DNA Language Model and Interpretable Graph Neural Network Identify Genes and Pathways Involved in Rare Diseases

Ali Saadat, Jacques Fellay
School of Life Sciences
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{ali.saadat, jacques.fellay}@epfl.ch

Abstract

Identification of causal genes and pathways is a critical step for understanding the genetic underpinnings of rare diseases. We propose novel approaches to gene prioritization and pathway identification using DNA language model, graph neural networks, and genetic algorithm. Using HyenaDNA, a long-range genomic foundation model, we generated dynamic gene embeddings that reflect changes caused by deleterious variants. These gene embeddings were then utilized to identify candidate genes and pathways. We validated our method on a cohort of rare disease patients with partially known genetic diagnosis, demonstrating the re-identification of known causal genes and pathways and the detection of novel candidates. These findings have implications for the prevention and treatment of rare diseases by enabling targeted identification of new drug targets and therapeutic pathways.

1 Introduction

The landscape of genomics research has undergone a profound transformation with the advent of high-throughput sequencing technologies (Metzker, 2009). The generation of a vast amount of genomics data offers unprecedented insights into human genetic diversity (Auton et al., 2015; Chen et al., 2023). However, this wealth of data brings significant challenges in terms of data analysis and interpretation. A main challenge in deciphering the underlying mechanisms of diseases is establishing a link between genotype and phenotype (Gallagher and Chen-Plotkin, 2018). This task becomes even harder in the context of rare diseases, where the scarcity of data reduces statistical power (Seaby and Ennis, 2020).

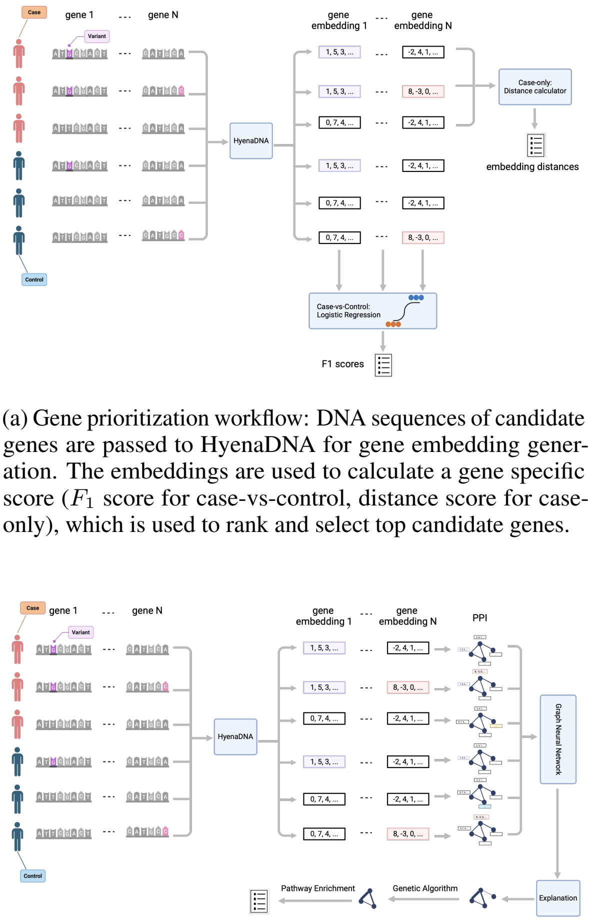
Traditional methods for finding disease-associated genes/pathways have predominantly relied on statistical approaches, such as correlating specific genetic variants with disease occurrence (Auer and Lettre, 2015; Uffelmann et al., 2021). These approaches show decent performance if the cohort size is large, which is often a big obstacle in rare disease studies. Moreover, these methods usually utilize basic variant statistics (such as number of variant carriers), and might not take into account the gene-specific impact of variants on the gene sequence (MacArthur et al., 2014).

Another family of computational approaches for gene/pathway prioritization rely on the concept of guilt-by-association, where genes/pathways are considered potentially relevant based on their similarity to known disease genes (Lee et al., 2011; Guala and Sonnhammer, 2017). These methods work well in scenarios where some underlying genetic factors of the phenotype are well-studied, which is not the case for many diseases (Amberger et al., 2018). Moreover, these methods might introduce bias since they look for similar genes, thereby missing novel disease-causing genes (Gillis and Pavlidis, 2012).

Recent years have seen a remarkable rise in the performance of language models, particularly in the field of natural language processing (NLP) (Devlin et al., 2018; Radford et al., 2019). These models 'learn' language by processing vast amounts of text data, enabling them to perform a wide range of downstream tasks such as translation, summarization, and question-answering with unprecedented accuracy and fluency (Zhao et al., 2023). Parallel to this development, the concept of language models has been applied to genomics, giving rise to DNA language models (DNA-LMs) (Zhou et al., 2023; Dalla-Torre et al., 2023; Benegas et al., 2023; Nguyen et al., 2023). Genomic sequences, much like textual data, comprise long chains of information, in this case nucleotides instead of words. DNA-LMs apply the principles of NLP to interpret and analyze these sequences, translating the 'language' of DNA into meaningful biological insights. By learning from extensive genomic data, these models can provide new perspectives on downstream biological processes (Consens et al., 2023; Marin et al., 2023).

HyenaDNA (Nguyen et al., 2023) is a long-range genomic foundation model pre-trained on the human reference genome at single nucleotide resolution. It can process long-range DNA sequences and represent them as embeddings in a high-dimensional space. For any genomic region such as a gene, HyenaDNA generates embeddings that capture the inherent information of the DNA sequence. These embeddings dynamically change in response to genetic variants, offering insights into how genetic alterations impact biological processes.

We hypothesize that variants with strong deleterious effects have a detectable impact on gene embeddings. We designed complementary methods to identify genes and pathways that contain such deleterious variants and could therefore play a causal role in the pathogenesis of rare diseases. For gene prioritization, we propose two approaches (case-vs-control and case-only) to quantitatively rank candidate genes (Figure 1a). For pathway identification, we propose a method that combines DNA-LM, interpretable graph neural networks (GNN) (Wu et al., 2021; Ying et al., 2019) and Genetic Algorithm (Katoch et al., 2020) (Figure 1b). We validate our methods on a cohort of rare disease patients with partially known genetic diagnosis, demonstrating the re-identification of known causal genes and the detection of novel candidates.



(a) Gene prioritization workflow: DNA sequences of candidate genes are passed to HyenaDNA for gene embedding generation. The embeddings are used to calculate a gene specific score (F_1 score for case-vs-control, distance score for case-only), which is used to rank and select top candidate genes.

(b) Pathway identification workflow: for each individual, a protein-protein interaction network is constructed with gene embeddings as node features. A GNN is trained to classify cases graphs from controls, and GNNExplainer is applied to score the importance of nodes and edges for graph classification. Afterwards, a Genetics Algorithm is used to find the most explainable subnetwork, and a pathway enrichment analysis is performed on that subnetowrk to find the over-represented biological pathways.

Figure 1: Overall summary of of the methods. Figures created with `BioRender.com`.

2 Methods

2.1 Study participants

We selected two cohorts from our in-house database of exome-sequenced individuals. The first cohort consists of 120 previously healthy children who were admitted to pediatric intensive care units (PICUs) with respiratory failure due to a common viral respiratory infection. This cohort serves as the “rare disease” patient group for this study. As control group, we selected a total of 172 healthy individuals. The studies were approved by the relevant ethics commissions and all study participants provided a signed informed consent for research including human genetic testing.

2.2 Short-read alignment and variant calling

Adapter sequences were trimmed from sequencing reads using fastp (Chen et al., 2018) and the reads were subsequently aligned against the human reference genome (hg38) using the maximum exact matches algorithm in Burrows-Wheeler Aligner (Li and Durbin, 2009). The Genome Analysis Software Kit (GATK4) best-practice pipeline was used to call variants in the multi-sample mode (DePristo et al., 2011). In summary, PCR duplicates were removed and base quality scores were recalibrated to correct for sequencing artifacts. We called individual-level variants with GATK HaplotypeCaller before combining single-sample callsets for joint genotyping. To exclude low quality variants, we applied variant quality score recalibration and manual filtering (depth $\geq$ 20, genotype quality $\geq$

20, and 0.2 $\leq$ heterozygous allele balance $\leq$ 0.8).

2.3 Variant annotation and filtering

To predict the potential impact of each variant, we used Variant Effect Predictor (VEP) (McLaren et al., 2016). To identify loss-of-function variants, we used Loss-of-Function Transcript Effect Estimator (LOFTEE) as a VEP plugin (Karczewski et al., 2020).

To classify the variant into putative pathogenicity groups, we implemented the ACMG/AMP guidelines (Richards et al., 2015) in R (`https://www.r-project.org`) (see full description Appendix A). A probability of pathogenicity (PoP) was assigned to each variant according to the ACMG/AMP Bayesian classification framework (Tavtigian et al., 2018). Variants with PoP $\geq$ 0.9 were considered as damaging. Genes with at least one pathogenic variant were included in the downstream analysis.

2.4 Gene embedding calculation

For candidate gene selection, we kept the genes that passed the following criteria: 1) At least one patient carries $\geq$ 1 pathogenic variant in the gene. 2) The length of the gene (including exons, introns, 3'-UTR, and 5'-UTR) is less than 450,000 nucleotides, which is the maximum input size of the medium-size HyenaDNA.

For each candidate gene, we obtained the reference gene sequence using biomaRt (Kinsella et al., 2011). Then for each study participant, we altered the reference alleles based on the position of the variants in the gene. The resulting DNA sequence was then fed into the medium-size HyenaDNA to get embeddings for each nucleotide. To construct a gene embedding, we extracted the nucleotide embeddings from positions of pathogenic variants, then we averaged them. All the gene embeddings were stored in a database to be used for the next steps. For loading pre-trained weights, we used the HuggingFace (Wolf et al., 2019) interface in Python (`https://www.python.org`). For model inference and embedding calculation, we used one Nvidia A100 (40GB) GPU.

2.5 Case-vs-control analysis

To assess the impact of pathogenic variants on the gene embeddings, we implemented a case-vs-control approach. For each gene, we trained a logistic regression (with L1 and L2 regularization) using the gene embeddings to classify patients from healthy controls. We used scikit-learn (Pedregosa et al., 2011) to train the model on 75% of the data and evaluate it with the remaining 25% resulting in a F_1 score for each gene. We compared the gene-specific F_1 scores and ranked genes based on this metric (Figure 1a).

For top candidate gene selection, we used Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) as an outlier detector. We applied DBSCAN on the calculated F_1 scores to find outliers and selected corresponding genes as top candidates.

Finally, to validate the results, we implemented a permutation test. We randomly shuffled the labels (case or control) for N=1000 times. Then we trained a logistic regression on 75% of the data and calculated a F_1 score on the other 25%. We counted the number of times that the random F_1 score was more than or equal than the observed F_1 score. We calculated a p-value as follow (with $\epsilon = 0.001$):

$$p = \frac{\text{count(random } F_1 \geq \text{Observed } F_1) + \epsilon}{N + \epsilon}$$

2.6 Case-only analysis

We also developed a case-only method to prioritize candidate genes if healthy controls are not available. In this approach, for each gene we divided the gene embeddings into mutant (if the patient carried a pathogenic variant) and non-mutant (if the patient was not a carrier). Then we calculated a distance score as the average Euclidean distance between mutant and non-mutant gene embeddings. We utilized these gene-specific distance scores to rank candidate genes (Figure 1a).

For top candidate gene selection, similar to the case-vs-control approach, we applied DBSCAN on the distance scores and selected outliers as the top candidate genes.

To validate the results, we implemented a statistical test as follow : For N=1000 times we generated random reference and alternative embeddings and calculated the distance score. We counted the number of times that the random distance score was more than or equal than the observed distance score. We calculated a p-value as follow (with $\epsilon = 0.001$):

$$p = \frac{\text{count(random distance score} \geq \text{Observed distance score)} + \epsilon}{N + \epsilon}$$

2.7 Graph neural network training

To understand the underlying mechanism of the disease, we designed an explainable approach based on graph neural network (GNN). A summary of the method can be found in figure 1b. First, we created a protein-protein interaction (PPI) network that indicates interactions between genes carrying pathogenic variants. We used the STRING database (Szklarczyk et al., 2023) and included interactions with confidence score ≥ 0.6.

Afterwards, we created individual-specific graphs, which include gene embeddings as node features. We trained a GNN to classify patients' graphs from controls. GNN architecture consists of two hidden graph convolution layers (Zhang et al., 2019) with 16 nodes for message passing and a global sort pooling (Zhang et al., 2018) for node feature aggregation. Pooling is essential because the model is trained for graph classification, therefore with pooling we can generate graph representations from node features. We used AdamW (Loshchilov and Hutter, 2019) optimizer with learning rate = 0.001 and weight decay = 0.001 for training. We used batch size = 32 and trained the model for 1000 epochs. We used PyTorch geometric (Fey and Lenssen, 2019) for implementing and training the GNN.

2.8 Subnetwork identification and pathway enrichment analysis

After training the GNN, we used GNNExplainer (Ying et al., 2019) to assign an explainability score to each node, showing how important they are for graph classification . We applied GNNExplainer for all the samples and averaged the explainability scores for each node across samples.

After obtaining the explainability scores, we used the Genetic Algorithm (GA) (Katoch et al., 2020) to identify the "best" subnetwork with maximum fitness, defined as the average of explainability scores of its nodes. GA is a bio-inspired algorithm that mimics evolution by implementing natural selection, chromosomal crossover, and mutation. Previous studies have successfully utilized GA for subnetwork identification (Ulgen et al., 2019; Wu et al., 2011). To summarize the GA, we start with a population of random subnetworks, then we select 50% of subnetworks with probabilities proportional to their fitness scores (roulette wheel selection). Afterwards, we create new subnetworks by mutating them (adding or removing edges) and crossovering them (connecting two subnetworks, if possible). We started with an initial population of 100 subnetworks and repeated the GA for 10 generations with a mutation rate of 0.5. At the end, we chose the "most fit" subnetwork at the last generation.

Finally, to gain biological insights into the selected subnetwork, we performed pathway enrichment analysis, a method for identifying biological functions that are over-represented in a group of genes (Chicco and Agapito, 2022). We used the GSEApy package (Fang et al., 2022), which uses Enrichr (Kuleshov et al., 2016) for over-representation analysis and the Reactome database (Milacic et al., 2023) as reference. We kept significantly enriched pathways with false discovery rate (FDR) ≤ 0.05.

3 Results

3.1 Study participants

As patient cohort (rare disease cases), we used exome data from 120 previously healthy children admitted to PICUs with respiratory failure due to a common viral respiratory infection. Their median age was 78 days, 50 (42%) were female, and 90 (78%) were of European ancestry. Respiratory Syncytial Virus (RSV) and Human Rhinovirus (HRV) were the most common detected pathogens, in 67 (56%) and 31 (26%) of the cases, respectively. As controls, we selected 172 healthy individuals from our in-house database of exome-sequenced individuals, representing a random subset of the general population. Since the phenotype we are studying is rare, we assume that the controls are not enriched in individuals with genetic risk factors for infectious disease susceptibility.

3.2 Variant classification

In the patient group, 55,300 variants were mapped to coding and splicing regions and were scored with the ACMG/AMP Bayesian classification framework. 48,875 variants had a PoP ≤ 0.1 and were considered benign. 5,838 variants had an intermediate PoP (between 0.1 and 0.9), resulting in their classification as variants of unknown significance (VUS). 587 variants (in 508 genes) exceeded the pathogenicity threshold (≥ 0.9) and were considered as damaging.

3.3 Gene prioritization

A total of 498 (98%) candidate genes passed the selection criteria (Methods, Gene embedding cal-

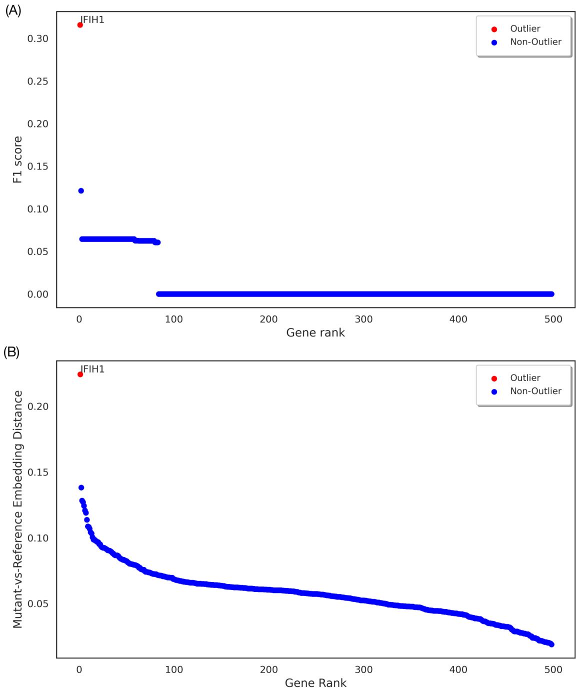


Figure 2: Gene prioritization results. (A) Genes ranked according to their corresponding F_1 score calculated based on case-vs-control workflow. (B) Gene ranking based on average distance of mutants and non-mutant embeddings, computed according to the case-only workflow.

culation). For each candidate gene, we calculated gene embeddings using the pre-trained HyenaDNA for all 292 study participants (120 cases and 172 controls), resulting in gene-specific embeddings in the embedding space. We then ranked candidate genes using two approaches:

1) Case-vs-control: We trained a logistic regression for each gene and calculated a gene-specific F_1 scores. We used these scores to rank the genes and find top candidates by applying DBSCAN for outlier detection. The top candidate gene with the highest F_1 score was *IFIH1* (Figure 2.A). We performed a permutation test which resulted in p-value=0.009 (Supplementary figure S1).

2) Case-only: in this scenario we used the gene-specific distance score (calculated based on the average Euclidean distance of mutant and non-mutant embeddings) for gene prioritization and top-candidate selection. *IFIH1* ranked first and was selected as an outlier using the DBSCAN method (Figure 2.B) and was significantly different from the expected distribution (p-value=10^{-6}, Supplementary figure S2).

3.4 PPI construction and graph neural network training

We constructed a high-quality PPI based on the interactions between the protein products of all candidate genes, resulting in a PPI with 138 nodes and 176 edges. For each participant, we initialized the same PPI structure, but used their personalized gene-embeddings as node features, resulting in 292 (120 cases and 172 controls) unique graphs. We used these graphs to train a GNN for classifying cases from controls. GNN structure consisted of 2 graph convolution layers with 16 nodes, and global sort pooling to generate graph representations from node features. We trained the GNN for 1000 epochs.

3.5 Subnetwork identification and pathway enrichment analysis

After training the GNN, we used GNNExplainer to assign an explainability score to each node, showing how important they are for graph classification. We applied GNNExplainer for all the samples and averaged the explainability scores for each node across samples. Figure 3.A shows the PPI with explainability scores reflected on the edges' widths. After obtaining the explainability scores, we used the Genetic Algorithm to identify the "best" subnetwork with maximum fitness. The fitness of a subnetwork was defined as the average of explainability scores of its nodes . This resulted in a subnetwork with 10 genes including *IFIH1*, *OAS1*, *OAS3*, *MX1*, *IFNAR1*, *IL10RB*, *ZNFX1*, *NLRC5*, *TRIM40*, and *ABCE1* (Figure 3.B). Finally, we performed pathway enrichment analysis using the Reactome database as reference and kept significantly enriched pathways with FDR ≤ 0.05. Top 10 resulting pathways are shown in figure 4.

4 Discussion

In this study we aim to harness the potential of DNA foundation models to translate the intricate 'language' of DNA into meaningful and actionable information. We propose a framework to utilize DNA-LMs for gene prioritization and pathway identification in rare disease studies. Based on the hypothesis that variants with strong deleterious effects alter the gene embeddings significantly in the embedding space, we demonstrate that it is possible to prioritize disease-associated genes/pathways in a cohort of 120 children requiring intensive care support because of a severe illness caused by a

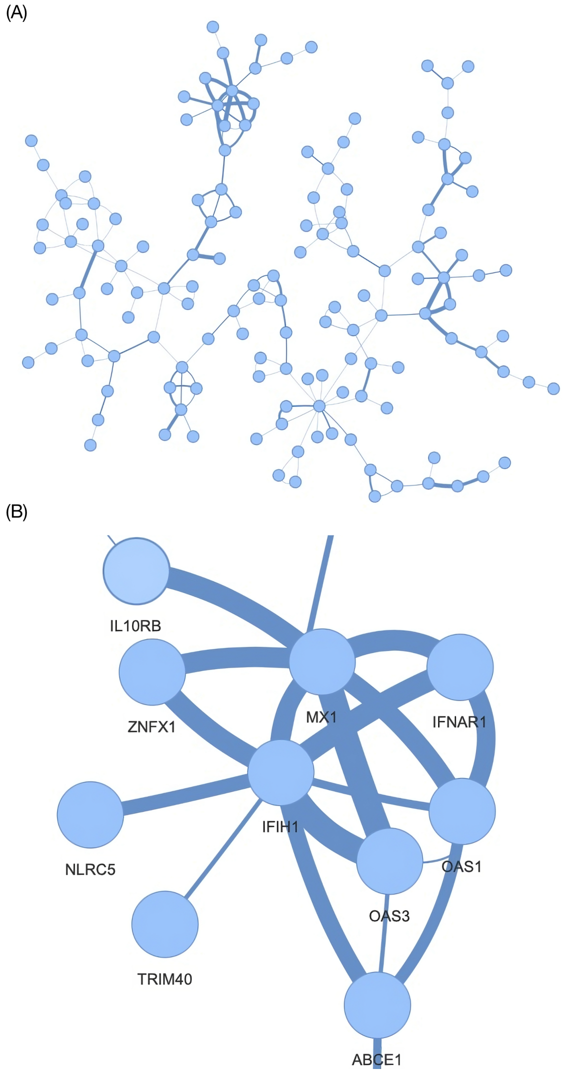


Figure 3: Subnetwork identification results. (A) PPI of candidate genes scored using GNNExplainer. The thickness of edges reflects the importance of nodes connected to it. (B) Selected subnetwork with maximum fitness, defined as the average nodes' importance scores. This subnetwork is identified via the Genetic Algorithm.

respiratory virus.

For gene prioritization, we propose two approaches to analyze the gene embeddings (Figure 1a): case-vs-control and case-only. The case-only approach is particularly promising for rare disease research, where finding a well-matched control group is often challenging. The ability of the method to differentiate between mutant and non-mutant gene embeddings within the same patient cohort is a novel and practical solution to this long-standing issue. By applying the gene prioritization workflow, we successfully re-identified *IFIH1* - which encodes an RIG-I-like receptor involved in the sensing of viral RNA (Rehwinkel and Gack, 2020) - as the top candidate gene in our patient cohort.

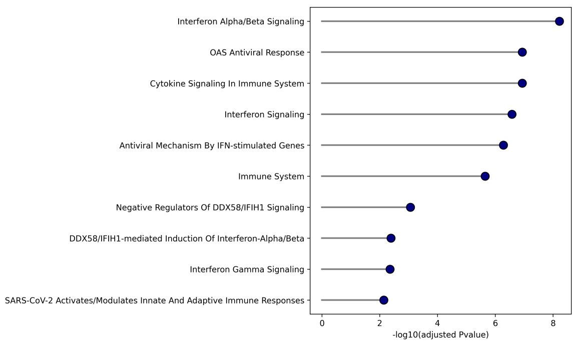


Figure 4: Top 10 significantly enriched pathways using the Reactome database. Genes in the selected subnetwork were used as input.

For pathway identification, we propose an integrative method, combining DNA-LM with interpretable GNN and Genetic Algorithm (Figure 1b). This approach takes into account various information such as PPI, number of variant carriers, and context-specific impact of variants on gene sequences. By applying this method, we were able to identify potentially relevant genes (*IFIH1*, *OAS1*, *OAS3*, *MX1*, *IFNAR1*, *IL10RB*, *ZNFX1*, *NLRC5*, *TRIM40*, and *ABCE1*) that can explain the disease pathogenesis.

All the identified genes are coding for molecules that play an important role in antiviral defense. *IFIH1* encodes MDA5, which is a cytoplasmic viral RNA sensor that recognizes single- or double-strand RNA to launch a type 1 interferon response (Rehwinkel and Gack, 2020). *OAS1* and *OAS3* encode enzymes that activate host RNase L to degrade viral RNA (Hornung et al., 2014). *ABCE1* encodes a protein that is involved in the regulation of OAS/RNase L pathway (Martinand et al., 1998). *MX1* encodes a guanosine-triphosphate-metabolizing protein that antagonizes the replication process of viruses (Haller and Kochs, 2019). *IFNAR1* and *IL10RB* encode cytokine receptors that mediate the antiviral immunity (Zanin et al., 2021; Moore et al., 2001). *ZNFX1* encodes a protein that binds to viral RNA and interacts with mitochondrial antiviral signaling (MAVS) protein, promoting the expression of interferon-stimulated genes (Vavassori et al., 2021). *NLRC5* and *TRIM40* encode regulators of antiviral signaling pathways (Kuenzel et al., 2010; Zhao et al., 2017). Deficiencies in some of these genes have been previously studied and shown to impair immunity against specific human viruses (Lamborn et al., 2017; Asgari et al., 2017; Chen et al., 2021; Abolhassani et al., 2022; Korol et al., 2023; Saadat et al., 2023; Lee et al., 2023).

In this study we focused on DNA-LMs, although protein language models (pLMs) such as ESM-1b (Brandes et al., 2023) have demonstrated state-of-the-art performance in scoring missense variants. The reason we used a DNA-LM instead of pLM is that DNA-LMs can model various variant types (e.g., splicing, stop-gained, etc.) while pLMs focus only on missense variants. Moreover, by using DNA-LMs, our method can be extended to other variant types such as those mapping to introns, branchpoint motives, or untranslated regions (UTRs).

While our method shows promise, there are inherent challenges and limitations. Our proposed workflow identifies genes with significant changes in their embeddings, yet a careful analysis is required to quantify the minimum embedding distortion to be detectable by the model. Moreover, the interpretation of gene embeddings requires careful consideration, since not all genetic variations captured in the embeddings might be clinically relevant.

The potential for integrating DNA-LMs with other techniques, such as multi-omics, could further enhance our understanding of genetic diseases. This has significant implications for the identification of disease-causing genes/pathways, potentially leading to more targeted and effective treatments in personalized medicine. The demonstration that DNA-LMs can accurately identify genes and pathways involved in rare diseases paves the way for further research and application of artificial intelligence in various genomics research domains.

Code Availability

The code for this study is available here.

References

Hassan Abolhassani, Nils Landegren, Paul Bastard, Marie Materna, Mohammadreza Modaresi, Likun Du, Maribel Aranda-Guillén, Fabian Sardh, Fanglei Zuo, Peng Zhang, Harold Marcotte, Nico Marr, Taushif Khan, Manar Ata, Fatima Al-Ali, Remi Pescarmona, Alexandre Belot, Vivien Béziat, Qian Zhang, Jean-Laurent Casanova, Olle Kämpe, Shen-Ying Zhang, Lennart Hammarström, and Qiang Pan-Hammarström. 2022. Inherited IFNAR1 deficiency in a child with both critical COVID-19 pneumonia and multisystem inflammatory syndrome. *J. Clin. Immunol.*, 42(3):471–483.

Joanna S Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. 2018. Omim.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, 47(D1):D1038–D1043.

Samira Asgari, Luregn J. Schlapbach, Stéphanie Anchisi, Christian Hammer, Istvan Bartha, Thomas Junier, Geneviève Mottet-Osman, Klara M. Posfay-Barbe, David Longchamp, Martin Stocker, Samuel Cordey, Laurent Kaiser, Thomas Riedel, Tony Kenna, Deborah Long, Andreas Schibler, Amalio Telenti, Caroline Tapparel, Paul J. McLaren, Dominique Garcin, and Jacques Fellay. 2017. Severe viral respiratory infections in children with ifih1 loss-of-function mutations. *Proceedings of the National Academy of Sciences*, 114(31):8342–8347.

Paul L Auer and Guillaume Lettre. 2015. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1):16.

Adam Auton, 1000 Genomes Project Consortium, Shane McCarthy, Gil A. McVean, and Goncalo R. Abecasis. 2015. A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song. 2023. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44).

Nadav Brandes, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye, and Vasilis Ntranos. 2023. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522.

Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890.

Siwei Chen, Laurent C. Francioli, Genome Aggregation Database (gnomAD) Consortium, Grace Tiao, Benjamin M. Neale, Daniel G. MacArthur, and Konrad J. Karczewski. 2023. A genomic mutational constraint map using variation in 76, 156 human genomes. *Nature*, 625(7993):92–100.

Yongkun Chen, Laura Graf, Tao Chen, Qijun Liao, Tian Bai, Philipp P Petric, Wenfei Zhu, Lei Yang, Jie Dong, Jian Lu, Ying Chen, Juan Shen, Otto Haller, Peter Staeheli, Georg Kochs, Dayan Wang, Martin Schwemmle, and Yuelong Shu. 2021. Rare variant MX1 alleles increase human susceptibility to zoonotic H7N9 influenza virus. *Science*, 373(6557):918–922.

Davide Chicco and Giuseppe Agapito. 2022. Nine quick tips for pathway enrichment analysis. *PLOS Computational Biology*, 18(8):e1010348.

Micaela E. Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. 2023. To transformers and beyond: Large language models for the genome. *arXiv preprint*.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics.

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. 2011. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*, 43(5):491–498.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. 2022. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1).

Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with pytorch geometric. *Preprint*, arXiv:1903.02428.

Michael D. Gallagher and Alice S. Chen-Plotkin. 2018. The post-gwas era: From association to function. *The American Journal of Human Genetics*, 102(5):717–730.

Jesse Gillis and Paul Pavlidis. 2012. "guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444.

Dimitri Guala and Erik L. L. Sonnhammer. 2017. A large-scale benchmark of gene prioritization methods. *Scientific Reports*, 7(1).

Otto Haller and Georg Kochs. 2019. Mx genes: host determinants controlling influenza virus infection and trans-species transmission. *Human Genetics*, 139(6–7):695–705.

Veit Hornung, Rune Hartmann, Andrea Ablasser, and Karl-Peter Hopfner. 2014. Oas proteins and cgas: unifying concepts in sensing and responding to cytosolic nucleic acids. *Nature Reviews Immunology*, 14(8):521–528.

Konrad J. Karczewski, FGenome Aggregation Database (gnomAD) Consortium, and Daniel G. MacArthur. 2020. The mutational constraint spectrum quantified from variation in 141, 456 humans. *Nature*, 581(7809):434–443.

Sourabh Katoch, Sumit Singh Chauhan, and Vijay Kumar. 2020. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 80(5):8091–8126.

R. J. Kinsella, A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek. 2011. Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database*, 2011(0):bar030–bar030.

Cecilia B Korol, Serkan Belkaya, Fahad Alsohime, Lazaro Lorenzo, Stéphanie Boisson-Dupuis, Joseph Brancale, Anna-Lena Neehus, Silvia Vilarinho, Alsum Zobaida, Rabih Halwani, Saleh Al-Muhsen, Jean-Laurent Casanova, and Emmanuelle Jouanguy. 2023. Fulminant viral hepatitis in two siblings with inherited IL-10RB deficiency. *J. Clin. Immunol.*, 43(2):406–420.

Sven Kuenzel, Andreas Till, Michael Winkler, Robert Häsler, Simone Lipinski, Sascha Jung, Joachim Grötzinger, Helmut Fickenscher, Stefan Schreiber, and Philip Rosenstiel. 2010. The nucleotide-binding oligomerization domain-like receptor nlrc5 is involved in ifn-dependent antiviral immune responses. *The Journal of Immunology*, 184(4):1990–2000.

Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97.

Ian T. Lamborn, Huie Jing, Yu Zhang, Scott B. Drutman, Jordan K. Abbott, Shirin Munir, Sangeeta Bade, Heardley M. Murdock, Celia P. Santos, Linda G. Brock, Evan Masutani, Emmanuel Y. Fordjour, Joshua J. McElwee, Jason D. Hughes, Dave P. Nichols, Aziz Belkadi, Andrew J. Oler, Corinne S. Happel, Helen F. Matthews, Laurent Abel, Peter L. Collins, Kanta Subbarao, Erwin W. Gelfand, Michael J. Ciancanelli, Jean-Laurent Casanova, and Helen C. Su. 2017. Recurrent rhinovirus infections in a child with inherited mda5 deficiency. *Journal of Experimental Medicine*, 214(7):1949–1972.

Danyel Lee, Jérémie Le Pen, Ahmad Yatim, Beihua Dong, Yann Aquino, Masato Ogishi, Rémi

Pescarmona, Estelle Talouarn, Darawan Rinchai, Peng Zhang, Magali Perret, Zhiyong Liu, Iolanda Jordan, Sefika Elmas Bozdemir, Gulsum Iclal Bayhan, Camille Beaufils, Lucy Bizien, Aurelie Bisiaux, Weite Lei, Milena Hasan, Jie Chen, Christina Gaughan, Abhishek Asthana, Valentina Libri, Joseph M Luna, Fabrice Jaffré, H-Heinrich Hoffmann, Eleftherios Michailidis, Marion Moreews, Yoann Seeleuthner, Kaya Bilguvar, Shrikant Mane, Carlos Flores, Yu Zhang, Andrés A Arias, Rasheed Bailey, Agatha Schlüter, Baptiste Milisavljevic, Benedetta Bigio, Tom Le Voyer, Marie Materna, Adrian Gervais, Marcela Moncada-Velez, Francesca Pala, Tomi Lazarov, Romain Levy, Anna-Lena Neehus, Jérémie Rosain, Jessica Peel, Yi-Hao Chan, Marie-Paule Morin, Rosa Maria Pino-Ramirez, Serkan Belkaya, Lazaro Lorenzo, Jordi Anton, Selket Delafontaine, Julie Toubiana, Fanny Bajolle, Victoria Fumadó, Marta L DeDiego, Nadhira Fidouh, Flore Rozenberg, Jordi Pérez-Tur, Shuibing Chen, Todd Evans, Frédéric Geissmann, Pierre Lebon, Susan R Weiss, Damien Bonnet, Xavier Duval, CoV-Contact Cohort§, COVID Human Genetic Effort¶, Qiang Pan-Hammarström, Anna M Planas, Isabelle Meyts, Filomeen Haerynck, Aurora Pujol, Vanessa Sancho-Shimizu, Clifford L Dalgard, Jacinta Bustamante, Anne Puel, Stéphanie Boisson-Dupuis, Bertrand Boisson, Tom Maniatis, Qian Zhang, Paul Bastard, Luigi Notarangelo, Vivien Béziat, Rebeca Perez de Diego, Carlos Rodriguez-Gallego, Helen C Su, Richard P Lifton, Emmanuelle Jouanguy, Aurélie Cobat, Laia Alsina, Sevgi Keles, Elie Haddad, Laurent Abel, Alexandre Belot, Lluis Quintana-Murci, Charles M Rice, Robert H Silverman, Shen-Ying Zhang, and Jean-Laurent Casanova. 2023. Inborn errors of OAS-RNase L in SARS-CoV-2-related multisystem inflammatory syndrome in children. *Science*, 379(6632):eabo3627.

Insuk Lee, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. 2011. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121.

Heng Li and Richard Durbin. 2009. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. 2023. Bend: Benchmarking dna language models on biologically meaningful tasks. *arXiv preprint*.

Camille Martinand, Tamim Salehzada, Michelle Silhol, Bernard Lebleu, and Catherine Bisbal. 1998. Rnase l inhibitor (rli) antisense constructions block partially the down regulation of the 2-5a/rnase l pathway in encephalomyocarditis-virus-(emcv)-infected cells. *European Journal of Biochemistry*, 254(2):248–255.

William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. The ensembl variant effect predictor. *Genome Biology*, 17(1).

Michael L. Metzker. 2009. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. 2023. The reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678.

Kevin W. Moore, Rene de Waal Malefyt, Robert L. Coffman, and Anne O'Garra. 2001. Interleukin-10 and the interleukin-10 receptor. *Annual Review of Immunology*, 19(1):683–765.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jan Rehwinkel and Michaela U. Gack. 2020. Rig-i-like receptors: their regulation and roles in rna sensing. *Nature Reviews Immunology*, 20(9):537–551.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl

Voelkerding, and Heidi L. Rehm. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–424.

Ali Saadat and Jacques Fellay. 2024. Fine-tuning the ESM2 protein language model to understand the functional impact of missense variants. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*.

Ali Saadat, Jérôme Gouttenoire, Paolo Ripellino, David Semela, Soraya Amar, Beat M. Frey, Stefano Fontana, Elise Mdawar-Bailly, Darius Moradpour, Jacques Fellay, and Montserrat Fraga. 2023. Inborn errors of type i interferon immunity in patients with symptomatic acute hepatitis e. *Hepatology*, 79(6):1421–1431.

Eleanor G Seaby and Sarah Ennis. 2020. Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Briefings in Functional Genomics*, 19(4):243–258.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. 2023. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.*, 51(D1):D638–D646.

Sean V. Tavtigian, Marc S. Greenblatt, Steven M. Harrison, Robert L. Nussbaum, Snehit A. Prabhu, Kenneth M. Boucher, and Leslie G. Biesecker. 2018. Modeling the acmg/amp variant classification guidelines as a bayesian classification framework. *Genetics in Medicine*, 20(9):1054–1060.

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1).

Ege Ulgen, Ozan Ozisik, and Osman Ugur Sezerman. 2019. pathfindr: An r package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Frontiers in Genetics*, 10.

Stefano Vavassori, Janet Chou, Laura Eva Faletti, Veronika Haunerdinger, Lennart Opitz, Pascal Joset, Christopher J. Fraser, Seraina Prader, Xianfei Gao, Luise A. Schuch, Matias Wagner, Julia Hoefele, Maria Elena Maccari, Ying Zhu, George Elakis, Michael T. Gabbett, Maria Forstner, Heymut Omran, Thomas Kaiser, Christina Kessler, Heike Olbrich, Patrick Frosk, Abduarahman Almutairi, Craig D. Platt, Megan Elkins, Sabrina Weeks, Tamar Rubin, Raquel Planas, Tommaso Marchetti, Danil Koovely, Verena Klämbt, Neveen A. Soliman, Sandra von Hardenberg, Christian Klemann, Ulrich Baumann, Dominic Lenz, Andreas Klein-Franke, Martin Schwemmle, Michael Huber, Ekkehard Sturm, Steffen Hartleif, Karsten Häffner, Charlotte Gimpel, Barbara Brotschi, Guido Laube, Tayfun Güngör, Michael F. Buckley, Raimund Kottke, Christian Staufner, Friedhelm Hildebrandt, Simone Reu-Hofer, Solange Moll, Achim Weber, Hundeep Kaur, Stephan Ehl, Sebastian Hiller, Raif Geha, Tony Roscioli, Matthias Griese, and Jana Pachlopnik Schmid. 2021. Multisystem inflammation and susceptibility to viral infections in human znfx1 deficiency. *Journal of Allergy and Clinical Immunology*, 148(2):381–393.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint*.

Jiaxin Wu, Mingxin Gan, and Rui Jiang. 2011. A genetic algorithm for optimizing subnetwork markers for the study of breast cancer metastasis. In *2011 Seventh International Conference on Natural Computation*. IEEE.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Preprint*, arXiv:1903.03894.

Natacha Zanin, Christine Viaris de Lesegno, Christophe Lamaze, and Cedric M. Blouin. 2021. Interferon receptor trafficking and signaling: Journey to the cross roads. *Frontiers in Immunology*, 11.

Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. 2018. An end-to-end deep learning architecture for graph classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1).

Chunyuan Zhao, Mutian Jia, Hui Song, Zhongxia Yu, Wenwen Wang, Qi Li, Lining Zhang, Wei Zhao, and Xuetao Cao. 2017. The e3 ubiquitin ligase trim40 attenuates antiviral immune responses by targeting mda5 and rig-i. *Cell Reports*, 21(6):1613–1623.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen

Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint*.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint*.

A Appendix

We used ACMG/AMP guidelines (Richards et al., 2015) to classify the variant into putative pathogenicity groups, as described in our previous works (Saadat et al., 2023; Saadat and Fellay, 2024). In summary, we gather all the available evidences for a variant. Table 1 summarizes all the ACMG/AMP criteria that we used.

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population and controls	MAF is high in the population (BS1)		MAF is rare in the population and absent from the controls (PM2_supporting)			
Computational		MIssense in a gene where mostly truncating variants cause disease (BP1) In-frame indel in a repeat region (BP3) Computational evidence suggest no impact on gene product (BP4)	Computational evidence supports a deleterious effect on the gene product (PP3)	Protein length changing variant (PM4) Novel missense change at an amino acid residue Where a different pathogenic missense change has been seen before (PM5)	Same amino acid change as an established pathogenic variant (PS1) Low confidence null variant (PVS1_strong)	High confidence null variant (PVS1)
Functional	Well-established functional studies show no deleterious effect (BS3)		Missense in a gene with low rate of benign missense variants and missense variants are common mechanism of disease (PP2)	Mutational hotspot or well-studied functional domain without benign variation (PM1)	Well-established functional studies show a deleterious effect (PS3)	

Table 1: the summary of ACMG/AMP criteria used for variant classification. MAF: minor allele frequency

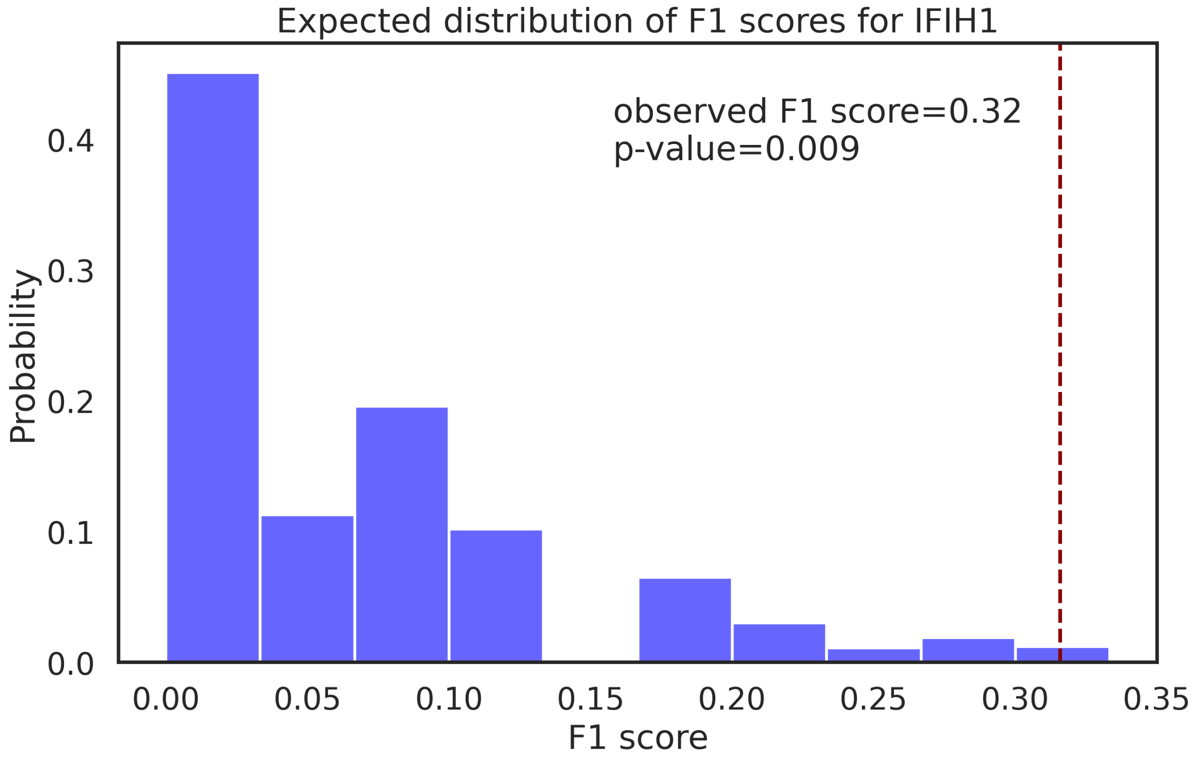
To calculate the probability of pathogenicity (PoP), we use the Bayesian framework developed by Tavtigian et al. (2018). For a given variant, the PoP is calculated as follow:

$$P_x = \text{number of pathogenic criteria applied at the level of } x$$
$$x \in \{\text{Very strong}, \text{Strong}, \text{Moderate}, \text{Supporting}\}$$

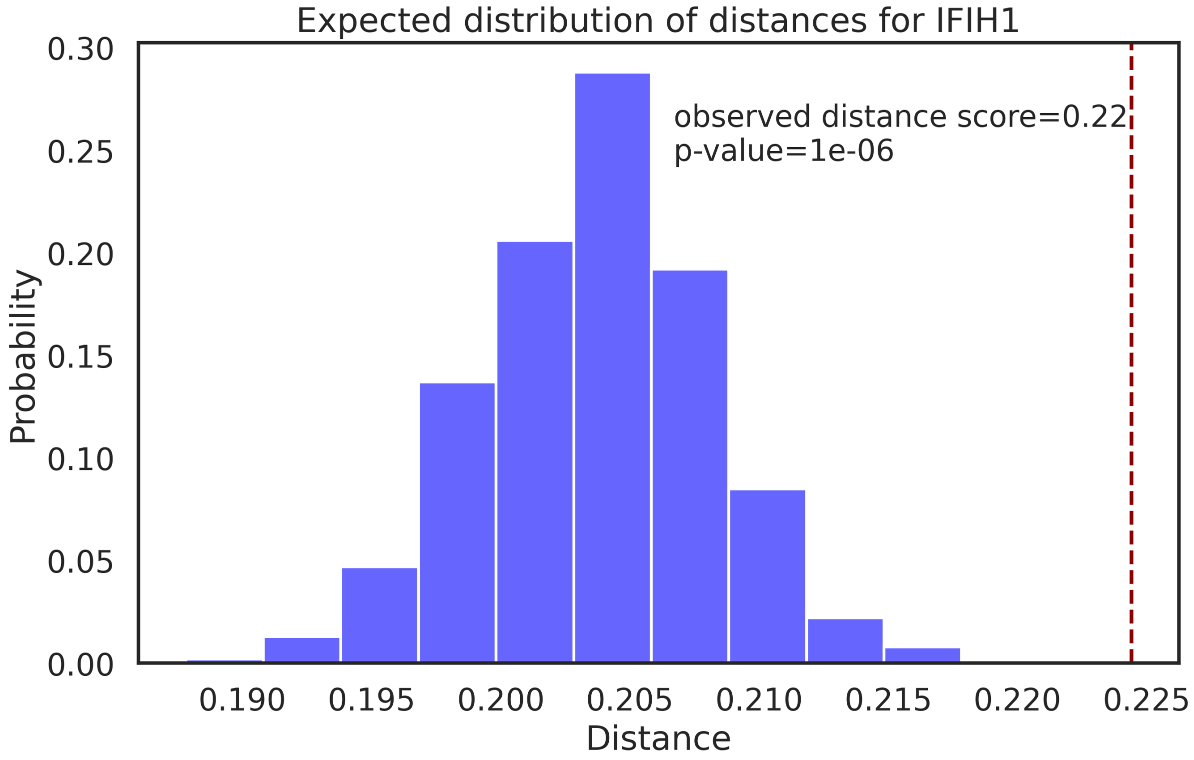
$$B_y = \text{number of benign criteria applied at the level of } y$$
$$y \in \{\text{Strong}, \text{Supporting}\}$$

$$\text{odds of pathogenicity (OP)} = 350^{\left(\frac{P_{\text{Very strong}}}{1} + \frac{P_{\text{Strong}}}{2} + \frac{P_{\text{Moderate}}}{4} + \frac{P_{\text{Supporting}}}{8} - \frac{B_{\text{Strong}}}{2} - \frac{B_{\text{Supporting}}}{8}\right)}$$

$$\text{probability of pathogenicity (PoP)} = \frac{OP \times 0.1}{((OP - 1) \times 0.1 + 1)}$$



Supplementary Figure S1: Permutation test results for case-vs-control approach. Expected distribution of F_1 scores for *IFIH1* is shown in blue. The red line indicates the observed F_1 score.



Supplementary Figure S2: Statistical test results for the case-only approach. Expected distribution of distance scores for *IFIH1* is shown in blue. The red line indicates the observed distance score.