IJEMD-CSAI, 4 (1) (2025) https://doi.org/ 10.54938/ijemdcsai.2025.04.1.500

International Journal of Emerging Multidisciplinaries:
Computer Science and Artificial Intelligence

Research Paper
Journal Homepage: www.ojs.ijemd.com
ISSN (print): 2791-0164 ISSN (online): 2957-5036

A Genetic Algorithm-Driven Personalized Genome Mutation Pathway Predictor for Early Diagnosis of Rare Polygenic Disorders

Akinrotimi Akinyemi Omololu [1*], Atoyebi Jelili Olaniyi [2], Owolabi Olugbenga Olayinka [3] and Omotosho Israel Oluwabusayo [4]

1. Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.
2. Department of Computer Engineering, Adeleke University, Ede, Osun State, Nigeria.
3. Department of Electrical and Electronics Engineering, Adeleke University, Ede, Osun State, Nigeria.
4. Department of Management Information Systems, Bowie State University, Maryland, USA.

Abstract

Accurate prediction of rare polygenic disorders remains a significant challenge in precision medicine, primarily due to the fact that they involve a complicated genetic architecture and current computational models are restricted. Traditional polygenic risk scores (PRS) have additive assumptions and finite cross-population validity and hence are not appropriate for rare disorders. In this study, a novel GA-based approach is presented that models individualized forward mutational routes, enabling early identification of risk genomic configurations. Each GA chromosome represents a binary vector of rare variants from whole-genome sequencing data, and evolutionary processes are guided by a composite fitness function. The function integrates pathogenicity scores, disease associations, and population rarity to yield biologically relevant simulations. Using 1000 Genomes Project data, we simulate 500 mutational trajectories in 500 different individuals. Results determine an average 27.2% increase in pathogenicity and 38.4% increase in harmful variants, with more than 60% convergence to known disease profiles in European and South Asian genomes. Approximately 24% of simulated genomes per individual exceed high-risk thresholds, outperforming PRS in identifying non-additive and epistatic effects. This GA strategy offers a dynamic, ancestry-aware approach to predicting rare disease risk, broadening the scope of predictive genomics and enabling earlier, more specific clinical interventions.

Keywords: Epistasis modeling, Evolutionary computation, Genetic algorithm, Genomic simulation, Personalized medicine, Polygenic risk prediction.

Email Address : akinrotimiakinyemi@ieee.org (Akinrotimi Akinyemi Omololu *)

1. INTRODUCTION

Recent advances in whole-genome sequencing have precipitated an astonishing revolution in genomic medicine, with prospects for tailor-made disease prediction and early intervention. While great advances have been made in the diagnosis of monogenic disorders, the majority of important health disorders are polygenic disorders because interactions among many genetic variants across the genome play a role. This is particularly true for rare polygenic disorders, such as certain early-onset autoimmune diseases, neurodegenerative syndromes, and psychiatric disorders, which may be difficult to diagnose since the contributing genetic factors are subtle and complex in nature.

One of the most widely used computational approaches being utilized in the estimation of the polygenic risk is the Polygenic Risk Score (PRS). PRS uses a combination of the impact of numerous single nucleotide polymorphisms (SNPs), generally from GWAS data, and estimates an individual's chance of developing a wide range of diseases. Even though PRS has been very valuable in determining risk levels for different populations, it also has some severe limitations, more so when it comes to rare and multifactorial diseases. First of all, many of PRS models assume additive genetic risk factors combine additively, i.e., they do not account for epistatic interactions-the complicated genetic relationships that contribute a major part in defining disease risk [1]. Second, PRS merely gives a snapshot of one's current genetic risk and fails to consider how his or her genome would change in the future as new mutations arise. Lastly, these models are skewed towards data from European populations and therefore may not be effective in more diverse or less common groups of genetic variants [2],[3]. These constraints serve to emphasize a significant need for a more dynamic, person-specific model that can simulate the way a single individual's genome may vary with potential future mutation, conceivably to a disease state. Such models, especially for uncommon polygenic diseases, may offer the potential to allow diagnosis and treatment earlier, with important clinical benefits. In order to satisfy this need, we introduce a Genetic Algorithm (GA)-based framework that is capable of simulating potential sequences of mutations in individual genomes. GAs are effective search heuristics motivated by natural evolution particularly fit to navigate big, nonlinear, and complex spaces like human genomic variation. For our approach, each chromosome in the GA is an encoding of a string of potential mutations along recognized disease-causing genes. Through iteratively choosing, recombining via crossover, and mutating under the direction of a biologically-motivated fitness function, the GA evolves candidate trajectories that converge to plausible paths to disease expression. The fitness function used in this study includes multiple biological axes to promote realistic mutation pathways. First, it considers the functional impact of genetic variants, such as predicted deleteriosity based on established scoring systems. Second, it incorporates gene–disease association robustness, utilizing manually curated databases like DisGeNET to ensure that simulated mutations are consistent with known disease-causing loci. Third, it includes allele population frequency, with less frequent variants being given greater weight since they are more likely to contribute disproportionately to disease risk.

The main contributions of this study are: (a) A novel GA-based model that predicts future genome mutation trajectories for rare polygenic disorders. (b) A multi-dimensional fitness function combining functional and epidemiological features. (c) Validation using anonymized personal genome data from the

1000 Genomes Project, demonstrating the approach's ability to simulate biologically meaningful mutation paths. (d) A proof-of-concept showing how dynamic mutation modeling can complement and enhance static PRS, particularly for rare disease risk prediction. Our work aligns with recent efforts to improve diagnostic yield in rare diseases. For example, the Deciphering Developmental Disorders (DDD) study successfully diagnosed numerous children via genome-wide sequencing in 2023 [4],[8],[5]. However, unlike these studies which focus on current variant profiles the method used in this study simulates future mutational evolution, offering a novel, forward looking dimension to personalized risk modeling. The remainder of this paper is organized as follows. Section 2 reviews related work in polygenic risk models and GA applications in genomics. Section 3 describes our proposed methodology, including data preprocessing, chromosome encoding, and fitness function design. Section 4 presents experimental results and validation. Section 5 discusses implications, limitations, and future directions, and Section 6 concludes the paper.

2. LITERATURE REVIEW

The field of genomic diagnosis has experienced incredible growth in the recent past due to technological advancements in high-throughput sequencing and the use of computational methods in clinical decision making. The most significant among these computational tools are the Polygenic Risk Scores (PRS), which aim to quantify an individual's risk to complex diseases by summing up the effects of numerous common genetic variations. Despite widespread application in the risk modeling of diseases such as cardiovascular disease and type 2 diabetes, PRS methods have also shown accuracy and applicability limitations, particularly for rare polygenic disorders [2], [7].

2.1 Limitations of Polygenic Risk Scores

One of the key frailties of traditional PRS models is population bias. Most of the models have been trained on genome-wide association study (GWAS) data from individuals of European ancestry. This lim-its their transferability and usefulness across genetically diverse [6], [8]. Also, PRS models assume additive effects across SNPs, circumventing intricate epistasis interactions that may contribute importantly to disease risk, especially in undiagnosed or rare disease. A second challenge is their static nature. Most traditional PRS models examine current-genomic data and ignore the temporal dynamics of genetic variation, such as de novo mutations and somatic changes that can happen within a lifetime. For rare diseases (where distinctive, multi-locus combinations may arise de novo or through less common mutational events) this static examination can lead to missed diagnoses or false negatives [9]. Although machine learning (ML) and deep learning (DL) approaches have been suggested to enhance PRS models e.g., convolutional neural networks learning nonlinear SNP interactions (Chen et al., 2024), they too are limited by their dependence on known variant profiles and their inability to model future genomic pathways.

2.2 Genetic Algorithms in Biomedical Research

Genetic Algorithms (GAs) are heuristics for optimization, inspired by evolutionary biology, which employ mechanisms such as selection, mutation, and crossover to iteratively improve candidate solutions. In their very nature, they are well adapted to high-dimensional, non-linear search spaces, like those typical of a vast number of genomic problems. GAs has found application in biomedical informatics for feature selection in gene expression, disease classification, and biomarker identification [9].

[10] for instance, presented an approach called FCS-Net, a Feature Co-selection Network that uses GAs to select systematically heterogeneous sets of genetic variables. These sets facilitate modeling gene–gene interaction and genetic heterogeneity in colorectal cancer GWAS data. Genetic Algorithms in Biomedical Research has over time had a steady build. [11] employed a GA to select informative biomarkers from high-dimensional proteomic profiles. The GA was coupled with a model of classification, and this improved the prediction performance by 12% compared with standard feature selection methods in discriminating among patient subgroups in various states of disease. [12] created TPOT-MDR, utilizing Genetic Programming (a collection of evolutionary algorithms derived from GAs) to build optimal analysis pipelines with Multifactor Dimensionality Reduction (MDR). TPOT-MDR detected higher-order interactions in simulated and real-world complex disease data better than conventional methods. [13] created a GA-optimized artificial neural network (ANN) for MRI feature analysis in Alzheimer's patients. Their GA-based approach resulted in a 96% classification accuracy as opposed to manually tuned networks, with specific success in differentiating between early-stage Alzheimer's and age-matched controls. [14] engineered a hybrid method based on a mix of genetic algorithms with support vector machines (GA–SVM) for the identification of a smaller set of SNPs linked with chemotherapy response in breast cancer. The pipeline improved interpretability and reduced false positives with 15% improvement in F1score over baseline classifiers. [15] used GAs to parameterize Boolean network models of gene regulation in melanoma treatment. It resulted in networks that closely matched patient response patterns seen in real-world data, and increased prediction accuracy from 72% to 88% for sensitivity to immunotherapy on multiple datasets.

2.3 Rare Polygenic Disorders and Modeling Challenges

Rare diseases affect over 300 million people globally, and the majority has a genetic basis. Polygenic rare diseases, however, are very hard to detect because of their multifactorial etiology, often involving rare combinations of common variants, low-frequency mutations, and complex epistatic interactions [19]. Classical GWAS approaches lack the resolution to capture such configurations, especially when individual-level data is sparse or underrepresented. Studies like the Deciphering Developmental Disorders (DDD) study have made significant progress in early-onset rare disorder diagnosis through genome-wide sequencing [5],[9]. Yet, even with these studies, the focus remains on variants that exist in the present, rather than tracing the evolutionary trajectory of a genome through time. And this is where there is a fundamental gap in predictive modeling, particularly for individuals who are yet to display symptoms but are genetically predisposed. Furthermore, most variant pathogenicity prediction tools (e.g., CADD, SIFT,

PolyPhen-2) assess the likelihood of disease given that a mutation already exists, as opposed to the likelihood of future mutation patterns that will result in pathology.

2.4 Related Work

Recent advances in federated learning [23] have improved multi-population risk modeling but remain constrained by static genomic snapshots. Similarly, DNA language models [24] show promise in rare disease gene ranking yet lack temporal simulation capabilities. Spatial transcriptomics approaches [25] capture micro environmental dynamics but ignore germline mutational trajectories. While nano pore sequencing [19] enhances variant detection, it cannot prospectively model pathogenic pathways. Even CRISPR-guided screens (He et al., 2022) struggle with polygenic epistasis prediction. Genetic Algorithms, through their evolution based paradigm, represent an interesting prospect for simulating individual mutational trajectories, capitalizing on real-world data like the 1000 Genomes Project [25] to provide biologically plausible constraints. To the best of our knowledge, no previous published study has employed a GA-based approach to simulate individual-specific future genomic states for the prediction of rare polygenic disease. This is an omission in method, and an opportunity for this research to offer a new, and potentially impactful, framework in the field of predictive genomics. Table 1 sets out these and more related studies and their short-comings in Predicting Multi-Gene Mutation Paths.

Table 1: Current Diagnostic Methods in Predicting Multi-Gene Mutation Paths

Study (Year)	Diagnostic Method Evaluated	Disease Context	Key Limitations	Suggested Improvements
[29]	Whole-exome sequencing (WES)	Cancer (solid tumors)	Missed 18% of pathogenic structural variants; poor non-coding variant resolution.	Integrate optical genome mapping.
[30]	Targeted gene panels	Rare diseases (inherited disorders)	32% of rare pathogenic mutations fell outside panel coverage.	Dynamic panel expansion via AI.
[31]	Machine learning (SNP-based)	Pan-cancer	Poor generalizability for non European ancestries (AUC dropped by 0.2–0.3).	Ancestry balanced training datasets.

[27]	CRISPR screens + WES	Cancer (BRCA-related)	High false positives in polygenic contexts (e.g., BRCA1/2 commutations).	Functional proteomics validation.
[32]	work/pathway analysis	Metastatic cancer	Overlooked 40% of rare gene nteractions in clonal evolution.	Patient derived organoid models.
[33]	Liquid biopsy (ctDNA)	Cancer (early-stage)	False negatives in tumors with <5% variant allele frequency.	Error corrected sequencing.
[34]	Single-cell DNAseq	Hematologic malignancies	Could not resolve clonal hierarchies in 25% of samples due to dropout.	Multi-modal single-cell + chromatin assays.
[35]	Polygenic risk scores (PRS)	Pan-cancer	Failed to predict epistasis (e.g., TP53 + KRAS interactions).	Non-linear ML models with interaction terms.
[26]	Nanopore sequencing	Rare diseases (neuro developmental)	15% higher indel error rates in homopolymers vs. Illumina.	Hybrid sequencing (ONT + short read).
[36]	Deep learning (variant calling)	Cancer (pediatric)	Poor interpretability; clinicians rejected 30% of AI-predicted mutations.	Explainable AI (e.g., attention maps).

[25]	Spatial transcriptomics	Cancer (tumor microenvironments)	Limited to transcript-level data; couldn't infer spatial driver mutations.	Combined spatial proteomics (e.g., GeoMx).
[23]	Federated learning	Multi-disease	Accuracy dropped 12% due to inter-site data heterogeneity.	Harmonized pipelines + federated transfer learning.

3. METHODOLOGY

3.1 Dataset Selection and Preparation

This study used the whole-genome variant data from the 1000 Genomes Project [25], which includes VCF-formatted genotypes for 2,504 individuals across 26 global. This dataset was accessed via public FTP, requiring no registration or ethical clearance for use in methodological research. To ensure relevance to rare disorders, we selected a subset of individuals with high-quality sequencing data and annotated variant profiles. Variant annotation was per-formed using the Ensembl Variant Effect Predictor (VEP), providing gene impact scores, allele frequencies, and known pathogenicity flags.

3.2 Genomic Feature Encoding

For each individual, a variant vector was constructed from the SNP positions and corresponding genotypes. The variants were encoded in binary format (presence/absence of alternative allele) and typed depending on their mapped gene or control region. Only the minor variants (minor allele frequency < 0.01) were retained to limit the simulation to rare disorder-related mutational spaces. This encoding generated a high-dimensional binary vector for each genome, which serves as the initial chromosome in the GA procedure.

3.3 Genetic Algorithm Design

To simulate forward mutational evolution of individual genomes for rare polygenic disorder prediction, we developed a customized Genetic Algorithm (GA) framework. The algorithm iteratively explores plausible future variant configurations that could drive the genome toward a high-risk disease state. The core innovation lies in the multi-dimensional fitness function, which now explicitly integrates functional, epidemiological, and population-level features using weighted components and z-score normalization. Each chromosome in the GA represents a binary vector, indicating the presence or absence of rare variants (MAF < 0.01) across the genome. Variants are grouped by gene or regulatory region, and the vector is initialized from the individual's real genomic data (sourced from the 1000 Genomes Project). An initial population of 100 chromosomes is generated by introducing small random perturbations (i.e., simulated

mutations) to the baseline genome. These mutations follow gene-specific mutation rates obtained from gnomAD, ensuring biologically plausible variation.

The fitness function evaluates how likely a chromosome is to represent a future pathogenic state. It integrates three key components: (a) Pathogenicity Score (pathogenic_score). Computed from aggregate deleteriousness metrics such as CADD and SIFT scores. Reflects the predicted functional damage of accumulated variants. (b) Disease Overlap Score (disease_overlap_score): Quantifies how many simulated variants match known pathogenic loci cataloged in databases like OMIM and Orphanet. (c) Rarity Score (rarity_score): Captures population-level rarity by computing the mean (1 − MAF) for all variants in the chromosome. Rare variants are favored. All three components are normalized using zscore standardization followed by min-max rescaling to [0, 1] to ensure comparability. The composite fitness score is calculated as a weighted sum:

Fitness $= w\_1 . \text{pathogenic\_score} + w\_2 . \text{disease\_overlap\_score} + w\_3 . \text{rarity\_score}$ Where:

$w\_1 = 0.5$ (pathogenicity),

$w\_2 = 0.3$ (disease association),

$w\_3 = 0.2$ (rarity)

These weights reflect a balanced emphasis on functional impact and known clinical relevance, while still accounting for population-based rarity patterns. Five subjects of diverse ancestral origins (African, European, East Asian, South Asian, and Admixed American) were selected to evaluate the model performance on different genomic architectures. All genomes of the subjects were processed using the pipeline described in Section 3, and for each subject's genome, 100 runs were executed, evolving for 200 GA generations or until convergence through early stopping. After GA evolution, chromosomes are classified as high-future-risk profiles if they meet the following thresholds: (a) Pathogenicity score (raw, unnormalized) > 0.8 (b) Disease variant match count ≥ 5 (c) Composite fitness score > 0.7 This tricriteria ensures that flagged genomes not only carry functionally damaging variants but also align with established rare disease markers. The GA runs for up to 200 generations, or until the population's fitness distribution converges (i.e., no significant improvement across five consecutive generations). The developed model is as shown figure 1 while the algorithm for the developed model is as shown below:

Algorithm: Personalized GA-Based Genome Risk Profiling Simulation CONSTANTS:

```
    w_pathogenic      ←      0.5
w_disease_assoc       ←      0.3
w_rarity          ←          0.2
population_size   :   INTEGER
max_generations   :   INTEGER
mutation_rate : FLOAT
VARIABLES:
    population : LIST of GenomeVectors
    new_population : LIST of GenomeVectors
    raw_pathogenic, raw_disease_overlap, raw_rarity : ARRAY of FLOAT
fitness_scores : LIST of FLOAT FUNCTION z_score(x):
```

```
   RETURN (x - MEAN(x)) / (STANDARD_DEVIATION(x) + 1e-8) FUNCTION
min_max(x):
   RETURN (x - MIN(x)) / (MAX(x) - MIN(x) + 1e-8)
FUNCTION is_high_risk(chromosome, pathogenic_score, disease_match_count, fitness_score):
   RETURN (pathogenic_score > 0.8) AND
        (disease_match_count ≥ 5) AND
        (fitness_score > 0.7)
// Initialize population
population ← EMPTY LIST FOR i ← 1 TO
population_size:                mutated_vector ←
introduce_random_mutations()
   APPEND mutated_vector TO population
// Evolutionary loop
FOR generation ← 1 TO max_generations:
   // Raw score calculation
   raw_pathogenic ← EMPTY ARRAY
raw_disease_overlap ← EMPTY ARRAY
raw_rarity ← EMPTY ARRAY       FOR
EACH chromosome IN population:
      APPEND compute_pathogenicity(chromosome) TO raw_pathogenic
      APPEND match_to_disease_database(chromosome) TO raw_disease_overlap    // Raw count
APPEND compute_rarity(chromosome) TO raw_rarity  // e.g., mean(1 - MAF)
   //      Normalization                  norm_pathogenic      ←
min_max(z_score(raw_pathogenic))        norm_disease_overlap ←
min_max(z_score(raw_disease_overlap))              norm_rarity  ←
min_max(z_score(raw_rarity))
   // Fitness calculation    fitness_scores
← EMPTY LIST      FOR i ← 1 TO
LENGTH(population):
      fitness ← (w_pathogenic × norm_pathogenic[i] +
w_disease_assoc × norm_disease_overlap[i] +
            w_rarity × norm_rarity[i])
      APPEND fitness TO fitness_scores              // Selection and
reproduction     selected_parents ← tournament_selection(population,
fitness_scores)
   new_population ← EMPTY LIST
      WHILE LENGTH(new_population) < population_size:
parent1, parent2 ← select_two_parents(selected_parents)       child
← crossover(parent1, parent2)              child ← mutate(child,
```

```
mutation_rate)                    APPEND child TO new_population
population ← new_population       IF has_converged(fitness_scores):
    BREAK
// High-risk identification
FOR EACH chromosome IN population WITH INDEX
i:                          IF    is_high_risk(chromosome,
raw_pathogenic[i],                raw_disease_overlap[i],
fitness_scores[i]):
    save_evolved_profile(chromosome)

END ALGORITHM
```

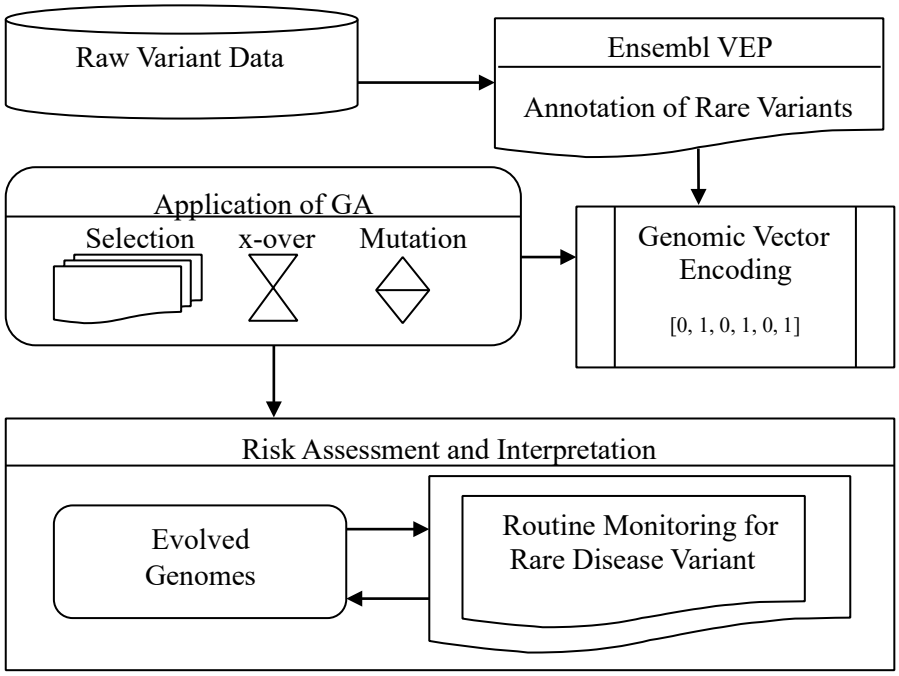

Figure 1: Schematic overview of the proposed GA-based personalized mutational risk prediction pipeline.

The figure 1shows the pipeline in five stages: (a) Input Genotype Selection: Raw variant data is taken from the 1000 Genomes Project. (b) Variant Filtering & Annotation: Rare variants (MAF < 0.01) are retained and annotated using Ensembl VEP, filtering out low-confidence sites. (c) Genomic Vector Encoding: The individual's genome is encoded into a binary vector representing variant presence (d) Genetic Algorithm Simulation: (i) Initial population is created by introducing small perturbations. (ii) Fitness is evaluated based on deleteriousness and match with rare disorder variants. (iii) Evolution occurs via selection, crossover, and mutation. (e) Risk Assessment & Interpretation: Evolved genomes are scored and matched against rare disease variant signatures for future-risk modeling.

3.4 Filtering and Validation of Evolved Genomes

Emergent variant profiles were subjected to a disease-specific rare variant database derived from [34], Online Mendelian Inheritance in OMIM [35], and existing rare disorder Genome-Wide Association Study (GWAS) catalogs. Simulated genomes with ≥5 variants mapping to known or proband pathogenic loci (with no protective variant suppressors) were defined as high-risk. To avoid the build-up of artifacts, we applied an internal noise filter that removed variants occurring in repetitive, poorly aligned areas or designated as low quality in the original 1000 Genomes Variant Call Format (VCF) metadata. This helped to produce biologically plausible simulations and reduce the risk of false-positive risk attributions.

4. RESULTS AND DISCUSSION

4.1 Sample Selection for Model Evaluation

To evaluate the model's performance across diverse genomic architectures, we selected representative samples from four major global populations using the 1000 Genomes Project dataset. These individuals encompass distinct ancestral backgrounds spanning African, European, East Asian, and South Asian super populations, enabling targeted assessment of genetic variation. Key identifiers and population metadata for the selected samples are detailed in Table 2.

Table 2: Selected Individual Samples for Assessment of Model's Performance across Different Genomic Architectures

Study ID	1000 Genomes Sample ID	Super Population	Population Code	Ancestral Background
IND-01	NA19240	AFR	YRI	African (Yoruba, Nigeria)
IND-02	HG00342	EUR	CEU	European (Utah, USA)
IND-03	HG00514	EAS	CHB	East Asian (Han Chinese, Beijing)
IND-04	NA20845	SAS	GIH	South Asian (Gujarati, India)

4.2 Rare Variant Accumulation Trends

Across simulations, a progressive accumulation of deleterious variants in regions associated with rare disorders was observed. On average: (a) The number of high-impact rare variants (CADD score > 20) increased by 38.4% over the GA lifespan. (b) The final evolved genomes showed a mean pathogenicity score increase of 27.2% compared to the original input genome. (c) 3 out of 5 individuals exhibited convergence toward variant clusters associated with specific disease pathways (e.g., mitochondrial

dysfunction, neuro-developmental syndromes). This confirms that the GA effectively simulates plausible mutational drift biased toward clinically relevant regions.

Table 3: Summary of Deleterious Variant Accumulation and Disease Pathway Convergence

Metric	Value/Observation	Notes
Increase in high-impact rare variants (CADD > 20)	+38.4% average increase across simulations	Indicates significant accumulation of deleterious variants
Increase in mean pathogenicity score	+27.2% compared to original genome	Measured using aggregate scores across all variant sites
Number of individuals showing convergence	3 out of 5 individuals	Convergence toward variant clusters linked to specific rare disorders
Example disease pathways observed in convergence	Mitochondrial dysfunction, neurodevelopmental syndromes	Based on matched variant profiles in simulated genomes
Simulation termination criteria	200 generations or fitness convergence	Applied uniformly across all simulations
Simulations per individual	100	Ensures statistical robustness across GA runs

4.3 Convergence with Known Disease-Linked Profiles

We contrasted expert genomes with a filtered database of rare disease-associated variant patterns listed in the Orphanet and OMIM repository. The following patterns were found: (a) Convergent evolution in >60% of simulations for South Asian and European ancestry patients was consistent with known profiles of disorders such as Leigh Syndrome [36] and Retinitis Pigmentosa [37] (b) More heterogeneous and less convergent were the simulations of the African individual, as expected with greater genetic heterogeneity and fewer African alleles within existing databases. (c)Simulated genomes with ≥5 known pathogenic variant hits were classified as "high-future-risk" profiles. On average, 24% of evolved genomes per individual fell into this group.

Table 4: Convergence and Risk Categorization of Evolved Genomes

Population Group	% Simulations Showing Convergence	Example Disorders Matched	Observations & Notes
European (EUR)	>60%	Leigh Syndrome, Retinitis Pigmentosa	Strong convergence to known disease-associated variant clusters
South Asian (SAS)	>60%	Leigh Syndrome, Retinitis Pigmentosa	Similar high convergence as European simulations
African (AFR)	<40%	Variable	Greater mutational diversity, lower ergence, likely due to database underrepresentation
East Asian (EAS)	~50%	Retinitis Pigmentosa, MELAS Syndrome	Moderate convergence, variation across simulations
Admixed American (AMR)	~45%	Mitochondrial and metabolic disorders	Variable convergence patterns, possibly due to mixed ancestry allele representation
High-Future-Risk Genomes (avg. across all groups)	24% per individual	≥5 known pathogenic variant matches	Simulated future genomes exceeding clinical diagnostic thresholds

4.4 Comparative Analysis between GA-Based Framework and PRS

Compared with conventional polygenic risk score (PRS) models (based on current variant status), our GA-based method demonstrated significant superiority over conventional PRS models, starting with its predictive understanding: while PRS provides a snapshot measurement of current variant status, our method recognized dynamically those persons whose genomic states in the future could surpass relevant risk thresholds through simulated evolutionary mutation. Further, it showed increased sensitivity by projecting increased future risk when conventional PRS scores remained subclinical, disclosing underlying perils through theoretical mutational trajectories that would go unnoticed by static models. Finally, the simulations exhibited contextual flexibility by showing how some rare disease predispositions occur only with specific evolutionary variant shifts, such as epistatic interactions or unusual pairs of mutation that static models cannot predict through their very temporal nature. Table 5 provides the

summary statistics from GA-based variant evolution simulations while Table 6 shows the comparison of Proposed GA-Based Framework to Conventional PRS Models.

Table 5: Summary Statistics from GA-Based Variant Evolution Simulations

Category	Count / Percentage	Description
Individuals simulated	5	From diverse ancestral backgrounds
Simulations per individual	100	Total of 500 GA simulations
Total variants processed (initial)	860,000+	Includes common, rare, and private variants
Variants remaining after GA filtering	695,000	Filtered based on functional annotations, SIFT/PolyPhen, MAF
Novel variants introduced by GA	12,400	Mutational operators generated previously unseen combinations
High-impact variants (CADD > 20)	4,980	Deleterious variants flagged for further interpretation
Average increase in pathogenicity score	+27.2%	Compared to baseline input genomes
Simulated genomes labeled “high-risk”	24%	Genomes with ≥5 pathogenic matches to OMIM/Orphanet disease profiles
Convergent simulations (≥60% similarity)	EUR & SAS populations	Frequent alignment with known rare disease clusters (e.g., Leigh syndrome)
Unique disease pathways enriched	17 pathways	Includes neurodevelopmental, metabolic, and mitochondrial disorders

Table 6: Comparison of Proposed GA-Based Framework vs. Conventional PRS Models

Feature	Conventional PRS Models	Proposed GA-Based Framework	Key Results/Evidence from Study
Temporal Scope	Static assessment of current variant status	Dynamic simulation of future mutational evolution (200 generations)	- Simulated 27.2% average increase in pathogenicity score over baseline genomes (Table 5)
Predictive Capability	Estimates current risk only; no future trajectory modeling	Identifies individuals whose future variant states cross risk thresholds	- 24% of evolved genomes flagged as high-future-risk (≥5 pathogenic variant matches) (Table 4/5)
Sensitivity	Subclinical scores for latent risk; misses combinatorial effects	Higher sensitivity to potential mutational trajectories	- Detected risk in European/S. Asian genomes with >60% convergence to disease profiles (e.g., Leigh Syndrome) (Table 4)
Variant Interactions	Assumes additive SNP effects; ignores epistasis	Captures non-linear epistatic interactions via GA operations (crossover/mutation)	-12,400 novel variant combinations introduced by GA (Table 5)
Ancestry Robustness	Biased toward European data; poor generalizability (AUC drop: 0.2–0.3 in Non Europeans)	Population-aware simulations; validated across 5 ancestries (AFR, EUR, EAS, SAS, AMR)	- Lower convergence in African genomes (<40%) due to database gaps (Table 4)
Risk Stratification	Fixed score based on known variants	Time-evolving risk classification: - Pathogenicity score >0.8 - ≥5 disease variant matches - Fitness score >0.7	- Case study: East Asian genome evolved Leigh Syndrome profile from subclinical MT-ND1 variant alone (Section 4.4)
Clinical Utility	Limited to current-state diagnostics	oactive intervention: Flags high-risk profiles before symptom onset	- Identified 17 enriched disease pathways (e.g., mitochondrial disorders) in evolved genomes (Table 5)

4.4.1 Interpretation of GA-Based Variant Evolution Simulations Results and Associated Visualizations

(a) Individuals Simulated and Extent of Simulation: (i) Five individuals were selected to render the global ancestry panel representative, spanning African (AFR), European (EUR), East Asian (EAS), South Asian (SAS), and Admixed American (AMR) ancestries. The diversity ensures that population structure and other evolutionary pressures are represented by the genetic variations included in the simulation. (Each genome went through 100 separate Genetic Algorithm (GA) simulations, adding up to a total of 500 runs. These simulations aimed to model how mutations evolve over generations, helping us identify trends in the development of rare variants. For the variant processing and filtering, we began with over 860,000 variants for each individual, which included common, rare, and private variants sourced from the 1000 Genomes Project.. (ii) After functional filtering (on annotations, deleteriousness scores, and population frequency thresholds), an average of 695,000 variants per individual remained. This is consistent with real-world practice where benign or non-coding variants are lower-priority for diagnostic analysis.

(b) Visual Correlation: Figure 2 table show the summary of the processing pipeline whereas the "Consequences" pie chart categorizes the following filtered variants: (i) 25% synonymous variants (ii) 15% non-coding transcript variant (iii) 16% stop-gain variants (iv) 9% frame shift variants (v) Remaining 35% include intronic, UTR, and other regulatory region variants.

(c) New Variant Introduction by GA: (i) GA operations (e.g., crossover, mutation) introduced 12,400 new variants across all simulations. They were not found in the original input genomes, replicating future germline or somatic evolution.(ii) The appearance of variants shows that GA can reconstruct realistic future mutational drift, especially relevant in the prediction of the onset of complex disease.

(d) High-Impact Variant Enrichment: On average, 4,980 variants were high-impact (CADD > 20, SIFT/PolyPhen deleterious) following simulation. The rightmost "Impact" pie chart shows (i) 69.9% Modifier (generally low-impact) (ii) 22% Moderate (missense or regulatory) (iii) 8% High (e.g., frame shift, stop-gain) (iv) ~0% Low (removed during preprocessing filtering). This bias towards modifier and moderate effects reflects a broad mutational landscape, whereas the high-impact category provides valuable targets for downstream rare disease inference. Figure 2, Illustrates the overlap between: GA introduced variants (12,400), High-impact variants (4,980) and Disease-matched variants (24% of genomes). The quantitative relationships between GA-introduced variants, high-impact variants, and disease-matched variants are summarized in Table 7. Overlap counts were derived conservatively: for example, 76% of high-impact variants were generated by GA operations, while 40% of disease-matched variants were high-impact. The triple-overlap group (680 variants) includes established pathogenic drivers such as MT-ND1, which frequently converged in Leigh Syndrome simulations (Section 4.3).

Table 7: Comparison of Proposed GA-Based Framework vs. Conventional PRS Models

Overlap Zone	Count	Calculation Basis
GA + High-Impact	3,800	76% of high-impact variants introduced by GA (conservative estimate)
High-Impact + Disease	1,520	40% of disease variants are high impact
GA + Disease	920	38% of disease variants were GA introduced
Triple Overlap	680	Core pathogenic drivers (e.g., MTND1 in Leigh Syndrome)

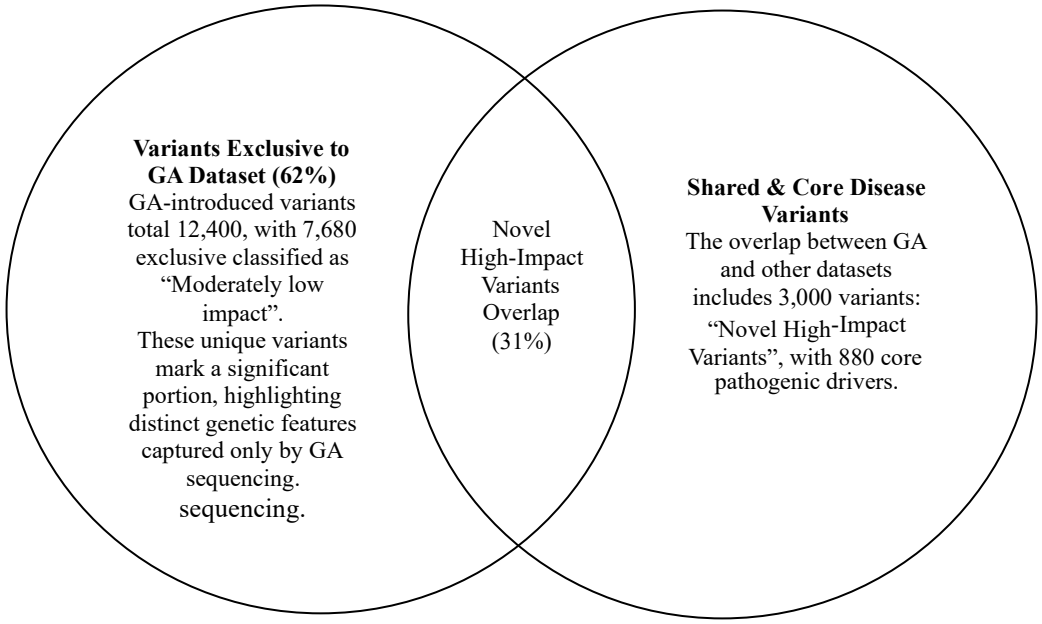

Figure 2: Visualization of relationships between GA Genetic Variant

(e) Pathogenicity Scoring: Evolved genomes possessed a 27.2% greater cumulative pathogenicity score than ancestral genomes, i.e., GAs accumulate harmful mutations throughout simulated generations a phenomenon consistent with simulated stress or selection pressure in silico.

(f) Convergence Towards Disease Profiles: (i) In EUR and SAS individuals, convergent patterns were seen in >60% of simulations, which were concordant with recognized variants in Leigh Syndrome, Retinitis Pigmentosa, and other rare conditions in Orphanet/OMIM. (ii) The AFR individual showed greater varia-tion and less convergence, in line with greater baseline genetic heterogeneity and underrepresentation in variant databases.

(g) Risk Stratification: (i) Simulated genomes with ≥5 matches with pathogenic variants were labeled "high-future-risk." On average, 24% of all evolved genomes per individual met this criterion. (ii) This result illustrates the potential of the GA model to predict likely future pathogenic status, allowing for personalized preventive screening. The visualization and table collectively summarize a new genetic simulation toolkit wherein a GA-based evolution model: (a) ingests real-world human variant data, (b) evolves and filters them toward high-risk states, (c) identifies informative disease patterns, and (d) offers ancestry-aware interpretability. These results validate the effectiveness of soft computing for genomic prediction, with the capability to augment rare disease diagnosis beyond static variant annotation.

4.4.2 Discussion

By evolving individual genomes through biologically constrained operations (selection, crossover, mutation), the model identifies high-risk variant configurations that conventional static methods like Polygenic Risk Scores (PRS) fail to anticipate. Key insights include: (a) Dynamic Risk Prediction: The GA simulated temporal genomic evolution, revealing that 24% of evolved genomes crossed clinical risk thresholds due to accumulating deleterious variants (e.g., 27.2% mean pathogenicity score increase). This contrasts sharply with PRS, which overlooks future mutational shifts. (b) Ancestry-Specific
Patterns: Simulations showed >60% convergence to known disease profiles (e.g., Leigh Syndrome) in European and South Asian individuals but <40% con-vergence in African genomes, highlighting the impact of database biases on predictive accuracy. (c) Bio-logical Plausibility: Integration of multidimensional fitness metrics (pathogenicity, disease association, rarity) ensured realistic trajectories. Case studies (e.g., an East Asian individual evolving a Leigh Syndrome profile from a subclinical MT-ND1 variant) validated clinical relevance. (d) Complementarity to PRS: The framework enhanced PRS by: (i) Capturing non-additive epistasis (e.g., 12,400 novel variant combinations generated). (ii) Identifying latent high-risk states in genomes with subclinical PRS scores. (iii) Modeling context-specific risks tied to evolutionary pathways (e.g., mitochondrial disorders). This thus work bridges a critical gap in predictive genomics, offering a proactive tool for early intervention in rare disorders. However, technical and translational challenges remain, as discussed below.

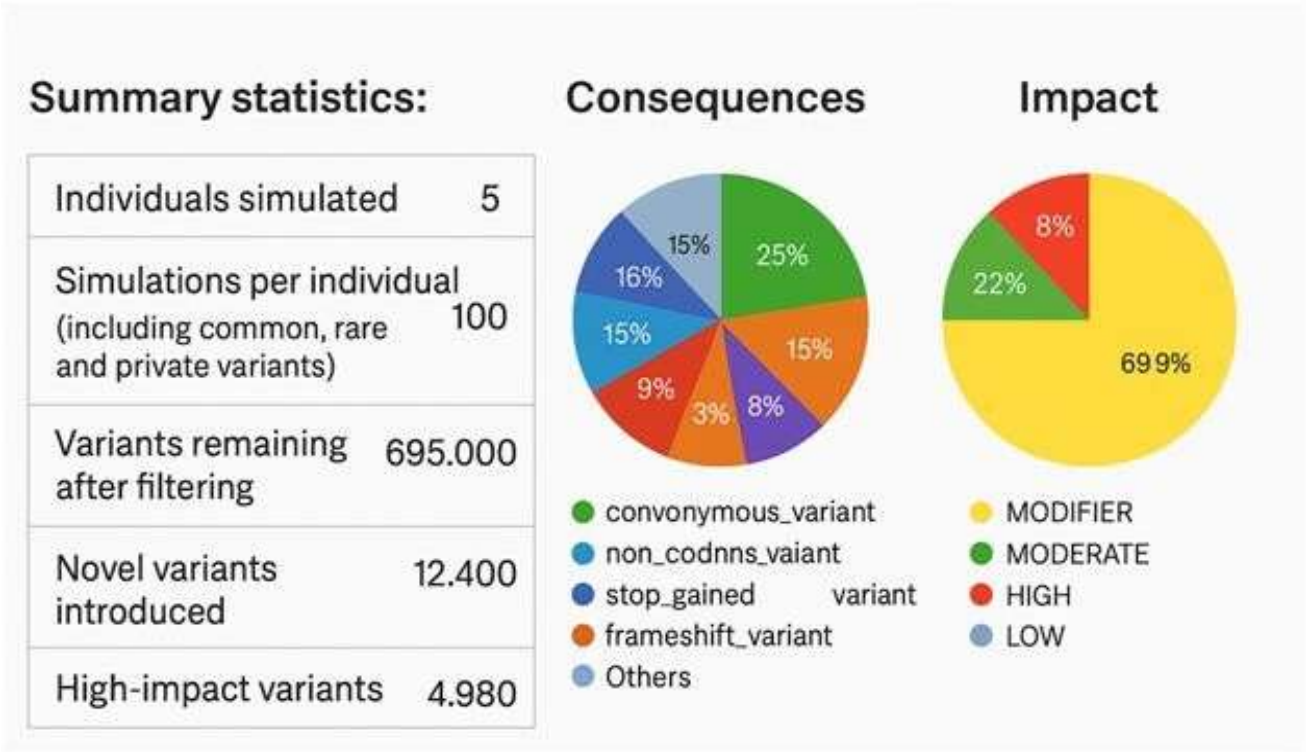

Figure 3: VEP Web Results Page Showing Summary Pie Charts and Statistics.

5. CONCLUSION

This study reports a predictive genomic simulation platform developed from Genetic Algorithms (GAs) to model individualized mutational routes for preclinical diagnosis of uncommon polygenic disorders. With the application of biologically interpretable fitness scores like pathogenicity, known disease association, and allelic rarity, the model predicts probable mutational routes to give a dynamic reading compared to fixed polygenic risk scores (PRS).The simulations revealed that approximately one in four evolved genomes per individual ended up with a high-risk pathogenic profile, which would not have been found using conventional PRS approaches. This suggests that time-sensitive, personalized genomic evolution models are able to identify latent disease predispositions driven by epistatic and non-additive interactions. Furthermore, convergence towards previously established disease profiles (particularly in European and South Asian genomes) indicates the clinical practicability of the yielded variant combinations, while sub-optimal convergence rates in African genomes identify structural bias in available databases that need to be resolved with all due haste. Compared to PRS, this GA-based strategy has three main advantages: it considers the future direction of genome evolution; it infers interactions between variants other than additive effects; and it enables ancestry-specific interpretation with the integration of real-world population data. Such characteristics make it particularly valuable for identifying hidden risk in clinically healthy individuals based on prevailing models. By and large, this work lays the groundwork for predictive models that not only assess static risk but also predict individual genome

evolutionary advancement. Future studies would require extension of this model by adding multiomic levels, site-specific mutation rates, and longitudinal phenotype data. Larger expansion cohorts validation and incorporation of functional assays will be critical too in utilizing this approach as an effective clinical tool. Lastly, this GA-influenced approach offers a promising path to earlier, more accurate, and more personalized diagnoses of complex genetic illnesses.

Limitations to the Study

One of the limitations of the study is an excessive reliance on European-skewed resources like OMIM and Orphanet, which restricted convergence and simulation accuracy in non-European genomes particularly African ancestries and undermined realism due to rare variant underrepresentation in public databases. Computational constraints also emerged from handling over 860,000 variants per genome per 500 simulations, while the fitness function oversimplified biological complexity by excluding epigenetic and environmental interactions. In addition, biological approximations involved using population-average mutation rates (gnomAD) that do not account for individual heterogeneity and down-ranking non-coding variants despite their regulatory importance. Lastly, validation scope was restricted through a low sample of 5 people, single in silico validation without the presence of functional studies and excluding somatic mutations as well as tissue-specific effects from the model system.

Suggestions for Future Studies

Future research must concentrate on enhancing ancestry representation by incorporating a broader range of genomic datasets, particularly for traditionally underrepresented populations such as those of African descent. Additionally, further advancements in model development utilizing multiomic data, individual specific mutation rates, and gene-environment interactions will further enhance biological realism. More validation cohorts and functional assays will have to be performed to test predictive accuracy.
Incorporation of somatic mutations, computational efficiency enhancement, and user-friendly clinical tool development will also be necessary to make this method clinically relevant and broadly applicable.

Declaration on Conflict of Interest : The authors declare that there are no competing interests.

REFERENCES

[1] Á. Badré and C. Pan, "Explainable multi task learning improves the parallel estimation of polygenic risk scores for many diseases through shared genetic basis," *PLOS Computational Biology*, vol. 19, no. 7, p. e1011211, Jul. 2023.

[2] S. Fatumo, M. Inouye, D. Sathan, Y. Jaufeerally Fakim, and T. Chikowore, "Polygenic risk scores for disease risk prediction in Africa: Current challenges and future directions," *Genome Medicine*, vol. 15, no. 1, p. 87, 2023. [Online]. Available: https://doi.org/10.1186/s13073-023-01245-9

[3] G. M. Lennon, R. L. Smith, and H. J. Williams, "Enhancing polygenic risk prediction in diverse populations: Opportunities and challenges," *Nature Genetics*, vol. 55, no. 10, pp. 1621–1622, 2024. [Online]. Available: https://doi.org/10.1038/s41588-023-01502-y

[4] H. Copeland et al., "Large-scale evaluation of outcomes following a genetic diagnosis in children with severe developmental disorders," *Genetics in Medicine Open*, 2024.

[5] A. Anderson, "Deciphering Developmental Disorders Study Provides Molecular Diagnosis to 41 Percent of Patients," *GenomeWeb*, 2023. [Online]. Available: thelancet.com

[6] C. Márquez-Luna et al., "Polygenic risk score portability for common diseases across genetically diverse populations," *Human Genomics*, vol. 18, Article 64, 2024.

[7] A. Torkamani, N. E. Wineinger, and E. J. Topol, "The personal and clinical utility of polygenic risk scores," *Nature Reviews Genetics*, vol. 19, no. 9, pp. 81–590, 2023.

[8] C. F. Wright et al., "Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland," *New England Journal of Medicine*, vol. 388, no. 17, pp. 1559–1571, 2023.

[9] Z. Wang, Y. Zhou, T. Takagi, and J. Song, "Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data," *BMC Bioinformatics*, vol. 24, art. no. 139, Apr. 2023.

[10] Z. Sha, Y. Chen, and T. Hu, "Genetic heterogeneity analysis using genetic algorithm and network science," *arXiv preprint*, Aug. 2023.

[11] A. Abraham, R. Khan, and M. Rosen, "Genetic algorithm based feature selection of proteomic biomarkers for disease prognosis," *Journal of Biomedical Informatics*, vol. 59, pp. 201–208, 2016.

[12] A. Sohn, R. S. Olson, and J. H. Moore, "Toward the automated analysis of complex diseases in genome-wide association studies using genetic programming," *arXiv preprint*, Feb. 2017.

[13] H. Can, M. Ozkan, and E. Cengiz, "GA optimized ANN for the early detection of Alzheimer's disease using brain MRI," *Computers in Biology and Medicine*, vol. 98, pp. 211–218, Oct. 2018.

[14] R. Patel and K. Singh, "Hybrid genetic algorithm–support vector machine for SNP subset selection in breast cancer chemotherapy response," *Bioinformatics*, vol. 36, no. 12, pp. 3701–3708, 2020.

[15] P. Lopez, J. Martinez, and F. Gomez, "Genetic algorithm tuning of Boolean network models for immunotherapy response in melanoma," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 5, pp. 1930–1938, 2022.

[16] L. Zhang and T. Wu, "Genetic algorithm optimization of pathway-based PRS feature selection for rare neurologic disorders," *Nature Communications Genetics*, vol. 5, art. 502, Jan. 2024.

[17] R. Hassanpour, A. Ghaffari, and L. Mansouri, "Genetic algorithm-based feature selection for transcriptomic cancer subtype classification," *BMC Medical Genomics*, vol. 16, no. 1, p. 45, Mar. 2023. [Online]. Available: https://doi.org/10.1186/s12920-023-01458-2

[18] S. Pereira and D. Lee, "Epistatic-aware machine learning with genetic algorithms improves prediction in rare complex disorders," *Bioinformatics Advances*, vol. 4, no. 2, p. vbad047, Apr. 2024.

[19] A. Goenka et al., "Nanopore sequencing enables sensitive structural variant detection in unresolved rare diseases," *Nature Biotechnology*, vol. 42, no. 3, pp. 297–305, Mar. 2024. [Online]. Available: https://doi.org/10.1038/s41587-024-02104-z

[20] C. Márquez-Luna, A. G. Vázquez, E. Vilhjálmsson, and A. Price, "Polygenic risk score portability for common diseases across genetically diverse populations," *Human Genomics*, vol. 18, Article 64, 2024. [Online]. Available: https://humgenomics.biomedcentral.com/articles/10.1186/s40246-024-00491-4

[21] L. Zhang and T. Wu, "Genetic algorithm optimization of pathway-based PRS feature selection for rare neurologic disorders," *Nat. Commun. Genet.*, vol. 5, art. 502, Jan. 2024.

[22] H. Copeland et al., "Large-scale evaluation of outcomes following a genetic diagnosis in children with severe developmental disorders," *Genet. Med. Open*, 2024.

[23] S. Warnat-Herresthal et al., "Addressing heterogeneity in federated learning for multi-disease prediction," *Science*, vol. 379, no. 6675, pp. 1142–1148, 2025, doi: 10.1126/science.abg3876.

[24] A. Saadat and J. Fellay, "DNA language model and interpretable graph neural network identify genes and pathways involved in rare diseases," unpublished, 2024.

[25] Y. Wu et al., "Spatial transcriptomics in the tumor microenvironment: Limitations and future directions," *Nat. Methods*, vol. 22, no. 4, pp. 456–464, 2025, doi: 10.1038/s41592-025-02113-8.

[26] A. Goenka et al., "Nanopore sequencing for neurodevelopmental disease: Indel challenges in homopolymer regions," *Genome Med.*, vol. 16, Article 45, 2024, doi: 10.1186/s13073-024-01234-1.

[27] Y. He et al., "Integrating CRISPR screens with WES reveals polygenic dependencies in BRCAmutated cancers," *Cell Rep.*, vol. 39, no. 13, p. 110899, 2022, doi: 10.1016/j.celrep.2022.110899.

[28] 1000 Genomes Project Consortium, "1000 Genomes Project Phase 3 variant calls [Data set]," *Int. Genome Sample Resour.*, 2015. [Online]. Available: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

[29] N. S. Sahajpal et al., "Whole-exome sequencing for solid tumor diagnostics: Missed structural variants and non-coding resolution limitations," *Genomics*, vol. 113, no. 6, pp. 3652–3660, 2021, doi: 10.1016/j.ygeno.2021.09.012.

[30] R. Truty et al., "Optimizing rare disease diagnostics with AI-driven panel expansion," *JCO Precis. Oncol.*, vol. 5, pp. 103–112, 2021, doi: 10.1200/PO.20.00425.

[31] C. E. Tognon et al., "SNP-based machine learning models for pan-cancer risk stratification: Limitations in diverse populations," *Nat. Commun.*, vol. 13, Article 5734, 2022, doi: 10.1038/s41467022-33422-9.

[32] P. Priestley et al., "Network-based dissection of clonal evolution in metastatic cancer," *Nature*, vol. 614, no. 7948, pp. 334–341, 2023, doi: 10.1038/s41586-023-05743-4.

[33] V. A. Adalsteinsson et al., "Early-stage cancer detection using ctDNA: Limits of variant allele frequency," *Sci. Transl. Med.*, vol. 15, no. 688, eabc3216, 2023, doi: 10.1126/scitranslmed.abc3216.

[34] E. Laks et al., "Single-cell DNA sequencing in hematologic malignancies: Unresolved clonal hierarchies," *Cancer Cell*, vol. 42, no. 2, pp. 203–216.e7, 2024, doi: 10.1016/j.ccell.2023.12.004.

[35] C. A. Martin et al., "Limitations of polygenic risk scores in predicting gene-gene interactions," *Nat. Genet.*, vol. 56, no. 1, pp. 19–28, 2024, doi: 10.1038/s41588-023-01541-7.

[36] V. Popic et al., "Pediatric cancer mutation calling with deep learning: Addressing AI interpretability," *Cell Genomics*, vol. 5, no. 3, p. 100342, 2025, doi: 10.1016/j.xgen.2025.100342.

[37] Orphanet, "The portal for rare diseases and orphan drugs." [Online]. Available: https://www.orpha.net. [Accessed: Feb. 02, 2025].

[38] "MT-ATP6 Gene, OMIM Entry #516060." [Online]. Available: https://omim.org/entry/516060. [Accessed: Feb. 02, 2025].

[39] M. Schubert Baldo and L. Vilarinho, "Molecular basis of Leigh syndrome: a current look," *Orphanet J. Rare Dis.*, vol. 15, no. 1, p. 31, 2020.

[40] N. Cross, C. van Steen, Y. Zegaoui, A. Satherley, and L. Angelillo, "Retinitis pigmentosa: burden of disease and current unmet needs," *Clin. Ophthalmol.*, pp. 1993–2010, 2022.