Computers in Biology and Medicine 185 (2025) 109560

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

UniVar: A variant interpretation platform enhancing rare disease diagnosis through robust filtering and unified analysis of SNV, INDEL, CNV and SV

Cherie C.Y. Au-Yeung [a], Yuen-Ting Cheung [a], Joshua Y.T. Cheng [a], Ken W.H. Ip [a], Sau-Dan Lee [a], Victor Y.T. Yang [a], Amy Y.T. Lau [a], Chit K.C. Lee [a], Peter K.H. Chong [a], King Wai Lau [a], Jurgen T.J. van Lunenburg [a], Damon F.D. Zheng [a], Brian H.M. Ho [a], Crystal Tik [a], Kingsley K.K. Ho [a], Ramesh Rajaby [a,b], Chun-Hang Au [a], Mullin H.C. Yu [a], Wing-Kin Sung [a,c,d,*]

[a] Hong Kong Genome Institute, Hong Kong Science Park, Shatin, Hong Kong, China
[b] Shibuya Laboratory, Division of Medical Data Informatics, Human Genome Center, University of Tokyo, Japan
[c] Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong, China
[d] Laboratory of Computational Genomics, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Variant interpretation
Variant prioritization
Genetic diagnosis
Rare diseases
SNV and INDEL
Copy number variants
Structural variants

ABSTRACT

Background: Interpreting the pathogenicity of genetic variants associated with rare diseases is a laborious and time-consuming endeavour. To streamline the diagnostic process and lighten the burden of variant interpretation, it is crucial to automate variant annotation and prioritization. Unfortunately, currently available variant interpretation tools lack a unified and comprehensive workflow that can collectively assess the clinical significance of these types of variants together: small nucleotide variants (SNVs), small insertions/deletions (INDELs), copy number variants (CNVs) and structural variants (SVs).

Results: The Unified Variant Interpretation Platform (UniVar) is a free web server tool that offers an automated and comprehensive workflow on annotation, filtering and prioritization for SNV, INDEL, CNV and SV collectively to identify disease-causing variants for rare diseases in one interface, ensuring accessibility for users even without programming expertise. To filter common CNVs/SVs, a diverse SV catalogue has been generated, that enables robust filtering of common SVs based on population allele frequency. Through benchmarking our SV catalogue, we showed that it is more complete and accurate than the state-of-the-art SV catalogues. Furthermore, to cope with those patients without detailed clinical information, we have developed a novel computational method that enables variant prioritization from gene panels. Our analysis shows that our approach could prioritize pathogenic variants as effective as using HPO terms assigned by clinicians, which adds value for cases without specific clinically assigned HPO terms. Lastly, through a practical case study of disease-causing compound heterozygous variants across SNV and SV, we demonstrated the uniqueness and effectiveness in variant interpretation of UniVar, edging over any existing interpretation tools.

Conclusions: UniVar is a unified and versatile platform that empowers researchers and clinicians to identify and interpret disease-causing variants in rare diseases efficiently through a single holistic interface and without a prerequisite for HPO terms. It is freely available without login and installation at https://univar.live/.

1. Background

Genomic data analysis plays a crucial role in identification of disease-causing variants in rare diseases, contributing to improved patient care and personalized treatment strategies [1]. Interpretation on the pathogenicity of genetic variants relies heavily on the use of professional judgement and evidence from literature search. To standardize the evaluation of variant pathogenicity, the American College of Molecular Genetics and Genomics (ACMG) and the Association of Molecular Pathologists (AMP) jointly published guidelines for the assessment of variants in genes associated with Mendelian disorders [2]. However, it is a very tedious and time-consuming task to manually

* Corresponding author. Hong Kong Genome Institute, Hong Kong Science Park, Shatin, Hong Kong, China.
E-mail address: kwksung@cuhk.edu.hk (W.-K. Sung).

https://doi.org/10.1016/j.compbiomed.2024.109560
Received 20 August 2024; Received in revised form 24 November 2024; Accepted 8 December 2024

interpret genetic variants from existing databases and the ever-growing literature. Variant annotation and prioritization thus need to be automated to narrow down candidate variants that require extensive interpretation. Lots of variation prioritization tools with user-friendly graphical user interface (GUI) were developed to help clinicians to identify disease-causing mutations. Nevertheless, we found that existing tools have three limitations.

First, among existing variation prioritization tools that do not require scripting or programming, they cannot prioritize single nucleotide variants (SNVs), small insertions/deletions (INDELs), copy number variants (CNVs) and structural variants (SVs) together. For example, for free accessible web-based tools, VarFish [3] and MutationDistiller [4] focus on the analysis of SNV and INDEL whereas AnnotSV [5] and CNVxplorer [6] focus on CNV/SV only. Note that many single-gene events combining an SV and SNV/INDEL in *trans* have been reported to cause autosomal recessive (AR) diseases [7–9]. However, existing variant interpretation tools lack a unified and comprehensive workflow for interpreting the clinical significance of all types of variants (SNV, INDEL, CNV and SV) in a holistic manner [10]. Technically, as SV also includes CNV like deletion/duplication, the term 'SV' in this paper encompasses both SV and CNV.

Second, filtering variants with high allele frequencies (AFs) in the healthy population [11] has been proved effective in excluding common variants, especially when combined with various gene annotations and the inheritance patterns of Mendelian disorders [12]. However, when it comes to SV, there is a scarcity of publicly accessible SV catalogues suitable for variant filtering. The current state-of-the-art SV catalogues are Genome Aggregation Database (gnomAD) [13] and the 1000 Genome Project (1KGP) [14]. As showed in the result section, they missed many SVs present in the general population. As a result, filtering SVs based on AF is less effective and more challenging compared to filtering SNVs and INDELs.

Third, most common variant prioritization tools, such as Exomiser [15], DeepPVP [16] and LIRICAL [17], adopted a phenotypic-driven approach to rank variants based on their association with the observed phenotype. These tools heavily depend on the Human Phenotype Ontology (HPO) [18] as the primary source of phenotypic information. However, assigning HPO terms to one patient requires clinical expertise and judgment. On average, it takes approximately 15 minutes for a clinician to extract HPO terms from clinical notes manually [19]. Also, the specificity and quantity of HPO terms used in prioritization can affect the ranking of the disease-causing variant [20]. Hence, assigning an optimal set of HPO terms for a patient is not an easy task.

To address the above limitations and difficulties, we have developed the Unified Variant Interpretation Platform (UniVar), a free web server tool that offers an automated and comprehensive workflow on annotation, filtering and prioritization for SNV, INDEL and SV collectively to identify disease-causing variants for rare diseases in one interface, which is not available in any existing variant interpretation tools. It is particularly useful for identifying disease-causing compound heterozygous variants involving both SNV/INDEL and SV in whole exome sequencing (WES) or whole genome sequencing (WGS) datasets. In addition, we have generated and built in a diverse SV catalogue of the global population that is more complete and accurate than the state-of-the-art SV catalogues, thus allowing users to filter out more common SVs. Furthermore, we have developed a novel computational method for deriving representative HPO terms from gene panels, enabling users to select gene panels for variant prioritization instead of inputting specific HPO terms. Such functionality is indispensable for cases without specific clinically assigned HPO terms, noting most other existing tools must require input of HPO terms. UniVar is freely available without login and installation at https://univar.live/. In summary, the aim of the study is to develop and present UniVar, which automates and streamlines the interpretation of genetic variants associated with rare diseases. The platform seeks to provide a comprehensive and user-friendly workflow for analyzing various types of variants (SNVs, INDELs, CNVs, and SVs) in one interface, improving accessibility and efficiency for researchers and clinicians. Additionally, it aims to enhance variant filtering accuracy and prioritize variants effectively, even without detailed clinical information, with an ultimate aim to increasing the diagnostic yield.

2. Materials and methods

2.1. Workflow of UniVar framework

Fig. 1 illustrates the comprehensive workflow of UniVar framework, which consists of four steps. The user journey begins in Step 1 with uploading (1) the lists of variants in variant call format (VCF) for the proband and, optionally, their family members; (2) the familial relationships among the samples (using the web form or a pedigree [PED] file) and (3) phenotypic information (if available). For each sample, the list of variants is represented by one VCF file for SNVs/INDELs and multiple VCF files for SVs. We allow multiple VCF files for SVs since different SV types are called by different SV callers, such as SurVIndel2 [21], INSurVeyor [22], Manta [23], etc. Two forms of phenotypic information are accepted when prioritization is necessary: HPO terms [18] and gene panels. The gene panels are retrieved from authoritative sources, including ClinGen [24], Genomics England PanelApp [25] and PanelApp Australia [25].

After uploading all mandatory files, our annotation workflow in Step 2 will be initiated. The annotation process for SNVs/INDELs and CNVs/SVs are executed with Ensembl Variant Effect Predictor [26] and Nirvana [27] respectively. The overlapping criteria between SVs and annotations are set to a reciprocal overlap of 0.9 by default. In addition, we adapted the classification of loss-of-function curations in gnomAD [13] and generated the predicted loss-of-function (pLoF) for SVs. The pLoF included loss of function, intragenic exonic duplication, whole-gene copy gain and whole-gene inversion.

In Step 3, our prioritization workflow begins by allowing the user to input phenotypic information, which can be provided either in the form of HPO terms or gene panels. Then, the phenotypic information is used to prioritize the variants with the help of the recent release of Exomiser (v13.3.0) [15]. In the case of utilizing gene panels for phenotypic information, as demonstrated in the subsequent section, our platform employs a novel method to convert the gene panel into HPO terms and prioritize the variants with Exomiser.

Lastly, once all the annotation and prioritization processes have been completed, the output is visualized in a web browser. To safeguard user data privacy and security, the link to the results is provided in the format of a long randomly generated text string, which is exclusively disclosed to the user.

The workflow of UniVar is implemented using Python, JavaScript and TypeScript. These source codes are available in https://github.com/kensung-lab/UniVar.

2.2. Data sources for annotation

A wide variety of annotations are integrated to facilitate the identification of disease-causing variants, which include gene, gene intolerance, known pathogenic/benign or reported variants, pathogenicity predictors, population AF, etc. All the annotation sources and their corresponding versions used in our current workflow are detailed in Appendix file 1: Table S1.

2.3. Inhouse SV catalogue of the global population

An accurate catalogue of SVs in healthy individuals is crucial for filtering common CNVs/SVs. Current prioritization tools rely on the state-of-the-art SV catalogues provided by gnomAD and/or 1KGP. However, as demonstrated in the results section, both gnomAD and 1KGP fail to capture numerous normal SVs, thereby impacting the

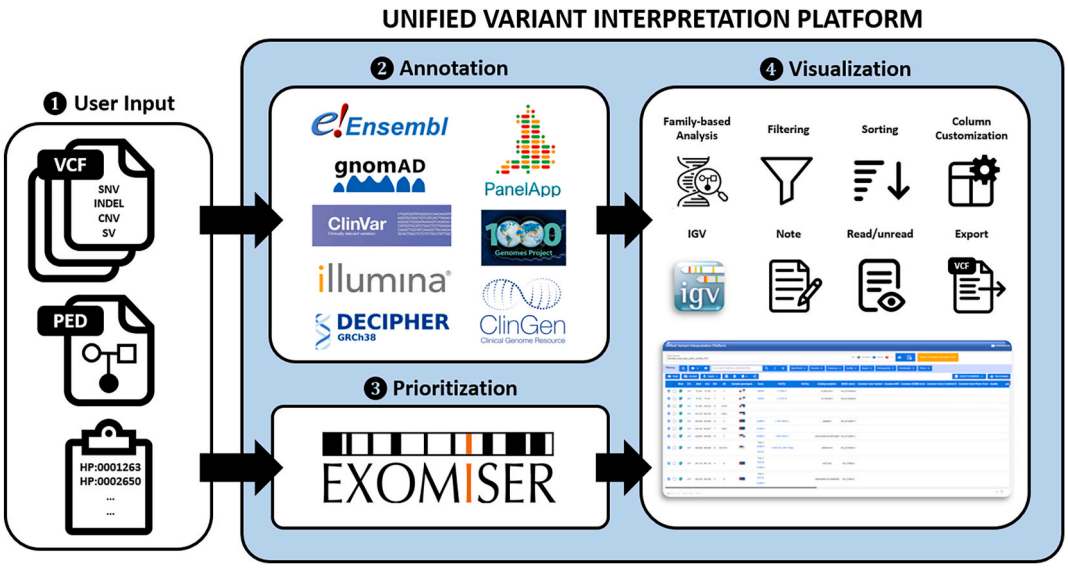

Fig. 1. A graphical illustration of UniVar's workflow. The UniVar workflow consists of four steps. Step 1: User input of SNV/INDEL/CNV/SV in VCF files, a PED file and HPO terms; Step 2: Annotation; Step 3: Prioritization; Step 4: Visualization, where a GUI web browser with a range of interactive functionalities is generated for variant interpretation.

effectiveness of SV filtering.

To enhance the existing SV annotation, we have developed an inhouse SV catalogue encompassing the global population. We chose two callers SurVIndel2 [21] and INSurVeyor [22] to detect SVs in all 2504 unrelated samples from the 1KGP [14]. The called SVs are jointly genotyped using SurVClusterer [28] and SurVTyper [29]. Subsequently, the allele count and AF for the global population and the five super-populations (African, Admixed American, East Asian, European [EUR], and South Asian) are computed.

2.4. Filtering parameters of UniVar

To narrow down the variant search, there are eight types of customizable filtering parameters in the UniVar interface which are applied to both SNVs/INDELs and CNVs/SVs. They are listed as follows: (1) In the genomic location, users can search by a genomic coordinate or by a certain gene. (2) For gene panel, users can search by genes in the associated gene panels from ClinGen [24], Genomics England PanelApp [25] and PanelApp Australia [25]. (3) The scenario sections can be filtered by the mode of inheritance based on the samples genotype, which include dominant, recessive, de novo, compound heterozygous, X-linked and any of the scenarios mentioned. (4) Variants can be filtered based on the three population frequency databases: 1KGP, gnomAD and our inhouse SV catalogue. (5) Users can filter variants by genotype quality or score in the quality section. (6) The impact section allows users to filter by the consequence of the variants, and (7) the pathogenicity section allows users to filter by the pathogenicity predictors such as the CADD and REVEL score. (8) If Exomiser is executed, users can filter by the Exomiser score in the prioritization section.

2.5. Output of variants

Upon completion of the UniVar workflow, the output is visualized directly in our web browser. UniVar displays all variants (SNVs, INDELs and SVs) in a tabular view. Each variant is displayed on a separate row, with its corresponding annotations listed across multiple columns. Users can also download the output to tab separated values (TSV) or VCF files for later use.

2.6. A novel computational method to derive the most representative HPO terms from gene panels

A gene panel is a specific set of genes associated with a particular disease or phenotype. The list of gene panels curated from ClinGen [24], Genomics England PanelApp [25] and PanelApp Australia [25] were downloaded. For genes in Genomics England PanelApp and PanelApp Australia, only green and amber genes were considered in the analysis.

For each gene panel, we developed an approach to select the five most representative HPO terms to represent it according to the following few steps.

First, the HPO terms associated with the genes in a gene panel were quantified. A score was assigned to each relevant HPO term through multiplying its information content (IC) by the number of genes associated with it. Here, we defined the relative IC of a term based on its frequency of annotated genes. The IC of an HPO term t is given by

$$IC(t) = -\log p_t,$$

where p_t is the frequency among annotation to all annotated genes. To ensure the exclusion of general terms that lack specificity, such as All (HP:0000001) and Phenotypic abnormality (HP:0000118), HPO terms that have an IC less than one and also less than three relative ancestor terms were excluded.

Second, we proceeded to select the five HPO terms with the highest scores within a gene panel. However, we set a constraint to prevent the chosen representative HPO terms from being excessively similar to one another. The semantic similarity is measured by the graphic-based approach proposed by Pesquita et at. [30], is defined as

$$sim(T_1, T_2) = \frac{\sum_{t \in T_1 \bigcap T_2} IC(t)}{\sum_{t \in T_1 \bigcup T_2} IC(t)}$$

Last, we only retained the HPO term with a higher score when the similarity between the two HPO terms was 0.8 or more.

2.7. Simulated a set of variants for diagnosed patients for benchmarking

Benchmarking was performed using simulated WGS datasets derived from 134 inherited retinal disease (IRD) patients who had confirmed clinical diagnosis and known disease-causing variants [31]. There are 160 unique disease-causing variants in 60 genes within the 134 IRD

patients. The loci of these disease-causing variants were converted from GRCh37 to GRCh38. As 13 disease-causing variants could not be converted, we therefore excluded 29 IRD patients who carry one of these variants. Finally, to simulate the VCF file for each of the included 105 IRD patients, we added the disease-causing variants of each patient to the VCF file of the sample HG00096 from 1KGP [14].

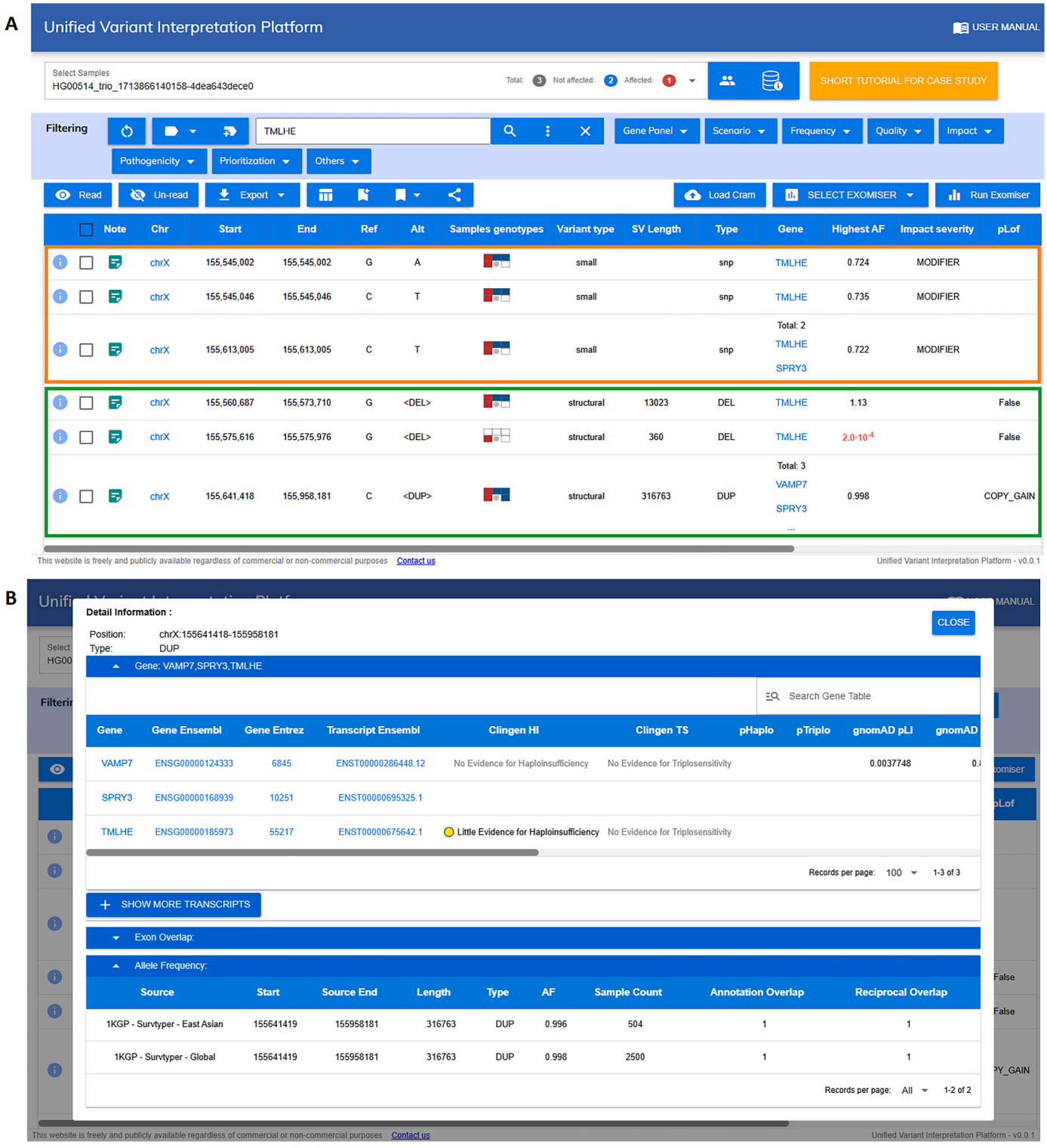

Fig. 2. Organization of UniVar's data visualization. UniVar's data visualization is organized into two layers of presentation framework. The SNVs and SVs presented here are variants filtered by *TMLHE* within a trio family. **A** The first layer is a tabular view that displays all user-supplied variants in a structured table format. The box with an orange border highlights three SNVs, while the green box highlights three SVs. From the columns, we can view the basic annotated information for these variants. **B** The second layer is a pop-up interface that contains detailed information compressed within the first layer. It shows a duplication overlapping with three genes (*VAMP7*, *SPRY3* and *TMLHE*) and the AF from our SV catalogue of 1KGP, where the global AF is 0.998 and the East Asian AF is 0.996.

3. Results

3.1. Highlights of the unique functionality of UniVar's web interface

UniVar is a unified web interface specifically designed for variant interpretation on WGS data of patients with rare diseases. One of the main features is that users can filter and prioritize SNV, INDEL and SV collectively to identify disease variants in one interface. Hence, this produces a synergy to jointly uncover compound heterozygous variants across SNV/INDEL and SV under one roof (Appendix file 2: Fig. S1), offering a unique functionality not found in any existing variant interpretation tools. Furthermore, our platform offers predefined filter parameters to help users to filter low risk disease-causing SNVs/INDELs and SVs. With just a single click, users can reduce the number of candidate variants significantly. Among them, our high-impact variant filter preset is the most effective and efficient filter and its filtering criteria are detailed in Appendix file 3. Additionally, our platform incorporated an inhouse SV catalogue and implemented a robust AF filtering mechanism specifically designed to assist users to filter common SVs. The evaluation of the performance of this filter is described in the subsequent sections, together with a case study to demonstrate the effectiveness of this filter (Appendix file 2: Fig. S1). In addition, UniVar offers the capability to prioritize variants when phenotypes are described using gene panels instead of HPO terms. This feature proves to be particularly valuable in cases where the precise HPO terms are not available or known because most other existing tools must require the input of HPO terms.

Next, we detail how UniVar presents the variants. UniVar displays all variants (SNVs, INDELs and SVs) in a tabular view. Each variant is represented in a separated row, with its corresponding annotations presented across multiple columns. Through a drag and drop interface in the 'Manage columns' panel, it offers users a high degree of flexibility in selecting the annotations to be displayed. Fig. 2A illustrates the variants overlapping with *TMLHE* in a few selected columns. The top three variants are SNVs, while the bottom three variants are SVs. Unlike SNV and INDEL, SV may affect multiple exons or multiple genes. For example, the duplication related to Charcot-Marie-Tooth disease found in 17p11.2-p12 covers over 100 genes. It is not feasible to display many genes for each SV in the 'Gene' column. To maintain the clarity of data presentation, we limit a maximum of two genes being displayed for each SV. For SV that involves more than two genes, the remaining genes would be hidden in the table and displayed in an additional layer (Fig. 2B) instead.

For each SV, the additional layer comprises a pop-up interface specially designed to show all genes and their annotations that are covered by the SV. Fig. 2B presents an example SV that covers three genes. These annotations are organized in the form of accordions, which included the detailed information of genes, related variants, external sources, exon overlap, clinical interpretation, AF and Exomiser results. Furthermore, each gene is annotated with its clinical significant, such as the haploinsufficiency (HI) and triplosensitivity classifications from ClinGen's dosage sensitivity curations [24] and the predicted probabilities of dosage sensitivity (pHaplo & pTriplo) [32].

UniVar also offers an interface to run Exomiser (v13.3.0) [15] to prioritize variants. Exomiser is a phenotypic-driven approach of prioritization that can improve the likelihood of identifying disease-causing variants in patient samples. The recent release is capable of prioritizing SNV, INDEL and SV together, which can improve the discovery of disease-causing SV that are underrepresented. UniVar allows users to update phenotypic information (in terms of HPO terms) by executing the 'Run Exomiser' service at any time, regardless it is done before or after the variants have been uploaded. However, Exomiser must require the input of specific HPO terms. For patients that do not have any specific clinical features, users can select gene panels (curated from ClinGen, Genomics England PanelApp and PanelApp Australia) to represent the diseases or phenotypes of the patients. We developed a novel computation method of deriving the five most representative HPO terms from a gene panel (see Section 2.6), and results in the subsequent section showed that these computed HPO terms can effectively prioritize disease-causing variants. For a comprehensive understanding of each functionality, a detailed manual can be found in Appendix file 3.

3.2. Benchmark the population frequency of the inhouse SV catalogue

There are not many population-based SV catalogues in the literature. For SV filtering, people currently use the SV catalogue of 1KGP [14] (which contains SVs from 2,504 samples generated by New York Genome Center) and gnomAD [13] (which contains SVs from 14,891 samples maintained by Broad Institute). These two SV catalogues are respectively referred to as 1KGP-SV and gnomAD-SV below. As shown below, both 1KGP-SV and gnomAD-SV are neither accurate nor complete. To fill in the gap, we developed an inhouse SV catalogue, referred as Inhouse-SV. The method to create the Inhouse-SV has been detailed in Section 2.3.

To compare our Inhouse-SV with 1KGP-SV and gnomAD-SV, we used the SV catalogue constructed from long-read sequencing data obtained from 3,622 Icelanders [33], which is abbreviated as Icelander-SV. This Icelander-SV dataset is selected as the benchmark dataset since SV calling from long reads is known to have high sensitivity and specificity, hence generally being regarded as a gold standard. Among the 55,649 deletions in Icelander-SV, 31% of them were found to overlap with Inhouse-SV, whereas 18% and 16% overlapped with 1KGP-SV and gnomAD-SV respectively. As for the 70,206 insertions (which technically also consists of duplications) in Icelander-SV, 70% of which overlapped with Inhouse-SV, compared to 25% and 28% overlapped with 1KGP-SV and gnomAD-SV respectively. The above results are shown in Fig. 3, indicates that gnomAD-SV and 1KGP-SV miss many SVs in Icelander-SV. On the other hand, Inhouse-SV covers over one fold more deletions and over 1.7 folds more insertions than gnomAD-SV. This clearly indicates that our Inhouse-SV is more complete than gnomAD-SV and 1KGP-SV.

To evaluate the accuracy of the AFs of the SVs in Inhouse-SV, 1KGP-SV and gnomAD-SV, we computed the Pearson's correlation coefficient between the AFs of the SVs in Icelander-SV and the corresponding AFs in each of the three SV catalogues overall and across superpopulation subgroups (EUR, Admixed American, South Asian, East Asian, African and other) respectively. The results are shown in Table 1. The AFs of the SVs of EUR population consistently achieves the highest correlation among the superpopulations across all three SV catalogues for both deletions and insertions. This is expected due to the ancestral connections between Icelanders and the EUR population [34]. More importantly, the correlation of the AFs of SVs between Icelander-SV and EUR population in Inhouse-SV is the highest compared to 1KGP-SV and gnomAD-SV. The correlation of 0.983 for deletions and 0.893 for insertions. This indicates that the AF from Inhouse-SV is more accurate than 1KGP-SV and gnomAD-SV. In all superpopulations, the correlation of AF for insertions was consistently lower than that of deletions. This discrepancy was mainly due to the inherent challenges associated with accurately detecting SVs within tandem repeat regions in short-read sequencing. However, upon excluding insertions within tandem repeat regions, the correlation of EUR population significantly improved to 0.942 from 0.890 in Inhouse-SV, to 0.921 from 0.722 in 1KGP-SV and to 0.921 from 0.774 in gnomAD-SV. In summary, Inhouse-SV has a more complete and accurate SV catalogue than 1KGP-SV and gnomAD-SV.

3.3. Effectiveness of the inhouse SV catalogue databases in filtering ClinVar benign but keeping pathogenic SVs

This section shows the effectiveness of Inhouse-SV's population frequency in filtering common benign SVs; and in contrast not filtering pathogenic SVs. We first downloaded the dataset of reported SV classified as benign or likely benign from ClinVar [35], which consisted of 14,068 deletions, 12,556 duplications and 67 insertions. Due to the

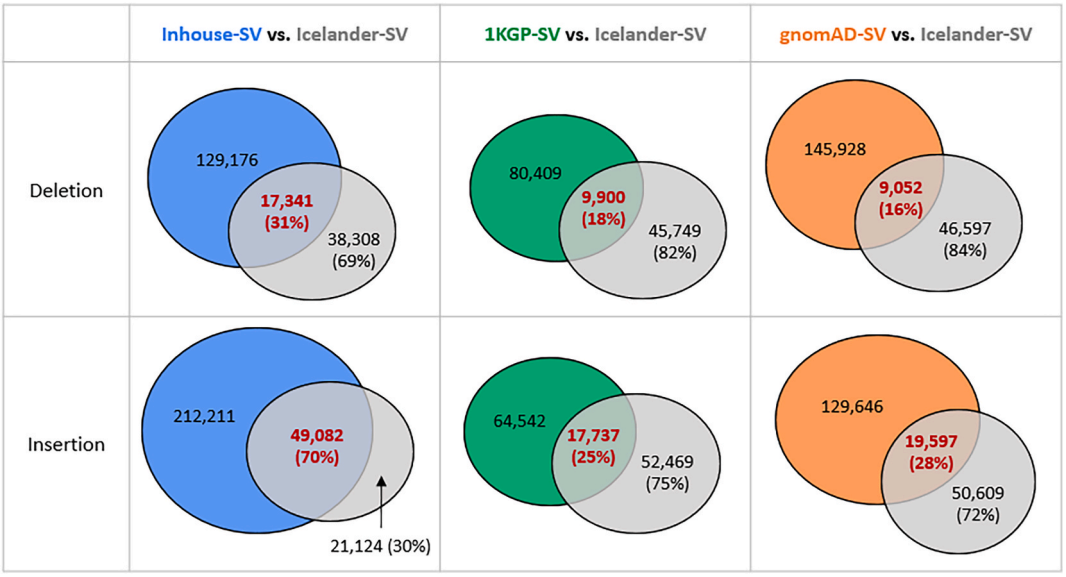

Fig. 3. Venn diagram of the number of overlaps in deletion and insertion between Icelander-SV and three SV catalogues. The proportion of overlapped and non-overlapped deletions and insertions in Icelander-SV across the three SV catalogues (Inhouse-SV, 1KGP-SV and gnomAD-SV) from short-read sequencing are presented in brackets.

Table 1
Correlation of AF between Icelander-SV and each of the three SV catalogues from short-read sequencing by superpopulation subgroup in terms of deletion and insertion.

		Global	EUR	AMR	SAS	EAS	AFR	OTH
Deletion	Inhouse-SV	**0.947**	**0.983**	**0.953**	**0.943**	**0.872**	**0.822**	–
	1KGP-SV	0.941	0.977	0.947	0.937	0.868	0.819	–
	gnomAD-SV	0.935	0.966	0.933	–	0.856	0.855	**0.946**
Insertion	Inhouse-SV	**0.846**	**0.890**	**0.852**	**0.840**	**0.763**	**0.698**	–
	1KGP-SV	0.677	0.722	0.688	0.680	0.611	0.547	–
	gnomAD-SV	0.732	0.774	0.735	–	0.649	0.636	**0.749**

Global: All genomes, EUR: European, AMR: Admixed American, SAS: South Asian, EAS: East Asian, AFR: African, OTH: Other.

underrepresented number of insertions in ClinVar, they were excluded from this analysis. We then assessed the effectiveness of each of the SV catalogues (Inhouse-SV, 1KGP-SV and gnomAD-SV) in filtering ClinVar benign deletions and duplications. Furthermore, we assessed the filtering effectiveness when these three SV catalogues were used together. For example, if we set the AF threshold to ≤0.01, any variant with an AF exceeding 0.01 in any of the three databases would be filtered out.

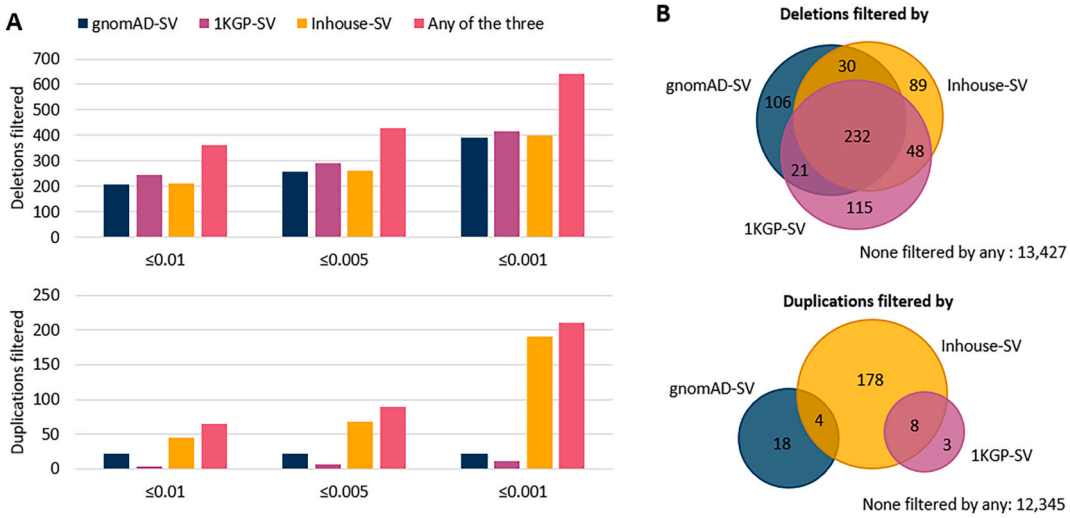

Fig. 4. The effectiveness of population frequency databases in filtering ClinVar benign SV. A The number of benign deletions and duplications filtered by applying AF thresholds of ≤0.01, ≤0.005, and ≤0.001 from each of the three SV catalogues and any of the three catalogues. **B** Venn diagram shows the overlaps in number of filtered deletions and duplications among the three SV catalogues at AF≤0.001.

Fig. 4A presents the number of benign deletions and duplications filtered by applying AF thresholds of ≤0.01, ≤0.005, and ≤0.001 from each of the three population frequency databases individually and in combination, whereas Fig. 4B provides the respective Venn diagrams showing the overlaps in number of filtered deletions and duplications among the three SV catalogues at AF≤0.001. Similar number of benign deletions were filtered by each of the three AF databases under all three AF thresholds. Moreover, when we combined all three SV catalogues at AF≤0.001, a unique total of 642 deletions were effectively filtered out, removing 54%–65% more benign deletions than when using any individual catalogue alone. On the other hand, Fig. 4A shows that the Inhouse-SV could effectively remove more benign duplications than the other two SV catalogues. As for both 1KGP-SV and gnomAD-SV, the number of filtered benign duplications was invariably small and showed little variation across three levels of AF threshold. Fig. 4B shows that 178 of them were exclusively filtered by using Inhouse-SV, as compared to 18 and 3 duplications exclusively by gnomAD-SV and 1KGP-SV respectively, whereas the remaining 12 by Inhouse-SV in tandem with either one of the two other databases.

Similarly, 14,533 deletions and 2,857 duplications classified as pathogenic or likely pathogenic, a review status with one star or above and no conflicting classification from ClinVar were retrieved. Since they are pathogenic, we assumed that most of these SVs should have very low AF and should not be filtered by any of the three AF databases. We assessed the number of ClinVar pathogenic SVs filtered at AF thresholds of ≥0.01, ≥0.005, and ≥0.001 from each of the three population frequency databases individually. Table 2 showed that among the three SV catalogues, Inhouse-SV filtered the least number of pathogenic SVs among the three SV catalogues. Only 9 SVs have an AF of ≥0.001 in Inhouse-SV, while there are 43 and 46 SVs with AF ≥0.001 in 1KGP-SV and gnomAD-SV respectively.

3.4. Benchmark the performance of variant prioritization based on annotated HPO terms derived from gene panels

On certain occasions, specific HPO terms corresponding to a patient are not well documented in clinical notes. In such cases, an alternative approach is to utilize a gene panel instead. Gene panels are typically used in targeted genetic testing tools in clinical and research settings. They consist of a curated set of genes that are known to be associated with specific disease conditions or phenotypes. For example, a gene panel associated with retinal disorders contains the list of disease genes that cause retinal disorder. However, the existing tools like Exomiser must require input of HPO terms to prioritize variants. To resolve this limitation, our proposed approach is firstly to identify five HPO terms that are most representative of each gene panel (see Section 2.6); and then using these representative HPO terms to run Exomiser to prioritize the variants.

One key question is whether our approach can prioritize the correct disease-causing variant. To benchmark our method, we obtained a list of 105 IRD patients [31] and their corresponding specific HPO terms ($HPO_{clinical}$) and disease-causing variants. The $HPO_{clinical}$ is highly accurate since it was chosen by three clinicians with expertise in IRD diagnosis. For each IRD patient, based on his/her $HPO_{clinical}$, we selected the Retinal disorders panel and/or Monogenic hearing loss panel from Genomic England PanelApp to be his/her gene panel. Then, we simulated a set of variant datasets seeded with the patient's corresponding disease-causing variants (see Section 2.7).

We compared the performance of variant prioritization between (1) utilizing the specific HPO terms ($HPO_{clinical}$) and (2) using the HPO terms derived from the corresponding gene panel (HPO_{panel}). HPO_{panel} for Retinal disorders and Monogenic hearing loss panel are listed in Table 3. To assess the performance of variant prioritization, we run Exomiser on these 105 sets of simulated variants seeded with disease-causing variants. Each IRD patient has been assigned to a set of $HPO_{clinical}$ and HPO_{panel}. We compared the ranking of disease-causing variants predicted by Exomiser between using $HPO_{clinical}$ and using HPO_{panel} as HPO terms. Our results show that the Exomiser ranking of using HPO_{panel} was comparable to that of $HPO_{clinical}$. We categorized the ranking results for the disease-causing variants into four mutually exclusive disease-causing ranking bins: 'Top', '2nd-5th', '6th-10th' and '>10th'. In Fig. 5A, the correct disease-causing variants were ranked the first for 75% of IRD patients when using HPO_{panel} as compared to 60% when using $HPO_{clinical}$. In addition, Fig. 5B shows that out of 105 samples' disease-causing variants, 94 samples were ranked within the top five by both methods.

The Exomiser ranking is determined by the combined score, which considers both variant and phenotype components. The variant component is computed based on the AF and predicted pathogenicity, while the phenotype component is based on the HPO terms. Hence, when running Exomiser using HPO_{panel} and $HPO_{clinical}$, the phenotype score of the same variant differs, while the variant score remains unchanged. As presented in Fig. 5C, the mean Exomiser phenotype score in each of the disease-causing ranking bins are higher when using HPO_{panel} than $HPO_{clinical}$. Hence the same pattern is observed for the mean combined score.

3.5. A case study on identification of disease-causing compound heterozygous variants to compare the capability of UniVar against other tools for genetic diagnosis

We utilized a case study of a Korean child with ataxia-telangiectasia (A-T) reported by Lee et al. [7]. A-T is a rare AR disorder characterized by progressive neurologic impairment related to multisystem abnormalities. The patient's clinical features were exhibited with ataxia, delayed cognitive and speech-language development, and oculomotor apraxia. Brain imaging also showed interval development of mild atrophy in the cerebellum. Here we investigated compound heterozygous pathogenic variants in *ATM* gene involving a SNV of c.742C>T (p.Arg248Ter) inherited from the father and a 31,460 bp deletion of exons 24–40 inherited from the mother identified by the authors. The WGS of this family trio is not available, therefore we used calls from another family in 1KGP (HG00514 as proband, HG00512 as father and HG00513 as mother) as background noises alongside the reported compound heterozygous pathogenic variants.

This dataset was used to compare the capability of UniVar against

Table 2
The number of ClinVar pathogenic SVs filtered at three different AF thresholds from each of the three SV catalogues.

	≥0.01	≥0.005	≥0.001
Inhouse-SV	0	1	9
1KGP-SV	1	3	43
gnomAD-SV	1	1	46

Table 3
The derived HPO terms from gene panels.

Gene panel	Derived HPO terms
Retinal disorders (4.12)	Abnormality of retinal pigmentation (HP:0007703), Abnormal electroretinogram (HP:0000512), Nyctalopia (HP:0000662), Reduced visual acuity (HP:0007663), Photophobia (HP:0000613)
Monogenic hearing loss (4.9)	Sensorineural hearing impairment (HP:0000407), Functional abnormality of the inner ear (HP:0011389), Hearing impairment (HP:0000365), Abnormality of the inner ear (HP:0000359), Congenital onset (HP:0003577)
Ataxia and cerebellar anomalies – narrow panel (4.13)	Cerebellar atrophy (HP:0001272), Ataxia (HP:0001251), Dysarthria (HP:0001260), Abnormal hindbrain morphology (HP:0011282), Gait ataxia (HP:0002066)

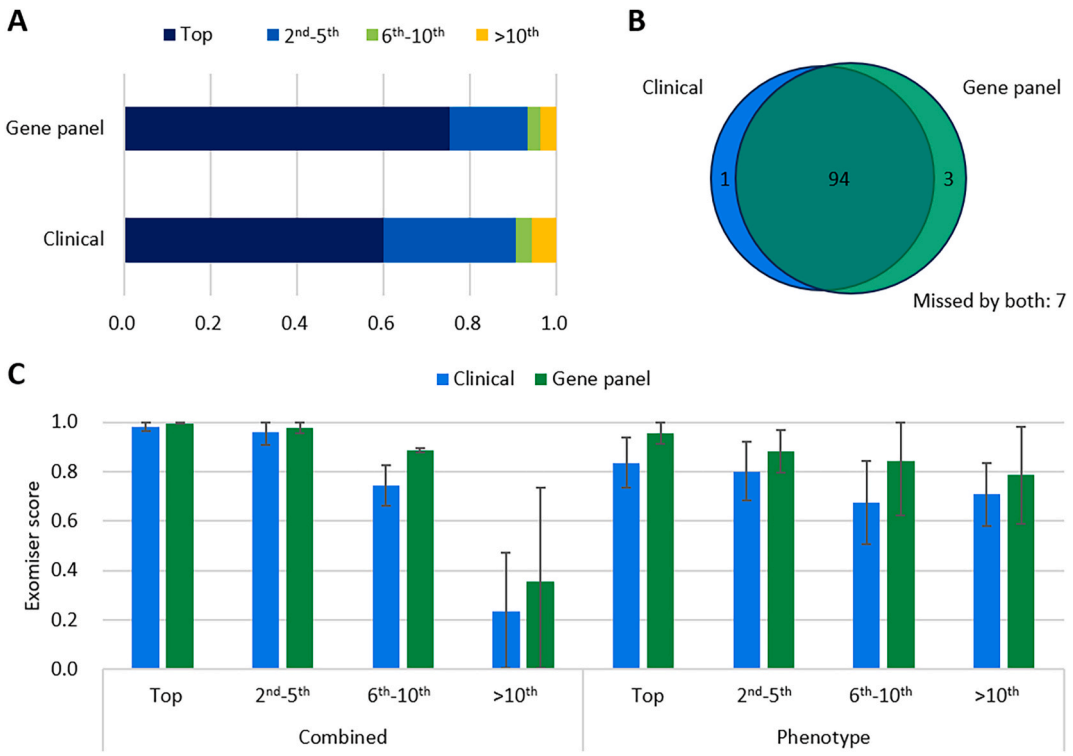

Fig. 5. Exomiser performance of using HPO terms annotated from different methods on simulated sets of variants. A The categorical percentage distribution of the disease-causing ranking according to four mutually exclusive disease-causing ranking bins under each of the two HPO annotated methods. **B** Venn diagram of the number of disease-causing variants prioritized within the top five by the two HPO annotated methods. **C** The mean Exomiser phenotype score and combined score between the two annotation methods on HPO terms in the four disease-causing ranking bins. The error bars represent the standard deviation of the Exomiser phenotype score and combined score.

other established tools. Each of the tool's capability is evaluated based on the web server's processing power, evidence presented for classifying pathogenicity in accordance with the ACMG guidelines [2,36], as well as the efficiency in identifying disease-causing variants reported in the case study, so to aid users in establishing a genetic diagnosis of A-T. We confined our comparison against those freely accessible web-based platforms which do not require installation and offer automatic annotation pipeline and a GUI designed for variant interpretation on rare diseases. We surfed for suitable tools, and selected AnnotSV [5], CNVxplorer [6] and VarFish [3] because these three tools have been most favourably compared against other tools.

The simulated SNV/INDEL, SV and PED files were uploaded to UniVar (v0.0.1) successfully. We first used the 'Select Gene Panel' approach from 'Run Exomiser' service to prioritize causal variants. '(uk) Ataxia and cerebellar anomalies – narrow panel 4.13' was chosen to represent the patient's phenotype. The five HPO terms representing this gene panel are listed in Table 3, and we defined these five HPO terms as HPO_{ataxia}. Next, the 'High risk (SNV/INDEL + SV)' preset filter was applied to narrow down the number of plausible causal variants, resulting in a total of 13 remaining variants. Lastly, the 'Exomiser Gene Combined Score' column was selected for sorting in descending order. The compound heterozygous SNV and SV were sorted at the top of the list with an Exomiser combined score of 0.991 as shown in Appendix file 2: Fig. S1 (Exomiser phenotype score is 0.856 and variant score is 0.993). The inheritance pattern of this trio-family is clearly shown in the 'Sample genotypes' column. In the genotype pop-up shown in Appendix file 2: Fig. S2, it is evident that the proband's heterozygous SNV was inherited from the father, while the heterozygous SV was inherited from the mother, indicating a compound heterozygous pattern. Exomiser's prediction indicated a high similarity of phenotype with A-T in a MOI of AR. We started by evaluating the clinical validity of gene-disease associations. The ClinGen website was assessed by clicking on the hyperlink provided in the pop-up for the *ATM* gene (Appendix file 2: Fig. S3). We found that ClinGen has suggested definite classification between *ATM* and A-T.

Next, we interpreted the SNV of *ATM* gene. The variant table showed that this is a stop gained variant (see the first row in Appendix file 2: Fig. S1). To confirm that this variant is not located at the last exon, we opened the IGV to check its position as shown in Appendix file 2: Fig. S4. The variant is in exon 7 of 63, therefore we can predict it to undergo nonsense-mediated decay. Further investigation of the 'Gene' section in second layer confirms that this variant is located in a biologically-relevant transcript, as denoted by the 'Is MANE Select' flag being marked as true (Appendix file 2: Fig. S5). Also, this variant presents in controls with very low frequency of 0.000008 in gnomAD v2 (0.000054 in East Asian) and absent in gnomAD v3 (Appendix file 2: Fig. S1). The evidence presented here was sufficient to classify this variant as likely pathogenic (PSV1, PM2_Supporting). To upgrade the variant as pathogenic, we can further apply PM3 by referring to the reported variants in ClinVar. The variant table showed that it is classified as a pathogenic/likely pathogenic variant in ClinVar. The ClinVar ID also includes a hyperlink to assess the corresponding variant in ClinVar.

Finally, we investigated the SV of *ATM* gene. The pLof is LOF (Appendix file 2: Fig. S1), therefore we investigated how the deletion is overlapping with *ATM*. In the second layer, the 'Clingen HI' column in 'Gene' section indicated that *ATM* is a gene with sufficient evidence for HI, thus we can refer it as an HI gene. We limited our search to a MANE select transcript (ENST00000675843) in the 'Exon Overlap' section (Appendix file 2: Fig. S6), it showed that this deletion spans from exon 24 to 40 thus confirmed that both breakpoints are within an established HI gene (Category 1A, 0 points; Category 2E, 0.9 points). Also, the SV did not overlap with any of the population database. Therefore, the evidence presented is enough to classify this SV as likely pathogenic (PSV1, PM2_Supporting). All in all, UniVar demonstrated a strong

capability that effectively aids users in establishing a genetic diagnosis of A-T.

AnnotSV (v3.4) is a tool that specializes in the analysis, interpretation, and prioritization of SV. We uploaded the simulated SV and PED file to AnnotSV. In addition, we enabled the compound heterozygosity analysis option by uploading the simulated SNV/INDEL dataset and the phenotype-drive analysis options by inputting the derived HPO terms HPO_{ataxia}. However, we needed to include a rank filtering of '3–5' (1 indicating benign and 5 indicating pathogenic). It is because without setting this filter, a page loading error would occur. As a result, the browser for the html visualization could only load 751 SVs of a total of 4,500 SVs being annotated. The disease-causing SV was ranked top without any further settings as shown in Appendix file 2: Fig. S7. The Exomiser phenotype score was 0.856, which matched the score obtained from UniVar. The genotypes of the SV clearly showed that it is inherited from the mother (HG00513). We switched the display mode to a 'single SV focus' (Appendix file 2: Fig. S8). The tooltip of *ATM* gene indicated a HI score (=3) and there's a definitive association between *ATM* and A-T. The location indicated that the SV overlapped from intron 23 to 40 in the *ATM* gene. Also, the SV did not overlap with any of the common SVs in any of the population databases. With the evidence presented, we were able to establish a likely pathogenic classification to this SV. Though AnnotSV can identify the disease-causing SV correctly (a capability at par with UniVar), it does not report it is an AR disease nor provide the pathogenic SNV like UniVar. To investigate the pathogenic SNV, users would need to utilize another interpretation tool in order to establish a positive genetic diagnosis of A-T.

CNVxplorer (v0.4) is a web server tool that is designed for functional assessment of CNV in rare disease patients. Since it is a tool that specializes in CNV only, 1,770 SVs that are neither deletion nor duplication within our simulated SV dataset were filtered. As the tool supports GRCh37 coordinates only, we converted them into GRCh37 assembly. The conversion failed on 44 of them, thus only the remaining 2,686 SVs out of a total of 4,500 were uploaded to CNVxplorer. We run the tool by selecting deletion as CNVscore's prediction. First, we assessed the performance of the tool's phenotypic analysis by inputting the HPO terms of HPO_{ataxia} in the 'Phenotypic similarity' tab. Without filtering any MOI, the pathogenic *ATM* gene was ranked the fifth highest in terms of clinical similarity to the patient's phenotype (Appendix file 2: Fig. S9), whereas it ranked the second highest when filtered by AR inheritance. However, there are no indications on which CNV is overlapping with the *ATM* gene. To narrow down the scope, we selected only the disease-causing SV to run the analysis again. There are nine non-pathogenic CNV overlaps that are either with a low frequency or a non-deletion SV (Appendix file 2: Fig. S10). There is no indication on the evidence for established HI genes, patterns of inheritance and breakpoints on any protein-coding elements. CNVxplorer lacks the capability of UniVar and AnnotSV to interpret and identify the disease-causing SV, and it is unable to analyze any SNV/INDEL.

VarFish (v0.23) is a web application for quality control, filtering, prioritization and analysis on SNV/INDEL. Since the web server tool only supports VCF files in GRCh37, we converted the simulated SNVs/INDELs into the corresponding assembly. The conversion failed on 295 out of 114,518 SNVs/INDELs and the remaining were uploaded to VarFish along with the PED file. Once the upload was completed, we first adjusted the filter parameters with the following: (1) any inheritance (default), (2) dominant strict in frequency (default), (3) AA change, splicing in impact (default) and (4) whole genome in chromosomes (default). We also enabled the phenotype-based prioritization and input in HPO terms of HPO_{ataxia}. Unfortunately, the disease-causing SNV in *ATM* was not prioritized within the top 200 variants. We then further restricted the filtering by adjusting the impact parameter to null variant. The disease-causing SNV was only ranked as the sixth highest, which was expected since there was a penalty in the incompleteness of a compound heterozygous inheritance (Appendix file 2: Fig. S11). There was an extensive list of hyperlinks provided to check the gene-disease associations, such as GeneCC and OMIM. The variant table clearly showed that this variant has a stop gained effect and a frequency lower than 0.00001. The evidence presented is enough to aid users in classifying the variant as likely pathogenic. In comparison with UniVar, VarFish not only performed less satisfactorily based on this case study's SNV/INDEL analysis result, it also lacks the capability to establish the genetic diagnosis of A-T and hence users need to resort to another tool to identify the pathogenic SV.

4. Discussion

Variant interpretation is widely recognized as the most time-consuming process in achieving a patient's genetic diagnosis due to the meticulous analysis and assessment of genomic variants to identify those with potential clinical significance [37]. In recognition of this challenge, considerable efforts have been dedicated to developing strategies and tools aimed at narrowing down the pool of candidate variants, thereby enhancing the efficiency of genetic diagnosis [38].

We presented UniVar as a unified and versatile platform for identification and interpretation of disease-causing variants for rare diseases that offers comprehensive automated annotation of SNV/INDEL and SV collectively in one interface, making it accessible to users who may not have any programming expertise. Many studies have consistently demonstrated the positive impact of integrating SNV/INDEL and SV analysis on the diagnosis yield in WES and WGS diagnostics [39–42]. This platform offers an opportunity to potentially identify causative compound heterozygous variants involving SNV/INDEL and SV, thereby enhancing the diagnostic capabilities for AR diseases. Through a practical case study of disease-causing compound heterozygous variants across SNV and SV, we demonstrated the unique utility and effectiveness in variant interpretation of UniVar as compared to three other well-known tools (AnnotSV [5], CNVxplorer [6] and VarFish [3]), showcasing its capabilities that are not available in any existing variant interpretation tools.

However, the integration of SNV/INDEL and SV also leads to an increased number of variants that need to be curated and interpreted. Therefore, effective utilization of filtering strategies becomes crucial in narrowing down candidate variants from WES or WGS data, as it streamlines the diagnostic process and alleviates the burden of variant interpretation. The ACMG has published best practice guidelines for filtering out common SNV/INDEL, such as filtering by predicted LOF, MOI and AF [43]. Studies have demonstrated that the implementation of effective combined filtering strategies can significantly reduce the number of candidate SNV/INDEL, leading to a more precise analysis [12,44]. In contrast, filtering SV is a more intricate and demanding task due to the innate challenges in detecting SV using short-read technologies. Till now, there is a limited understanding of their population frequency which hinders the assessment of their severity and impact. Although large-scale studies such as 1KGP and gnomAD have contributed to expanding our comprehension and knowledge of SV, these studies employed SV detection tools such as DELLY [45], Manta [23] and MELT [46] which are not sensitive enough. This limited the effectiveness of the AF filtering on SVs when using these databases. In a more recent development, novel tools such as INSurVeyor [22] and SurvIndel2 [21] have emerged, showcasing superior performance when compared to these state-of-the-art callers. We therefore generated an SV catalogue of the global population using these two tools. Although this SV catalogue is more complete, it is noteworthy that certain SVs may still be missed since the database is merely created from genomes of 2,504 individuals. Therefore, to leverage the comprehensive knowledge provided by different tools, through the combination of population frequency databases in UniVar, it can bring about a more robust SV filtering approach.

The HPO is frequently used as the phenotypic information in variant prioritization tools. It is a standardized vocabulary for describing phenotypes, and many studies have shown its importance in capturing the

patient's condition(s) when analyzing and interpreting WGS data [16,20,31]. However, assigning an optimal set of HPO terms for a patient requires time and clinical expertise. Alternatively, the approach of using a gene panel is a simpler and easier method that saves time. Therefore, we developed a novel computational method for deriving representative HPO terms based on gene panels from most authoritative sources. Users can select one or multiple gene panels to initiate the platform's variant prioritization instead of inputting HPO terms. As most other existing tools must require the input of HPO terms, this functionality of UniVar is indeed indispensable for cases without clinically assigned HPO terms. More importantly, our results showed that the derived HPO terms output from UniVar can prioritize disease-causing variants as effective as specific clinically assigned HPO terms. It may even open an option to dispense with this laborious task for consideration under manpower constraint circumstances.

We introduced the 'High risk (SNV/INDEL + SV)' filter as a preconfigured filter parameter within UniVar, which is a combined filtering strategy on disease-causing SNV/INDEL and SV. This filtering approach has been implemented as an integral part of the Hong Kong Genome Project (HKGP), enhancing the efficiency and efficacy in identifying disease-causing variants. A diagnostic yield of 28% in identifying disease-causing variants was achieved, surpassing other comparable genome projects worldwide [47]. Among the disease-causing SNVs and SVs associated with HI genes within the positive cases, 80% of the disease-causing variants were ranked top one and 84% ranked the top three. However, not all the disease-causing variants can be captured by our high-impact variant filter preset. By relaxing the HI condition, a coverage of 97% of disease-causing variants were attained, where 63% of them were ranked the top one and 74% ranked the top three. These remaining 3% disease-causing variants correspond to those common low-penetrant variants listed in the BA1 exception list [48].

Our study has one major limitation, in which our current workflow does not support all variant types, in particular omitting mitochondrial variants and non-coding variants. Patients that carry disease-causing variants of these types cannot be identified with our tool. Other than this study limitation, it is indisputable that UniVar must require enhancement and updating over time. It is because the field of genomics is undergoing a profound and rapid evolution, leading to a transformative impact on research and medical domains [1]. Therefore, genomic analysis tools must continuously evolve to meet the latest professional standards and demands. To ensure up-to-date information, annotation sources have to be regularly reviewed and updated.

To address the study limitation and fast developments in the genomics field, our future plan focuses on further enhancing UniVar's overall capabilities to aid users in interpreting and identifying disease-causing variants, including the incorporation of the latest release of population frequency from gnomAD v4 and the analytics support towards non-coding variants. It is because interpreting the association between causal non-coding variants and their affected target genes is one of today's major challenges. Unlike coding variants, the pathogenicity mechanisms of non-coding variants are not well understood due to the complexity of regulatory mechanisms and the lack of functional annotations [49]. With a continued advancement in the understanding of non-coding and unexplored regions, the entire scientific community will have stronger capabilities to unveil a broader spectrum of genetic diagnosis in rare diseases.

5. Conclusions

UniVar is a unified and versatile platform for identification and interpretation of disease-causing variants that improve rare diseases diagnosis. It is a free web server tool that offers a comprehensive and secured workflow on annotation, filtering, and prioritization for SNV, INDEL, CNV and SV together. UniVar also provides a user-friendly GUI that consists of a range of interactive functionalities. By applying a predefined filter for high-impact variants, users can effortlessly uncover pathogenic compound heterozygous variants across SNV/INDEL and CNV/SV under one roof in a single click, which is not available in any existing variant interpretation tools. Furthermore, our diverse SV catalogue of the global population is more complete and accurate than the state-of-the-art SV catalogues, thereby enabling a robust AF filtering for common SVs. In comparison among three databases, our inhouse SV catalogue could filter most benign SVs as well as keep most pathogenic SVs. Lastly, when specific HPO terms for the patient are not available, users can opt to select the gene panel(s) instead. This is also a unique functionality not available in any other existing tools. A novel computational algorithm of UniVar will derive representative HPO terms from the selected gene panel(s), which are used for prioritization of disease-causing variants. This feature is particularly useful and time-saving in cases where the assigned HPO terms are not readily available or detailed clinical information is not available, having regard to our study results that this approach could prioritize disease-causing variants as effective as using HPO terms assigned by clinicians.

CRediT authorship contribution statement

Cherie CY. Au-Yeung: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Yuen-Ting Cheung:** Software, Data curation. **Joshua YT. Cheng:** Software. **Ken WH. Ip:** Software. **Sau-Dan Lee:** Software, Data curation. **Victor YT. Yang:** Software. **Amy YT. Lau:** Software, Data curation. **Chit KC. Lee:** Software. **Peter KH. Chong:** Software. **Nathan Lau:** Software. **Jurgen TJ. Lunenburg:** Software. **Damon FD. Zheng:** Software. **Brian Ho:** Project administration. **Crystal Tik:** Project administration. **Kingsley KK. Ho:** Project administration. **Ramesh Rajaby:** Writing – review & editing. **Chun-Hang Au:** Writing – review & editing. **Mullin HC. Yu:** Writing – review & editing, Supervision. **Wing-Kin Sung:** Writing – review & editing, Supervision, Conceptualization.

Availability of data and materials

Datasets generated and/or analyzed during the current study are publicly available in the URLs below.

ClinGen: https://www.clinicalgenome.org/

ClinVar: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/

gnomAD: https://gnomad.broadinstitute.org/

Genomics England PanelApp: https://panelapp.genomicsengland.co.uk/

PanelApp Australia: https://panelapp.agha.umccr.org/

HPO: https://hpo.jax.org/app/

SV AF from 1KGP: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20210124.SV_Illumina_Integration/1KGP_3202.gatksv_svtools_novelins.freeze_V3.wAF.vcf.gz

Clinical diagnosis and disease-causing variants in IRD patients: https://doi.org/10.3390/genes11040460

Phased SNV/INDEL/SV VCFs from 1KGP: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/

Long-read sequencing of SV from 3622 Icelanders: https://doi.org/10.1038/s41588-021-00865-4

Funding

This study received no external funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Hong Kong Genome Institute.

List of abbreviations

1KGP	1000 Genomes Project
ACMG	American College of Molecular Genetics and Genomics
AF	Allele frequency
AMP	Association of Molecular Pathologists
A-T	Ataxia-telangiectasia
AR	Autosomal recessive
CNV	Copy number variant
CRAM	Compressed reference-oriented alignment map
EUR	European
gnomAD	Genome Aggregation Database
GUI	Graphical user interface
HI	Haploinsufficiency
HKGI	Hong Kong Genome Project
HPO	Human Phenotype Ontology
IC	Information content
IGV	Integrative Genomics Viewer
INDEL	Small insertion/deletion
IRD	Inherited retinal disease
MANE	Matched Annotation from NCBI and EMBL-EBI
MOI	Mode of inheritance
OMIM	Online Mendelian Inheritance in Man Catalog
PED	Pedigree
pLoF	Predicted loss-of-function
SNV	Single nucleotide variation
SV	Structural variant
VCF	Variant call format
WES	Whole exome sequencing
WGS	Whole genome sequencing

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109560.

References

[1] J.J. McCarthy, H.L. McLeod, G.S. Ginsburg, Genomic medicine: a decade of successes, challenges, and opportunities, Sci. Transl. Med. 5 (2013).
[2] S. Richards, et al., Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of medical genetics and genomics and the association for molecular pathology, Genet. Med. 17 (2015) 405–423.
[3] M. Holtgrewe, et al., VarFish: comprehensive DNA variant analysis for diagnostics and research, Nucleic Acids Res. 48 (2020) W162–W169.
[4] D. Hombach, et al., MutationDistiller: user-driven identification of pathogenic DNA variants, Nucleic Acids Res. 47 (2019) W114–W120.
[5] V. Geoffroy, et al., AnnotSV: an integrated tool for structural variations annotation, Bioinformatics 34 (2018) 3572–3574.
[6] F. Requena, et al., CNVxplorer: a web tool to assist clinical interpretation of CNVs in rare disease patients, Nucleic Acids Res. 49 (2021) W93–W103.
[7] H.Y. Lee, et al., Compound heterozygous variants including a novel copy number variation in a child with atypical ataxia-telangiectasia: a case report, BMC Med. Genom. 14 (2021) 204.
[8] M.I. Alvarez-Mora, et al., Novel compound heterozygous mutation in TRAPPC9 gene: the relevance of whole genome sequencing, Genes 12 (2021) 557.
[9] M. Rodríguez-Hidalgo, et al., ABCA4 c.6480-35A>G, a novel branchpoint variant associated with Stargardt disease, Front. Genet. 14 (2023) 1234032.
[10] C.A. Austin-Tse, et al., Best practices for the interpretation and reporting of clinical whole genome sequencing, npj Genom. Med. 7 (2022) 1–13.
[11] K.J. Karczewski, et al., The mutational constraint spectrum quantified from variation in 141,456 humans, Nature 581 (2020) 434–443.
[12] B.S. Pedersen, et al., Effective variant filtering and expected candidate variant yield in studies of rare human disease, NPJ Genom Med 6 (2021) 60.
[13] R.L. Collins, et al., A structural variation reference for medical and population genetics, Nature 581 (2020) 444–451.
[14] M. Byrska-Bishop, et al., High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, Cell 185 (2022) 3426–3440.e19.
[15] D. Smedley, et al., Next-generation diagnostics and disease-gene discovery with the Exomiser, Nat. Protoc. 10 (2015) 2004–2015.
[16] I. Boudellioua, M. Kulmanov, P.N. Schofield, G.V. Gkoutos, R. Hoehndorf, DeepPVP: phenotype-based prioritization of causative variants using deep learning, BMC Bioinf. 20 (2019) 65.
[17] P.N. Robinson, et al., Interpretable clinical genomics with a likelihood ratio paradigm, Am. J. Hum. Genet. 107 (2020) 403–417.
[18] M.A. Gargano, et al., The Human Phenotype Ontology in 2024: phenotypes around the world, Nucleic Acids Res. 52 (2024) D1333–D1346.
[19] J.M. Havrilla, et al., PheNominal: an EHR-integrated web application for structured deep phenotyping at the point of care, BMC Med. Inf. Decis. Making 22 (2022) 198.
[20] D. Smedley, P.N. Robinson, Phenotype-driven strategies for exome prioritization of human Mendelian disease genes, Genome Med. 7 (2015) 81.
[21] R. Rajaby, WK. Sung, SurVIndel2: improving copy number variant calling from next-generation sequencing using hidden split reads, Nat Commun 15 (2024) 10473.
[22] R. Rajaby, et al., INSurVeyor: improving insertion calling from short read sequencing data, Nat. Commun. 14 (2023) 3243.
[23] X. Chen, et al., Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications, Bioinformatics 32 (2016) 1220–1222.
[24] H.L. Rehm, et al., ClinGen–the clinical genome resource, N. Engl. J. Med. 372 (2015) 2235–2242.
[25] A.R. Martin, et al., PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels, Nat. Genet. 51 (2019) 1560–1565.
[26] W. McLaren, et al., The Ensembl variant effect predictor, Genome Biol. 17 (2016) 122.
[27] M. Stromberg, et al., Nirvana: clinical grade variant annotator, in: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics* 596, Association for Computing Machinery, New York, NY, USA, 2017, https://doi.org/10.1145/3107411.3108204.
[28] Mesh89, Mesh89/SurVClusterer, 2023.
[29] kensung-lab/SurVTyper, KenSung-Lab, 2023.
[30] C. Pesquita, D. Faria, H. Bastos, A. Falco, F. Couto, Evaluating GO-based semantic similarity measures. Proceedings of 10th Annual Bio-Ontologies Meeting, 2007, pp. 37–39.
[31] V. Cipriani, et al., An improved phenotype-driven tool for rare mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data, Genes 11 (2020) 460.
[32] R.L. Collins, et al., A cross-disorder dosage sensitivity map of the human genome, Cell 185 (2022) 3041–3055.e25.
[33] D. Beyter, et al., Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits, Nat. Genet. 53 (2021) 779–786.
[34] A. Helgason, S. Sigurðardóttir, J.R. Gulcher, R. Ward, K. Stefánsson, mtDNA and the origin of the Icelanders: deciphering signals of recent population history, Am. J. Hum. Genet. 66 (2000) 999–1016.
[35] M.J. Landrum, et al., ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (2018) D1062–D1067.
[36] E.R. Riggs, et al., Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen), Genet. Med. 22 (2020) 245–257.
[37] A. Niroula, M. Vihinen, Variation interpretation predictors: principles, types, performance, and choice, Hum. Mutat. 37 (2016) 579–597.
[38] S. Pabinger, et al., A survey of tools for variant analysis of next-generation genome sequencing data, Briefings Bioinf. 15 (2014) 256–278.
[39] B. Yuan, et al., CNVs cause autosomal recessive genetic diseases with or without involvement of SNV/indels, Genet. Med. 22 (2020) 1633–1641.
[40] R. Pfundt, et al., Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders, Genet. Med. 19 (2017) 667–675.
[41] A. Lindstrand, et al., From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability, Genome Med. 11 (2019) 68.
[42] J. Kim, et al., KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses, Sci. Rep. 8 (2018) 5677.
[43] S.M. Harrison, L.G. Biesecker, H.L. Rehm, Overview of specifications to the ACMG/AMP variant interpretation guidelines, Curr Protoc Hum Genet 103 (2019) e93.
[44] M.A. Field, V. Cho, T.D. Andrews, C.C. Goodnow, Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies, PLoS One 10 (2015) e0143199.
[45] T. Rausch, et al., DELLY: structural variant discovery by integrated paired-end and split-read analysis, Bioinformatics 28 (2012) i333–i339.
[46] E.J. Gardner, et al., The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology, Genome Res. 27 (2017) 1916–1929.
[47] Legislative Council Panel on Health Services, Implementation of the Hong Kong Genome Project, 2024.
[48] A.M. Oza, et al., Expert specification of the ACMG/AMP variant interpretation guidelines for genetic hearing loss, Hum. Mutat. 39 (2018) 1593–1613.
[49] D.S. Paul, N. Soranzo, S. Beck, Functional interpretation of non-coding sequence variation: concepts and challenges, Bioessays 36 (2014) 191–199.