OXFORD

Briefings in Bioinformatics, 2025, **26(5)**, bbaf475

https://doi.org/10.1093/bib/bbaf475

Problem Solving Protocol

Clinical and data-driven optimization of Genomiser for rare disease patients: experience from the Hong Kong Genome Project

Anson Man Chun Xi[‡], Denis Long Him Yeung[‡], Wei Ma, Dingge Ying, Amy Hin Yan Tong, Dicky Or, Shirley Pik Ying Hue, Hong Kong Genome Project, Annie Tsz-Wai Chu, Brian Hon-Yin Chung[*]

Hong Kong Genome Institute, 2/F, Building 20E, Hong Kong Science Park, Hong Kong Special Administrative Region, China

[*]Corresponding author. Hong Kong Genome Institute, Hong Kong Special Administrative Region, China. E-mail: bhychung@genomics.org.hk

[‡]Anson Man Chun Xi and Denis Long Him Yeung contributed equally to this work.

The Hong Kong Genome Institute oversees Hong Kong's first large-scale genome sequencing project, the Hong Kong Genome Project, to advance genomic research and drive personalized genomic medicine.

Abstract

Genomiser is a phenotype-driven tool that prioritizes coding and non-coding variants by relevance in rare disease diagnosis; yet comprehensive evaluation of its performance on real-life whole genome sequencing data is lacking. The Hong Kong Genome Project had initially incorporated Exomiser in the diagnostic pipeline. This study evaluated the feasibility of upgrading from Exomiser to Genomiser with three modifications: extension of the interval filter to include ±2000 bp from transcript boundaries, adjusting minor allele frequency (MAF) filter to 3%, and the inclusion of SpliceAI. A total of 985 patients with disclosed whole genome sequencing test results were included in this study, of which 207 positive cases (14 attributed to non-coding variants) were used for Genomiser parameter optimization by means of sensitivity evaluation. Under the default parameter setting, Genomiser achieved lower sensitivity compared to Exomiser (70.15% vs. 72.14%, top-3 candidates; 74.63% vs. 80.60%, top-5 candidates). Further investigation noted that this was attributed to non-coding variant noise influenced by Regulatory Mendelian Mutation (ReMM) scoring metrics. This issue was mitigated when a previously optimized ReMM score was applied as a filtering cut-off (ReMM = 0.963), improving Genomiser's sensitivity (92.54% vs. 89.55%, top-15 candidates). We further evaluated the optimized parameter in a cohort of 778 negative cases and detected 20 non-coding variants (2.6% added yield), with 5 validated to be disease-causing. Our proposed approach adheres to American College of Medical Genetics and Genomics/Association for Molecular Pathology and ClinGen variant interpretation guidelines to ensure interpretable results and integrates non-coding variant analysis into clinical pipelines.

Keywords: whole genome sequencing; short-read genome sequencing; Genomiser; Exomiser; ReMM; variant prioritization; Hong Kong Genome Project; rare disease

Introduction

The clinical utilization of whole genome sequencing (WGS) as first-tier diagnostic testing for rare disease (RD) has become increasingly relevant, attributed to its improved sensitivity [1, 2] and high-throughput sequencing data generation that allows the detection of technically challenging variants in comparison to conventional molecular testing [3]. WGS may also offer superiority in re-analysis as novel gene-disease association arises [4]. To enhance the utility of WGS, a wide range of *in silico* tools have been developed to prioritize clinically significant variants based on probabilistic functional-prediction algorithms, conservativeness measurement, or by employing ensemble and linear machine learning methods [5–7].

Exomiser is a variant prioritization tool that integrates phenotypic information encoded with Human Phenotype Ontology (HPO) terms and genotypic information from next-generation sequencing (NGS) data to rank candidate variants by their relevance [8]. The tool was initially developed for whole-exome sequencing (WES) analysis [9, 10] which permits automated functional annotation of single nucleotide variants (SNVs) in coding regions, leveraging on pathogenicity prediction for missense conservation by PolyPhen [11], MutationTaster [12], and SIFT [13]. Over the years, later versions of Exomiser have incorporated more sophisticated prediction models including Rare Exome Variant Ensemble Learner (REVEL) [14], Missense Variant Pathogenicity Prediction (MVP) [15], and AlphaMissense [16] to enhance variant prioritization performance. Various national-scale NGS projects have adopted Exomiser as part of the diagnostic framework for RD patients, including the 100 000 Genome Project by Genomics England [17], the TRANSLATE NAMSE project under the genomDE initiative of Germany [18], and the Hong Kong Genome Project (HKGP) by the Hong Kong Genome Institute [19].

With the implementation of WGS, Genomiser was introduced as an extension of Exomiser with the capability to process

Received: May 6, 2025. **Revised:** July 21, 2025. **Accepted:** August 20, 2025

WGS data. The expanded tool incorporates a bespoke machine learning method known as the Regulatory Mendelian Mutation (ReMM) framework to score for regulatory elements, alongside SpliceAI, which predicts potential disruptions in splicing. These integrations significantly enhanced the capability of Genomiser to uncover novel gene-disease association and improved clinical diagnosis of non-coding variants [20]. Despite advances in prioritization tools that offer non-coding variant analysis, the interpretation of non-coding variants presents unique challenges that complicate clinical application. Current limitations include both a scarcity of validated datasets for benchmarking and inherent biases in existing tools, which often focus narrowly on specific genomic contexts—as exemplified by SpliceAI and MaxEntScan's exclusive emphasis on splice-altering variants. While these tools are ClinGen endorsed and demonstrate robust performance for their intended purposes [21, 22], their restricted scope and the frequent need for functional validation highlight critical gaps in comprehensive non-coding variant analysis. These challenges collectively underscore the importance of systematic parameter optimization to improve the reliability of non-coding variant prioritization in clinical practice.

To the best of our knowledge, there are a limited number of studies that evaluate Genomiser and its sensitivity performance using real-world WGS data. Most existing research has relied solely on the default settings of Genomiser, without optimization of its parameters. Moreover, there is limited evaluation concerning the incorporation of later introduced pathogenicity prediction sources, such as SpliceAI and AlphaMissense, in the context of clinical diagnosis [16, 23]. It has been previously shown that Genomiser suffered from reduced sensitivity without proper constitution of parameter settings [20]. It has been demonstrated that if only simple filtering of common variants with a minor allele frequency >1% was applied, ReMM failed to prioritize any causative variant as the top hit and placed only 7% of causative variants within the top-10 ranks [20]. Moreover, in the latest edition of the Critical Assessment of Genome Interpretation assessment (CAGI6), Genomiser showed lower sensitivity compared to Exomiser, despite demonstrating potential in identifying unsolved diagnoses attributed to non-coding variants [24]. This suggests that the parameter settings in Genomiser, merely extended from Exomiser's default without adjustments for the newly added predictive tools limits its effectiveness. Therefore, meticulous optimization of Genomiser's default parameters would be prudent to fully harness its diagnostic potential.

In this study, we evaluated the potential of Genomiser to interpret non-coding variants and enhance WGS diagnostic yield in a real-world clinical setting. We applied Genomiser in a WGS-based analytical framework and compared its sensitivity to that of Exomiser in our existing pipeline.

Materials and methods

Study participants and consent

The HKGP is the first large-scale genome sequencing project in Hong Kong, conducted to enhance the clinical application of WGS and to promote genomic medicine research. The implementation of the project was coordinated by the Hong Kong Genome Institute (HKGI) and proceeded through collaborations with the Hong Kong Department of Health, the Hong Kong Hospital Authority, and medical schools of local universities [19, 25]. Patients were recruited through the three Partnering Centers (PCs), as described previously [19]. Clinical details of patients, concerning family history, personal medical history, and past investigations conducted were recorded by genetic counselors. Phenotypic information of patients was recorded using HPO terms to standardize clinical data input for Exomiser and Genomiser analysis. Genetic counseling practices were aligned with centralized training for genetic counselors fostered by the HKGI to ensure accurate selection of HPO terms. An internal clinical portal, which supports the housing and handling of patients' clinical data, was developed to promote efficient information exchange between the HKGI and PCs. The portal has a built-in search engine for HPO terms that further secures the consistency in HPO term input [19].

Written informed consent was obtained from all participants at pre-test counseling sessions, and ethics approvals were obtained from the Central Institutional Review Board (IRB) (HKGP-2021-001, HKGP2022–001) and IRBs of the Department of Health (L/M 257/2021), Joint Chinese University of Hong Kong New Territories East Cluster (2021·423, 2023·120), and The University of Hong Kong/Hospital Authority Hong Kong West Cluster (UW 21–413, UW 23–289).

Illumina short-read GS and secondary analysis

The detailed workflow for srGS primary and secondary analysis has been described previously [25]. The HKGP follows in-house standard operation protocols for whole blood sample collection and transfer from the Partnering Centre to the HKGI laboratory, followed by genomic DNA extraction (gDNA) and quality measurements (QC) in the laboratory [19, 25].

Briefly, the extracted gDNA samples were first quality checked to ensure a minimum concentration of 28 ng/μL, without degradation, an OD 260/280 between 1.7–2.0, and an OD 260/230 between 1.8 and 2.5. They were proceeded for PCR-free library preparation with manufacturer's protocol of the KAPA HyperPlus Kit (Kapa Biosystems Inc.). Size selection was performed using KAPA HyperPure Beads (Roche). The final libraries were checked for library concentration and quantified by qPCR. The final fragment size of the libraries was determined using Agilent 4200 TapeStation (Agilent Technologies Inc.) to confirm a final library range of 400–700 bp. Libraries passing QC were sequenced with Illumina NovaSeq™ 6000 or NovaSeq™ X Plus sequencers (Illumina Inc.). Sequencing data required a mean coverage of ≥30X in reads, with a depth of >15X for ≥95% of the human reference genome (GRCh38/hg38). The generated srGS data then underwent sequence quality checks, and secondary analysis was performed by an in-house bioinformatics pipeline with base calling on the DRAGEN server (v4.3.6); sequence alignment of GRCh38 reference genome by BWA (v0.7.17) [26]; and duplicates removal by Picard (v2.27.4). SNV calling was performed by GATK HaplotypeCaller (v4.2.6.1) [27], and the output was used to run Exomiser and Genomiser [20, 28].

Exomiser-based curation workflow and tertiary analysis approach in HKGP

The detailed workflow for tertiary analysis has been described previously, where sensitivity plateaued by examining only the top-10 variants, and expansion to the top-15 was unnecessary [25]. Briefly, the top-10 variants for each mode of inheritance (MOI) pattern generated from Exomiser prioritization output are first reviewed; this is equivalent to the screening of contributing variants in ~40 prioritized candidates concerning autosomal dominant, autosomal recessive, X-linked dominant, and X-linked recessive MOI patterns. In complementary, additional screening was performed with the application of virtual gene panels from Genomic England PanelApp and PanelApp Australia [29, 30]. The pathogenicity of variants was interpreted according to

the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) guidelines and ClinGen sequence variant interpretation working group (SVI WG) recommendations.

Evaluation and optimization of Genomiser setting

Genomiser V14.0.0 with the July 2024 database release (V.2406) and Exomiser V12.1.0 with the February 2021 database release (V.2102) were used for sensitivity performance evaluation. The two tools were run using configurations described in the default setting of the official Exomiser documentation [31], with minor modifications. Briefly, the interval filter was defined based on the 'RefSeq Select and MANE' (ncbiRefSeqSelect) track obtained from the UCSC table browser, and was extended by 2000 base pairs upstream of the transcription start site and downstream of the transcription end site. The Frequency Filter was adjusted to a maximum frequency of 3%. All samples were run on singleton VCF input to avoid reduction in variant score for inherited pathogenic variants with reduced penetrance [25]. The original ranking listed in the output and the candidate ranking of identified variants after filtering based on ReMM (V0.4) score threshold were recorded and the assessment was conducted on prioritized variants of the top-1, top-3, top-5, top-10, top-15, and top-20 ranked candidates regardless of the suspected MOI.

Between December 2019 and December 2023, WGS test results from 985 patients were issued by HKGP. Among these 985 patients, causative variants were identified in 207 patients, confirming the genetic diagnoses. Causative variants were classified as pathogenic or likely pathogenic according to ACMG/AMP guidelines and ClinGen recommendations, fourteen of which were validated as non-coding variants. In the remaining 778 patients with negative WGS reports, no causative variant(s) were identified. We utilized the 207 positive cases for Genomiser parameter optimization based on pathogenicity prediction sources for coding and non-coding variants; the optimized parameter was later applied to the reanalysis of the 778 negative cases to evaluate Genomiser's performance (Fig. 1). Available pathogenicity prediction scores in Genomiser for coding variants include PolyPhen, MutationTaster, SIFT, REVEL, MVP, and AlphaMissense. Only REVEL, MVP, and AlphaMissense were utilized in our analysis as PolyPhen, MutationTaster, and SIFT have been incorporated into REVEL based on the ensemble prediction model. This also aligns with the Exomizer default recommendation in enabling REVEL and MVP for coding variants [31].

Cumulative sensitivity differences in prioritizing candidates within the top-1, top-3, top-5, top-10, top-15, and top-20 rankings were assessed across the different combinatory uses of REVEL, with iterations of adding MVP, or AlphaMissense, or both. For non-coding variants, the pathogenicity prediction sources of ReMM and SpliceAI were investigated to detect candidate regulatory variants and intronic splicing variants.

Results

Overview of HKGP benchmarking cohort

Primary clinical indications of the 985 probands were diverse, with neurologic disorders (17.7%), neurodevelopmental disorders (16.8%), and cardiovascular disorder (13.5%) being the most common referred disease categories (Supplementary Table S1).

Evaluation of pathogenicity prediction sources for coding variants

We tested different iterations of pathogenicity prediction sources for coding variants. REVEL was included in all iterations and served as a baseline due to its superior performance using an ensemble method compared to individual prediction sources such as MutationTaster, Polyphen2, and SIFT, as demonstrated in ClinGen recommendations for PP3/BP4 criteria guideline [32]. Its application has been largely integrated into the ACMG/AMP guideline [33, 34]. MVP and AlphaMissense have been demonstrated to have improved accuracy and therefore were also tested in this study [15, 16]. We observed a marginal increase in sensitivity when REVEL, MVP, and AlphaMissense were all enabled compared to the default setting of only REVEL being enabled. Based on this evaluation, all three available coding variant pathogenicity prediction sources were incorporated into the subsequent performance evaluation of Genomiser to maximize sensitivity in detecting pathogenic coding variants (Fig. 2).

Evaluation of non-coding pathogenicity prediction sources

The pathogenicity prediction sources for non-coding variants (ReMM and SpliceAI) were evaluated based on 14 cases with established disease-causing non-coding variants. These variants were not detected by Exomiser and were only diagnosed using our virtual gene panels targeting the patient's specific disease presentation after manual curation (Table 1). SpliceAI scored all splice variants ≥0.46, demonstrating high sensitivity. It is also a validated tool endorsed by the ClinGen SVI WG [35]. Therefore, it was included in all subsequent analyses. For variants affecting regulatory elements, such as the *GJB1* 5′UTR variant in patients HKGP001248-1 and HKGP002113-1, and the *SMARCB1* 3′UTR variant in patient HKGP002428–1, only the ReMM score contributed to the final variant score in Genomiser [36–38]. Further interrogation of the ReMM scores detected varying performance of the tool to detect true positive variants, as demonstrated by difference in the positive *SMARCB1* 3′UTR variant with a lower ReMM score of 0.66624 compared to 0.995748 of the positive *GJB1* 5′UTR variant. To maximize the variant types detectable by Genomiser, ReMM was also included for downstream parameter optimization.

Sensitivity of Exomiser and Genomiser under default setting

Under the default setting, Exomiser prioritized 51.74%, 72.14%, 80.60%, 88.06%, 89.55%, and 90.05% of the 207 positive diagnoses in the top-1, top-3, top-5, top-10, top-15, and top-20 ranked candidates, respectively. In comparison, Genomiser showed a poorer performance and was able to prioritize 58.71%, 70.15%, 74.63%, 80.10%, 85.07%, and 86.57% at the top-1, top-3, top-5, top-10, top-15, and top-20 ranked candidates, respectively (Fig. 3). This relative underperformance of Genomiser was observed in 60 cases (out of 207 positive cases) despite comparable variant scores were assigned to the causative variant in Exomiser and Genomiser. Further investigation noted that this underperformance was attributed to high ReMM scores in many non-coding variants other than the true causative variant, which led to a lower rank of the causative variant in Genomiser (Supplementary Table S2). For example, when considering candidate variants prioritized by Genomiser in patient HKGP000304–1 with intellectual disability and behavioural abnormalities, the true causative frameshift

Table 1. Non-coding variants undetected by Exomiser in the initial diagnosis.

Patient ID	Referral indications	Variant(s) identified	MOI	ClinVar ID	Exomiser ranking	Genomiser ranking	ALL_PATH	ACMG classification
HKGP000500–1	Developmental epileptic encephalopathy	NM_020320.5(*RARS2*):c.685C>T p.(Arg229Ter)	AR	N/A	Undetected	172nd	N/A	Likely pathogenic (PVS1, PM2_P)
		NM_020320.5(*RARS2*):c.1238-28T>G p.?		N/A			REMM = 0.164809, SPLICE_AI = 0.46	Likely pathogenic (PVS1(RNA), PM3, PM2_P)
HKGP000625–1	Neurofibromatosis type 1	NM_001042492.3(*NF1*):c.2410-16A>G p.?	AD	572474	Undetected	1st	REMM = 0.809865, SPLICE_AI = 0.96	Pathogenic (PS4, PS2, PP3, PM2_P)
HKGP002117–1	Neurofibromatosis type 1	NM_001042492.3(*NF1*):c.5812 + 332A>G p.?	AD	216065	Undetected	1st	REMM = 0.642750, SPLICE_AI = 0.75	Pathogenic (PVS1(RNA), PS4, PP1_S, PS2_M, PM2_P)
HKGP002602–1	Neurofibromatosis type 1	NM_001042492.3(*NF1*):c.1392 + 751T>G p.?	AD	N/A	Undetected	1st	REMM = 0.995589, SPLICE_AI = 0.97	Pathogenic (PP4_S, PM2_P, PP1, PP4)
HKGP001367–1	Neurofibromatosis type 1	NM_001042492.3(*NF1*):c.1260 + 1604A>G p.?	AD	428941	Undetected	1st	REMM = 0.98019, SPLICE_AI = 0.76	Pathogenic (PVS1(RNA), PS4, PM2_P)
HKGP000370–1	Cone rod dystrophy	NM_001298.3(*CNGA3*):c.540dup p.(Tyr181LeufsTer2)	AR	N/A	Undetected	1st	N/A	Likely pathogenic (PVS1, PM2_P)
		NM_001298.3(*CNGA3*):c.396-11C > G		2734247			REMM = 0.905352, SPLICE_AI = 0.87	Likely pathogenic (PM3_S, PS3_P, PM2_P)
HKGP000402–1	Arthrogyrosis multiplex	NM_004826.4(*ECEL1*):c.1810G>A p.(Gly604Arg)	AR	1705569	Undetected	25th	REVEL = 0.955, MVP = 0.953464, SPLICE_AI = 0.69, ALPHA_MISSENSE = 0.9956	Likely pathogenic (PP3_S, PM3, PM2_P)
		NM_004826.4(*ECEL1*):c.2152-15C>A p.?		N/A			REMM = 0.0277 SPLICE_AI = 1.0	Likely pathogenic (PM3_S, PP3, PM2_P)
HKGP002217–1	Congenital adrenal hyperplasia	NM_000500.9(*CYP21A2*):c.1069C>T p.(Arg357Trp)	AR	12152	Undetected	1st	REVEL = 0.68, MVP = 0.9751358, SPLICE_AI = 0.13, ALPHA_MISSENSE = 0.4711	Pathogenic (PM3_VS, PS3_P, PM2_P, PP3)
		NM_000500.9(*CYP21A2*):c.293-13C>G p.?		12155			REMM = 0.00711508, REVEL = 0.044, MVP = 0.24569798, SPLICE_AI = 0.75, ALPHA_MISSENSE = 0.07	Pathogenic (PM3_VS, PP1_M, PP3)
HKGP001248–1	Charcot Marie Tooth disease; multiple exotosis	NM_000127.3(*EXT1*):c.1469del p.(Leu490fsTer9)	AD	265131	1st	3rd	N/A	Pathogenic (PVS1, PS4, PM2_P)
		NM_000166.6(*GJB1*):c.-103C>T p.?	XL	217166	Undetected	1st	REMM = 0.995748	Pathogenic (PS4, PP1_S, PM2_P)
HKGP002113–1	Charcot Marie Tooth disease	NM_000166.6(*GJB1*):c.-103C>T p.?	XL	217166	Undetected	2nd	REMM = 0.995748	Pathogenic (PS4, PP1_S, PM2_P)
HKGP000706–1	Multiple congenital anomalies	NM_017780.4(*CHD7*):c.5405-17G > A p.?	AD	195978	Undetected	1st	REMM = 0.134139, SPLICE_AI = 1.0	Pathogenic (PS2_VS, PS4, PVS1_M(RNA), PM2_P)
HKGP000995–1	Polycystic kidney disease	NM_001009944.3(*PKD1*): c.10167 + 25_10,167 + 43del p.?	AD	586238	Undetected	1st	REMM = 0.698712, SPLICE_AI = 0.67	Pathogenic (PS4_S, PVS1_M(RNA), PP4_M, PM2_P, PP1)
HKGP001361–1	High myopia	NM_001844.5(*COL2A1*):c.1527 + 135G>A p.?	AD	812294	Undetected	1st	REMM = 0.0, SPLICE_AI = 0.57	Likely pathogenic (PP1_S, PS4_P, PP3, PM2_P)
HKGP002428–1	Tetroma, edometriotic cyst	NM_003073.5(*SMARCB1*):c.*82C>T p.?	AD	239481	Undetected	8th	REMM = 0.666224	Likely pathogenic (PS4_M, PP1_M, PS3_P, PM2_P)

Non-coding variants undetected by Exomiser in the initial diagnosis of fifteen cases of optimization cohort. ALL_PATH signifies the scores predicted by available non-coding variant pathogenicity sources. SpliceAI and ReMM scores are based on a probabilistic scale of 0 to 1; the highest score of 1 signifies the variant's position as having the greatest potential of imposing regulatory consequence or being splice-altering if mutated for ReMM and SpliceAI, respectively. MOI, mode of inheritance; AD, autosomal dominant; AR, autosomal recessive; XL, X-linked.

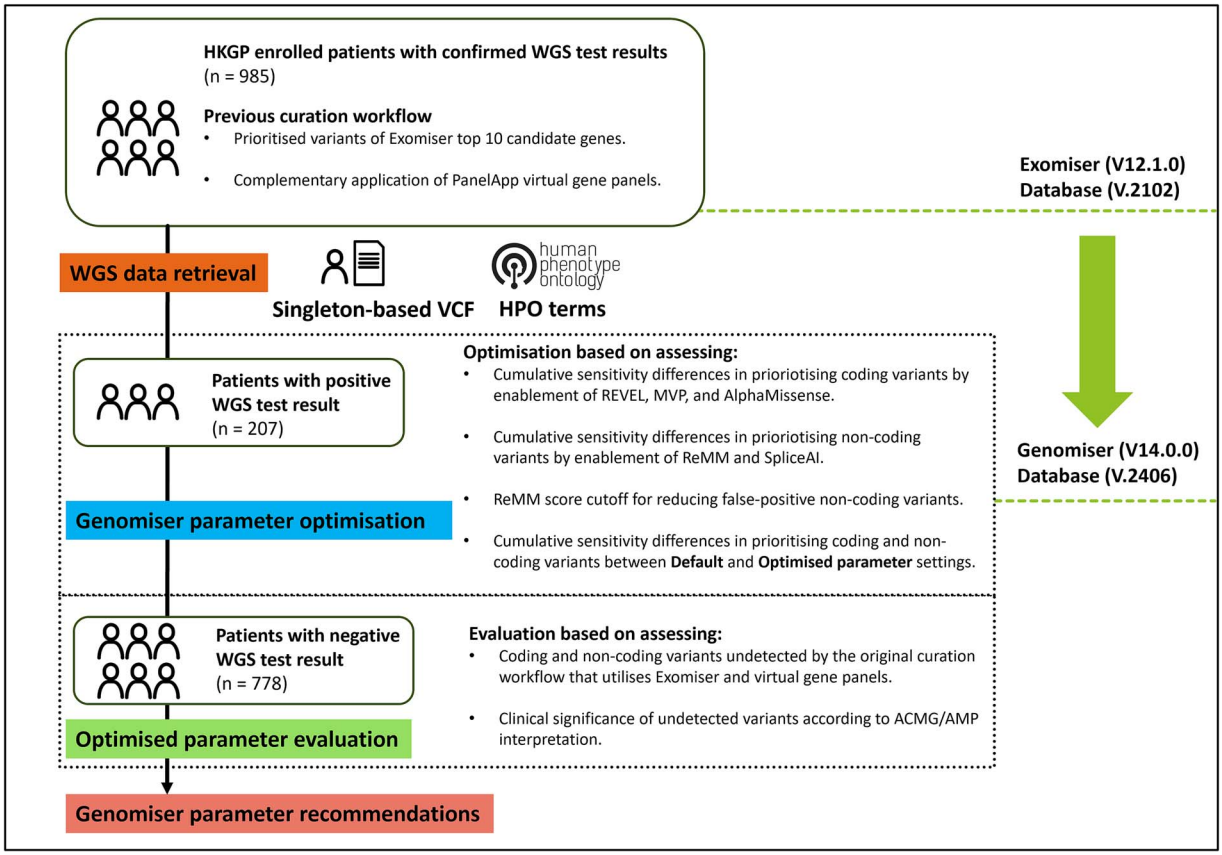

Figure 1. Overview of the study's workflow.

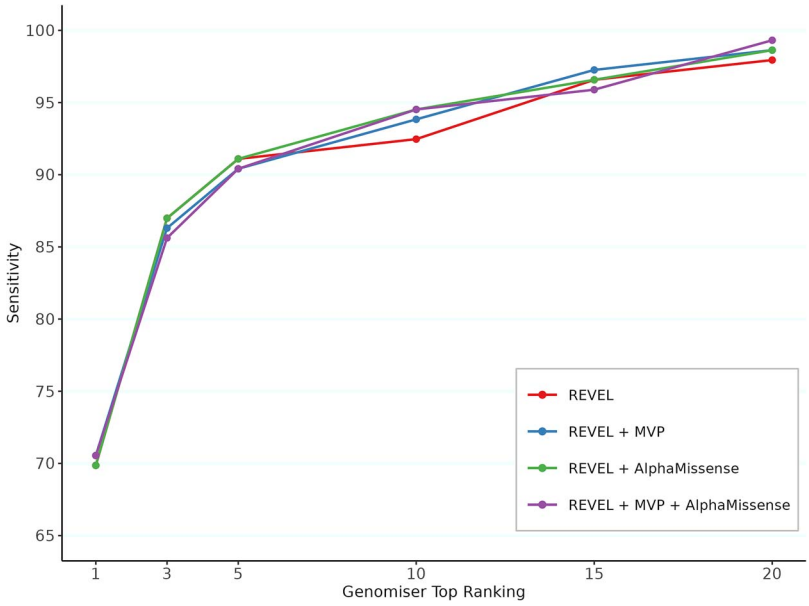

Figure 2. Cumulative sensitivity distribution curve of Genomiser using different coding variant pathogenicity source combinations.

variant of *PHIP* was ranked seventh, which was outranked by six other non-coding variants with scores of >0.9, which ReMM was the sole pathogenicity prediction source contributing to the final variant score. In patients having non-specific phenotypes with high genetic heterogeneity, such as non-syndromic intellectual disability, many genes received high phenotype scores due to increased likelihood of having overlapped phenotypes with the patient. Combining with the array of variants with high ReMM scores, this led to frequent inclusion of false-positive non-coding variants into the top-ranked list (Supplementary Table S3).

Optimization of filtering parameter for ReMM score

To improve the implementation of Genomiser into our curation pipeline, we next investigated the best strategy to minimize the prioritization of these false-positive non-coding variants based

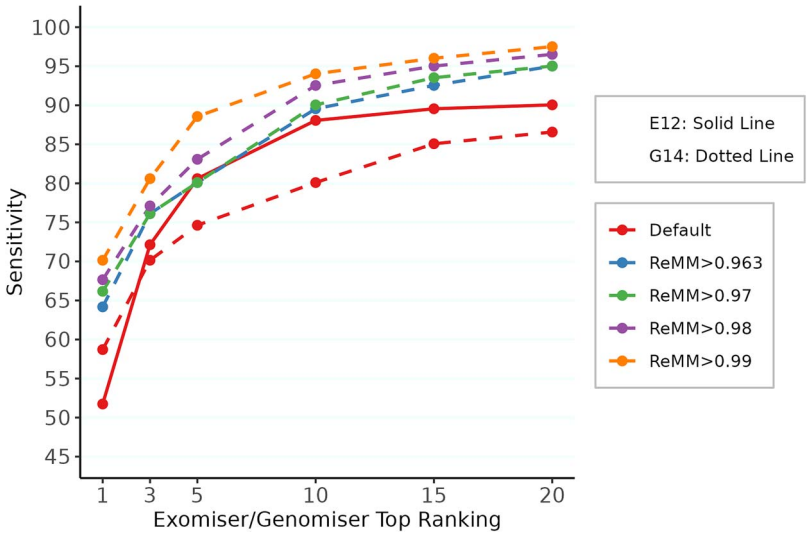

Figure 3. Cumulative sensitivity distribution curve of Exomiser under default setting, and Genomiser under default and optimized settings. E12, Exomiser version 12.1.0; G14, Genomiser version 14.0.0.

on the ReMM score. A previously optimized ReMM score (0.963) associated with the highest F1-score by one of the developers of Genomiser was selected as the baseline threshold for evaluation [39]. Three additional thresholds at 0.1 intervals (0.97, 0.98, and 0.99) were also included in the Genomiser sensitivity evaluation. After applying the ReMM threshold filter, monoallelic candidate variants associated with autosomal recessive disease were filtered. Using this approach, the performance of Genomiser has improved, with a sensitivity of 92.54% being reached in prioritizing true positive variant in the top 15 rank, compared to 89.55% achieved by Exomiser (Fig. 3). Overall, there is a positive correlation between Genomiser sensitivity and increasing ReMM score thresholds (Fig. 3). The sensitivity performance difference between Genomiser and Exomiser was most noticeable at the highest score threshold of ReMM >0.99, which the former reaching a sensitivity of 96.02% in the top-15 rank compared to 89.55% of Exomiser (Fig. 3). Nonetheless, when concerning the aforementioned ReMM validation study [39], we noted a significant reduction in recall rate when the ReMM score is leaning towards a more stringent threshold of 0.99; this would result in a likelihood of missing the true-positive pathogenic and likely pathogenic non-coding variants in the prioritization despite a higher specificity is achieved. Following this evaluation and balancing specificity, the threshold of ReMM score was set at 0.963.

Performance evaluation of optimized Genomiser parameter and identification of novel clinically significant variants

Performance of Genomiser using the optimized parameters was validated in the re-analysis of the set of 778 HKGP cases that received negative WGS test results. Under the optimized parameter, Genomiser identified 20 clinically significant deep intronic variants with splicing implications predicted by SpliceAI that matched well with the patients' phenotypes. MaxEntScan, another widely used splicing prediction tool used by ClinGen VCEPs, was used to confirm the prediction of splicing effect (Table 2). No clinically significant variants of regulatory implications were identified by ReMM. Altogether, this has added a yield of 2.6% (20/778) following the updated curation pipeline, mainly concerning the upgrade from Exomiser to Genomiser alongside the optimized parameter setting of the latter.

Notably, five out of the twenty deep intronic variants can be directly classified as pathogenic or likely pathogenic according to the ACMG/AMP guidelines and contributed to positive diagnoses. These variants were not prioritized by Exomiser and were previously missed (Table 2). A notable example is patient HKGP000035–1 with suspected Dravet syndrome, who received extensive investigation in the *SCN1A* gene by both targeted Sanger sequencing, MLPA, and NGS using targeted epilepsy panel with all negative results. Additional WGS analysis conducted through HKGP using Exomiser for prioritization identified no significant findings. The optimized Genomiser successfully prioritized a *de novo SCN1A* deep intronic variant (c.2415 + 431C>G) at the eighth rank, which was predicted by SpliceAI and MaxEntScan to produce a cryptic splice donor (Table 2). The variant was classified as likely pathogenic owing to the patient's phenotypic specificity and its *de novo* occurrence, which permits the application of PS2 and PS4_Moderate according to the ClinGen *SCN1A* criteria specific interpretation guideline. This diagnosis has better informed patient's seizure control and treatment implications, concerning that conventional anti-seizure mediations, such as Carbamazepine, may worsen *SCN1A*-related seizures [40].

In addition, the remaining fifteen deep intronic variants prioritized by Genomiser were identified to be related to the referral indications of the patients and were regarded to be of potential clinical significance (Supplementary Table S4). Unlike the aforementioned patient, phenotypic presentation for these cases was not discriminative enough to enable the application of phenotypic specificity criterion, and functional validation data concerning molecular consequences for these variants were not available at the time of assessment to allow functional interpretation. An example is patient HKGP000602-1 presented with infantile-onset developmental epileptic encephalopathy and progressive myoclonus. The patient was suspected of having a mitochondrial disorder and was offered WES analysis of movement disorder-related genes alongside NGS-based whole mitochondrial genome analysis by the referral clinician previously. The optimized Genomiser identified a paternally inherited deep intronic variant of *NPRL3* at the 13th rank which may explain the patient's neurological and neurodevelopmental features. The identification of this variant through the optimized Genomiser demonstrated the utility of its phenotype-driven prioritization approach in detecting variants that extend beyond

Table 2. Pathogenic/likely pathogenic intronic variants identified in the reanalysis of performance validation cohort.

Patient ID	Referral indications	Variant(s) identified	MOI	Genomiser ranking	ClinVar ID	Pathogenicity sources		Additional validation	ACMG classification
						SpliceAI	ReMM	MaxEntScan	
HKGP000647–1	Neurofibromatosis type 1	NM_001042492.3(*NF1*):c.5812 + 332A>G p.?	AD	1st	216065	0.75	0.642750	Cryptic donor (10.16)	Pathogenic (PVS1 (RNA), PS2_M, PM2_P)
HKGP000340–1	Neurofibromatosis type 1	NM_001042492.3(*NF1*):c.6921 + 586A>G p.?	AD	1st	N/A	0.77	0.596687	Cryptic donor (9.55)	Likely pathogenic (PP4_S, PS2_M, PM2_P, PP3)
HKGP000462–1	Werner syndrome	NM_000553.6(*WRN*):c.3244del p.(Val1082TyrfsTer17)	AR	1st	1076156	N/A	N/A	N/A	Likely pathogenic (PVS1, PM2_P)
		NM_000553.6(*WRN*):c.724 + 515A>G p.?			2081507	0.49	0.707267	Cryptic donor (7.23)	Likely pathogenic (PP4_S, PM3, PM2_P, PP3)
HKGP000035–1	Dravet syndrome	NM_001165963.4(*SCN1A*):c.2415 + 431C>G p.?	AD	8th	N/A	0.94	0.025674	Cryptic donor (9.11)	Likely pathogenic (PS2, PS4_M, PM2_P, PP3)
HKGP000302–1	Duchenne muscular dystrophy	NM_004006.3(*DMD*):c.6290 + 3076A>G p.(Thr3055SerfsTer1)	XL	3rd	3377324	0.27	0.705289	Cryptic acceptor (8.83)	Likely pathogenic (PVS1(RNA), PM2_P)

Pathogenic/likely pathogenic intronic variants identified in the reanalysis of validation cohort (778 negative cases). The SpliceAI and MaxEntScan scores of the five identified intronic variants showed concordant predictive results that suggest the variants' function implication on splicing effect. The variants have been curated to be pathogenic/likely pathogenic according to the ACMG/AMP interpretation guideline and ClinGen SVI recommendation. MOI, mode of inheritance; AD, autosomal dominant; AR, autosomal recessive; XL, X-linked.

the initial clinical suspicion of mitochondrial disease. Although the variant was classified as a variant of uncertain significance (VUS) due to lack of supportive ACMG/AMP criterion assignment, its detection has informed new follow-up actions regarding functional validation by RNA sequencing and fuel discussions on the variant's impact.

Discussion

To the best of our knowledge, there is a limited number of benchmarked prioritization frameworks like Genomiser that simultaneously analyze both coding and non-coding variants [41–43] and is updated regularly [44–47]. Quality control measures for variant input and accurate HPO term entry are imperative to ensure consistency and feasibility of Genomiser. The first part is secured by strict QC parameters set for sequencing runs at HKGI. Effective alignment of practice and training provided to genetic counselors ensured the accuracy and uniformity of HPO term selection. Owing to the high quality of our data, a high diagnostic yield was achieved by Genomiser, compared to other studies even with the use of default setting [24]. In this study, we evaluated and optimized Genomiser's parameters using a large cohort of real-world data consisting 985 HKGP patients to further enhance its performance for clinical implementation. The default Genomiser configuration from the developer was applied with two modifications (Supplementary Fig. S1). A genomic region interval filter including all genes and ±2000 bp from gene boundaries was applied to include proximal enhancers, core promoters, and all intronic regions. The MAF filter was set at 3% to include variants with reduced penetrance.

While evaluating prioritization performance of Genomiser, we observed that non-coding variants with relatively higher ReMM score tend to outrank the true causative variant under Genomiser's default setting. This practical observation should be noted when utilizing Genomiser prioritization for the interpretation of non-coding regulatory variants in clinical settings. While disabling the ReMM is one option, we caution that this approach may limit a comprehensive assessment of non-coding variants, as that ReMM and SpliceAI approaches are trained on different rationales. Through our evaluation of the impact brought about by ReMM score towards variant prioritization using real-world data, our study extends beyond the current pool of evidence on Exomiser sensitivity performance for coding variants [5–7, 18] and offers new insights for the practical application of Genomiser to prioritize regulatory variants leveraged from WGS data. Several thresholds of ReMM score filter were therefore tested, aiming at reducing the false-positive variants prioritized by it. On applying the threshold of 0.963 for ReMM, Genomiser sensitivity towards prioritizing true-positive variants has increased by 7.5%, when compared to that of Genomiser under default setting when considering the top-15 ranked candidates. Further increasing the ReMM score threshold did result in higher sensitivity toward prioritizing true causative variants but also led to a higher likelihood of missing the true-positive non-coding variants, with a significantly lower recall rate [39]. While evaluating parameter optimization, we also noted that the ReMM scores differed significantly for the causative 5′UTR variant of *GJB1* and 3′UTR variant of *SMARCB1* (patients HKGP001248–1, HKGP002113–1, and HKGP002428–1, respectively). This is perhaps due to the notion that only about 1 in 294 regulatory variants are classified as pathogenic with high confidence in the ClinVar database [48]. This limitation in validated non-coding variant datasets limits accuracy for training *in silico* tools. Genomiser, like any

other prioritization tools that support functional annotation of non-coding variants, is subjected to this intrinsic limitation and may be biased when predicting the deleteriousness of certain subsets of variants [48]. Nonetheless, the 3′UTR variant of *SMARCB1* will still be incorporated into the output due to Genomiser's formulated whitelist, which will consider the variant to be pathogenic regardless of the underlying pathogenicity prediction score. Through our evaluation using real-world data with conclusive WGS results that were strictly founded on the consensus reached by ACMG/AMP, we have shown that the incorporation of ReMM into Genomiser is feasible and may facilitate clinical diagnosis, attributed to its success in prioritizing disease-causing regulatory variants. On the other hand, SpliceAI has consistently demonstrated better performance in predicting pathogenicity of spliceogenic variants compared to other splicing impact prediction tools [22, 35, 49]. This approach would only require a slight adjustment to the parameters, mainly concerning ReMM score filtering. On top of that, our optimized parameter has incorporated prediction sources that have been endorsed by ClinGen SVI WG and various ClinGen VCEPs, namely REVEL and SpliceAI. This provides the basis for assigning strong level of evidence for pathogenicity when curating candidate variants and permits the standardized adaptation of Genomiser across different clinical diagnostic laboratories. Following this evaluation, we recommend running Genomiser with the enablement of pathogenicity sources—REVEL, MVP, AlphaMissense, SpliceAI, and ReMM with consideration of applying ReMM score threshold of 0.963 to minimize the inclusion of false-positive non-coding variants. This would garner a sensitivity of 92.54% to be reached that only requires the manual reviewing of the top-15 candidates generated by a single run of Genomiser. This approach is simple and clinically applicable, which requires only one round of analysis of prioritized variants from Genomiser to handle both coding and non-coding variants compared to other sophisticated workflows proposed by other groups [50].

To benchmark the performance of the optimized Genomiser settings, a cohort of 778 HKGP patients with previous negative WGS findings were reanalyzed. Since prediction tools endorsed by ClinGen expert groups were already used during variant prioritization in Genomiser, this allowed ACMG/AMP-based variant classification to be applied straightly after prioritization and minimized the curation time. In the reanalysis of the 778 negative cases, we have identified 20 new candidates that were previously missed [25]. Considering the 20 newly identified deep intronic variants that were interpreted to be of clinical significance through our optimized Genomiser parameter, the added yield of 2.6% (20/778) is comparable to that of the prevalence of deep intronic variants identified in RD cohorts across literature, which may vary between 1% and 7%, and such variants are more likely to be classified as variants of VUSs [17, 48, 51–57] (Supplementary Table S5). Indeed, 15 out of 20 deep intronic variants that we have identified in these negative cases were yet to have been reported in the literature or submitted into the ClinVar or LOVD databases. These variants could only be classified as VUSs at the time of assessment, and functional validation is necessary to fully comprehend their contribution to the clinical presentation of patients. In this regard, in-depth discussion regarding these 15 novel deep intronic variants will be conducted with referral clinicians by means of multi-disciplinary meetings to decide on follow-up actions concerning additional phenotyping and experimental investigation. Nonetheless, our optimized parameter shows the potential of identifying and prioritizing variants based on phenotypic features, which may be helpful in facilitating clinical diagnosis, especially when functional evidence of variants may not always be present.

One limitation of Genomiser is its dependence on *in silico* predictions for non-coding variants, which are developed using datasets that are biased towards splicing variants. This reflects broader challenges in training prioritization tools on limited validated datasets of non-coding variants. This could be partially solved by the phenotype-driven scoring in Genomiser which helps recover clinically relevant candidates. For instance, the widely used prediction tools for splicing variants, SpliceAI, has been shown to have a sensitivity of 41% in detecting deep intronic spliceogenic variants [23], and likely due to the highly imbalance distribution of pathogenic and benign variants for training [58]. This notion echoes with the recommendation from expert panels, which states that performance of *in silico* tools are often dependent on their positive training sets, and functional studies often remain as the only way to determine a variant's impact effectively [48].

Overall, this study has demonstrated how Genomiser may serve as part of a robust analytic framework once its parameter setting has been optimized, thereby allowing the diagnostic power of WGS to be fully harnessed. This has allowed us to expand the scope of our variant prioritization pipeline so that we may now perform routine pursuit of non-coding variant diagnosis without compromizing specificity towards coding variants. Using real-life cases with negative test results, we have identified novel non-coding variants and demonstrated that Genomiser may uplift WGS diagnosis. Following this study, we have now incorporated the optimized Genomiser into our curation pipeline to help identify undetected non-coding variants in WGS analysis. For candidate regulatory variants prioritized by the optimized Genomiser that are associated with consistent phenotypes with patients in the 778 negative cases, these are shortlisted for further investigation. Moving forward, the HKGP will be expanding towards applying multi-omics approaches such as DNA methylation and transcriptomics analysis to validate the 15 novel deep intronic VUSs. Furthermore, to better represent the practicality of the Genomiser under clinical setting, the tool will continually be re-evaluated in the future once more positive cases that concern disease-causing non-coding variants have accumulated.

Key Points

- We proposed an optimized parameter and curation approach of Genomiser that specifically addresses non-coding variant noise attributed to ReMM prediction, thereby allowing variant prioritization at a whole-genome level.
- By evaluating whole genome sequencing data of real-world patients with rare diseases, the optimized Genomiser successfully prioritized 92.54% of true-positive causative variants in the top-15 candidate ranks.
- Reanalysis of 778 previously undiagnosed rare disease cases using the optimized Genomiser identified 20 novel deep intronic variants predicted to disrupt splicing.
- The optimized Genomiser adheres to the American College of Medical Genetics and Genomics/Association for Molecular Pathology and ClinGen variant interpretation guidelines to ensure interpretable results and integrates non-coding variant analysis into clinical pipelines.

Acknowledgments

The authors would like to thank Dr. Su-vui Lo, Chief Executive Officer of the Hong Kong Genome Institute, for her instrumental leadership and guidance. We also thank all patients and family members for participating in the Project. We wish to acknowledge the support of the healthcare and recruitment teams at the Partnering Centres of HKGI: The University of Hong Kong/Queen Mary Hospital, The Chinese University of Hong Kong/Prince of Wales Hospital, and Hong Kong Children's Hospital. Finally, the authors would like to express their gratitude to the HKGI Board of Directors and Advisory Committees and HKGP stakeholders: The Health Bureau, Hospital Authority, and Department of Health.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: All authors declare that there are no competing interests.

Funding

The HKGP is a publicly funded genome sequencing initiative commissioned by the Health Bureau of the HKSAR Government. The sponsor had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data.

Data availability

The details of variants identified in this study are available at the Leiden Open Variation Database (LOVD).

References

1. Sun Y, Liu F, Fan C. *et al.* Characterizing sensitivity and coverage of clinical WGS as a diagnostic test for genetic disorders. *BMC Méd Genom* 2021;**14**:102. https://doi.org/10.1186/s12920-021-00948-5
2. Ewans LJ, Minoche AE, Schofield D. *et al.* Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis. *Eur J Hum Genet* 2022;**30**:1121–31. https://doi.org/10.1038/s41431-022-01162-2
3. Boycott KM, Hartley T, Biesecker LG. *et al.* A diagnosis for all rare genetic diseases: the horizon and the next frontiers. *Cell* 2019;**177**:32–7. https://doi.org/10.1016/j.cell.2019.02.040
4. Cirino AL, Lakdawala NK, McDonough B. *et al.* A comparison of whole genome sequencing to multigene panel testing in hypertrophic cardiomyopathy patients. *Circ Cardiovasc Genet* 2017;**10**:e001768. https://doi.org/10.1161/CIRCGENETICS.117.001768
5. Yuan X, Wang J, Dai B. *et al.* Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Brief Bioinform* 2022;**23**:bbac019. https://doi.org/10.1093/bib/bbac019
6. Yuan X, Su J, Wang J. *et al.* Refined preferences of prioritizers improve intelligent diagnosis for Mendelian diseases. *Sci Rep* 2024;**14**:2845. https://doi.org/10.1038/s41598-024-53461-x
7. Schlüter A, Vélez-Santamaría V, Verdura E. *et al.* ClinPrior: an algorithm for diagnosis and novel gene discovery by network-based prioritization. *Genome Med* 2023;**15**:68. https://doi.org/10.1186/s13073-023-01214-2
8. Danzi MC, Dohrn MF, Fazal S. *et al.* Deep structured learning for variant prioritization in Mendelian diseases. *Nat Commun* 2023;**14**:4167. https://doi.org/10.1038/s41467-023-39306-7
9. Bone WP, Washington NL, Buske OJ. *et al.* Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet Med* 2016;**18**:608–17. https://doi.org/10.1038/gim.2015.137
10. Cipriani V, Pontikos N, Arno G. *et al.* An improved phenotype-driven tool for rare mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data. *Genes* 2020;**11**:460. https://doi.org/10.3390/genes11040460
11. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002;**30**:3894–900. https://doi.org/10.1093/nar/gkf493
12. Schwarz JM, Rödelsperger C, Schuelke M. *et al.* MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**:575–6. https://doi.org/10.1038/nmeth0810-575
13. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812–4. https://doi.org/10.1093/nar/gkg509
14. Ioannidis NM, Rothstein JH, Pejaver V. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;**99**:877–85. https://doi.org/10.1016/j.ajhg.2016.08.016
15. Qi H, Zhang H, Zhao Y. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat Commun* 2021;**12**:510. https://doi.org/10.1038/s41467-020-20847-0
16. Cheng J, Novati G, Pan J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492. https://doi.org/10.1126/science.adg7492
17. Smedley D, Smith KR, Martin AR. *et al.* 100,000 Genomes pilot on rare-disease diagnosis in health care—preliminary report. *N Engl J Med* 2021;**385**:1868–80. https://doi.org/10.1056/NEJMoa2035790
18. Schmidt A, Danyel M, Grundmann K. *et al.* Next-generation phenotyping integrated in a national framework for patients with ultrarare disorders improves genetic diagnostics and yields new molecular findings. *Nat Genet* 2024;**56**:1644–53. https://doi.org/10.1038/s41588-024-01836-1
19. Chu ATW, Tong AHY, Tse DMS. *et al.* The Hong Kong genome project: building genome sequencing capacity and capability for advancing genomic science in Hong Kong. *J Transl Genet Genom* 2023;**7**:196–212. https://doi.org/10.20517/jtgg.2023.22
20. Smedley D, Schubach M, Jacobsen JOB. *et al.* A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am J Hum Genet* 2016;**99**:595–606. https://doi.org/10.1016/j.ajhg.2016.07.005
21. Lee K, Krempely K, Roberts ME. *et al.* Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Hum Mutat* 2018;**39**:1553–68. https://doi.org/10.1002/humu.23650
22. Riepe TV, Khan M, Roosing S. *et al.* Benchmarking deep learning splice prediction tools using functional splice assays. *Hum Mutat* 2021;**42**:799–810. https://doi.org/10.1002/humu.24212
23. Jaganathan K, Panagiotopoulou SK, McRae JF. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* 2019;**176**:535–548.e24. https://doi.org/10.1016/j.cell.2018.12.015
24. Stenton SL, O'Leary M, Lemire G. *et al.* Critical assessment of variant prioritization methods for rare disease diagnosis within the rare genomes project.
25. Lam WKJ, Lau CS, Luk HM. *et al.* The implementation of genome sequencing in rare genetic diseases diagnosis: a pilot study

from the Hong Kong genome project. *Lancet Reg Heal - West Pac* 2025;**55**:101473. https://doi.org/10.1016/j.lanwpc.2025.101473
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324
27. McKenna A, Hanna M, Banks E. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303. https://doi.org/10.1101/gr.107524.110
28. Smedley D, Jacobsen JOB, Jäger M. *et al.* Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat Protoc* 2015;**10**:2004–15. https://doi.org/10.1038/nprot.2015.124
29. Martin AR, Williams E, Foulger RE. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet* 2019;**51**:1560–5. https://doi.org/10.1038/s41588-019-0528-2
30. Robertson AJ, Tran K, Patel C. *et al.* Evolution of virtual gene panels over time and implications for genomic data reanalysis. *Genet Med Open* 2023;**1**:100820. https://doi.org/10.1016/j.gimo.2023.100820
31. Jacobsen J, Smedley D, Robinson P. *Exomiser Documentation Release 14.0.0*. https://exomiser.readthedocs.io/_/downloads/en/14.0.0/pdf/ (5 May 2025, date last accessed).
32. Pejaver V, Byrne AB, Feng B-J. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet* 2022;**109**:2163–77. https://doi.org/10.1016/j.ajhg.2022.10.013
33. Johnston JJ, Dirksen RT, Girard T. *et al.* Variant curation expert panel recommendations for RYR1 pathogenicity classifications in malignant hyperthermia susceptibility. *Genet Med* 2021;**23**:1288–95. https://doi.org/10.1038/s41436-021-01125-w
34. Parsons MT, de la Hoya M, Richardson ME. *et al.* Evidence-based recommendations for gene-specific ACMG/AMP variant classification from the ClinGen ENIGMA BRCA1 and BRCA2 variant curation expert panel. *Am J Hum Genet* 2024;**111**:2044–58. https://doi.org/10.1016/j.ajhg.2024.07.013
35. Ha C, Kim J-W, Jang J-H. Performance evaluation of SpliceAI for the prediction of splicing of NF1 variants. *Genes* 2021;**12**:1308. https://doi.org/10.3390/genes12091308
36. Smith MJ, Walker JA, Shen Y. *et al.* Expression of SMARCB1 (INI1) mutations in familial schwannomatosis. *Hum Mol Genet* 2012;**21**:5239–45. https://doi.org/10.1093/hmg/dds370
37. Hudder A, Werner R. Analysis of a Charcot-Marie-tooth disease mutation reveals an essential internal ribosome entry site element in the Connexin-32 gene*. *J Biol Chem* 2000;**275**:34586–91. https://doi.org/10.1074/jbc.M005199200
38. Tsai P-C, Chen C-H, Liu A-B. *et al.* Mutational analysis of the 5′ non-coding region of GJB1 in a Taiwanese cohort with Charcot–Marie–Tooth neuropathy. *J Neurol Sci* 2013;**332**:51–5. https://doi.org/10.1016/j.jns.2013.06.011
39. Schubach M, Nazaretyan L, Kircher M. The regulatory Mendelian mutation score for GRCh38. *GigaScience* 2023;**12**:giad024.
40. Ceulemans B, Boel M, Claes L. *et al.* Severe myoclonic epilepsy in infancy: toward an optimal treatment. *J Child Neurol* 2004;**19**:516–21. https://doi.org/10.1177/08830738040190070701
41. Vitsios D, Dhindsa RS, Middleton L. *et al.* Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat Commun* 2021;**12**:1504. https://doi.org/10.1038/s41467-021-21790-4
42. Tabarini N, Biagi E, Uva P. *et al.* Exploration of tools for the interpretation of human non-coding variants. *Int J Mol Sci* 2022;**23**:12977. https://doi.org/10.3390/ijms232112977
43. Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol* 2019;**20**:32. https://doi.org/10.1186/s13059-019-1634-2
44. Jacobsen JOB, Kelly C, Cipriani V. *et al.* Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease. *Hum Mutat* 2022;**43**:1071–81. https://doi.org/10.1002/humu.24380
45. Amlie-Wolf A, Tang M, Mlynarski EE. *et al.* INFERNO: inferring the molecular mechanisms of noncoding genetic variants. *Nucleic Acids Res* 2018;**46**:8740–53. https://doi.org/10.1093/nar/gky686
46. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015;**12**:931–4. https://doi.org/10.1038/nmeth.3547
47. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2014;**31**:761–3. https://doi.org/10.1093/bioinformatics/btu703
48. Ellingford JM, Ahn JW, Bagnall RD. *et al.* Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med* 2022;**14**:73. https://doi.org/10.1186/s13073-022-01073-3
49. Walker LC, de Hoya M, Wiggins GAR. *et al.* Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: recommendations from the ClinGen SVI splicing subgroup. *Am J Hum Genet* 2023;**110**:1046–67. https://doi.org/10.1016/j.ajhg.2023.06.002
50. Cooperstein IB, Marwaha VOP, Ward A. *et al. An Optimized Variant Prioritization Process for Rare Disease Diagnostics: Recommendations for Exomiser and Genomiser* 2025. medRxiv. https://doi.org/10.1101/2025.04.18.25326061
51. Qian X, Wang J, Wang M. *et al.* Identification of deep-intronic splice mutations in a large cohort of patients with inherited retinal diseases. *Front Genet* 2021;**12**:647400. https://doi.org/10.3389/fgene.2021.647400
52. Weisschuh N, Mazzola P, Zuleger T. *et al.* Diagnostic genome sequencing improves diagnostic yield: a prospective single-Centre study in 1000 patients with inherited eye diseases. *J Méd Genet* 2024;**61**:186–95. https://doi.org/10.1136/jmg-2023-109470
53. Lionel AC, Costain G, Monfared N. *et al.* Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med* 2018;**20**:435–43. https://doi.org/10.1038/gim.2017.119
54. Laurie S, Steyaert W, de Boer E. *et al.* Genomic reanalysis of a pan-European rare-disease resource yields new diagnoses. *Nat Med* 2025;**31**:478–89. https://doi.org/10.1038/s41591-024-03420-w
55. Pagnamenta AT, Camps C, Giacopuzzi E. *et al.* Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases. *Genome Med* 2023;**15**:94. https://doi.org/10.1186/s13073-023-01240-0
56. Bullich G, Matalonga L, Pujadas M. *et al.* Systematic collaborative reanalysis of genomic data improves diagnostic yield in neurologic rare diseases. *J Mol Diagn* 2022;**24**:529–42. https://doi.org/10.1016/j.jmoldx.2022.02.003
57. Bagnall RD, Ingles J, Dinger ME. *et al.* Whole genome sequencing improves outcomes of genetic testing in patients with hypertrophic cardiomyopathy. *J Am Coll Cardiol* 2018;**72**:419–29. https://doi.org/10.1016/j.jacc.2018.04.078
58. Barbosa P, Savisaar R, Carmo-Fonseca M. *et al.* Computational prediction of human deep intronic variation. *GigaScience* 2023;**12**:giad085.https://doi.org/10.1093/gigascience/giad085

Briefings in Bioinformatics, 2025, **26(5)**, bbaf475
https://doi.org/10.1093/bib/bbaf475
Problem Solving Protocol