhammer EL. FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res* (2014) 42:D380–8. doi:10.1093/nar/gkt984
15. Itan Y, Zhang SY, Vogt G, Abhyankar A, Herman M, Nitschke P, et al. The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci U S A* (2013) 110(14):5558–63. doi:10.1073/pnas.1218167110
16. Johnston JJ, Biesecker LG. Databases of genomic variation and phenotypes: existing resources and future needs. *Hum Mol Genet* (2013) 22(R1):R27–31. doi:10.1093/hmg/ddt384
17. Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. *J Mol Biol* (2013) 425(21):4047–63. doi:10.1016/j.jmb.2013.08.008
18. Itan Y, Mazel M, Mazel B, Abhyankar A, Nitschke P, Quintana-Murci L, et al. HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics* (2014) 15:256. doi:10.1186/1471-2164-15-256
19. Kohler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Ayme S, et al. The human phenotype ontology in 2017. *Nucleic Acids Res* (2017) 45 (D1):D865–76. doi:10.1093/nar/gkw1039
20. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Sov Phys Dokl* (1966) 10(8):707–10.
21. Boisson B, Laplantine E, Dobbs K, Cobat A, Tarantino N, Hazen M, et al. Human HOIP and LUBAC deficiency underlies autoinflammation, immunodeficiency, amylopectinosis, and lymphangiectasia. *J Exp Med* (2015) 212(6):939–51. doi:10.1084/jem.20141130
22. Crequer A, Picard C, Patin E, D'Amico A, Abhyankar A, Munzer M, et al. Inherited MST1 deficiency underlies susceptibility to EV-HPV infections. *PLoS One* (2012) 7(8):e44010. doi:10.1371/journal.pone.0044010
23. Casrouge A, Zhang SY, Eidenschenk C, Jouanguy E, Puel A, Yang K, et al. Herpes simplex virus encephalitis in human UNC-93B deficiency. *Science* (2006) 314(5797):308–12. doi:10.1126/science.1128346

24. Kuehn HS, Boisson B, Cunningham-Rundles C, Reichenbach J, Stray-Pedersen A, Gelfand EW, et al. Loss of B cells in patients with heterozygous mutations in IKAROS. *N Engl J Med* (2016) 374(11):1032–43. doi:10.1056/NEJMoa1512234
25. Cottineau J, Kottemann MC, Lach FP, Kang YH, Vely F, Deenick EK, et al. Inherited GINS1 deficiency underlies growth retardation along with neutropenia and NK cell deficiency. *J Clin Invest* (2017) 127(5):1991–2006. doi:10.1172/JCI90727
26. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature* (2015) 526 (7571):68–74. doi:10.1038/nature15393
27. Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Velez M, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci U S A* (2015) 112(44):13615–20. doi:10.1073/pnas.1518646112
28. Itan Y, Shang L, Boisson B, Ciancanelli MJ, Markle JG, Martinez-Barricarte R, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat Methods* (2016) 13(2):109–10. doi:10.1038/nmeth.3739
29. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: online Mendelian inheritance in man [OMIM(R)], an online catalog of human genes and genetic disorders. *Nucleic Acids Res* (2015) 43:D789–98. doi:10.1093/nar/gku1205

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.