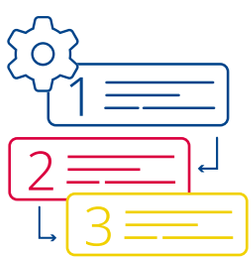
InheriNext

A fast AI-based variant prioritization system for NGS data
in the context of inherited disease analysis

Introduction

There are about 480 million people who suffered from inherited diseases. The incidence of inherited diseases is 6%. The number of newborn babies is about 120 million every year, indicating 7.2 million rare disease patients increased every year. A total of 6,172 known rare diseases have been identified and 85% of rare diseases are caused by altered genes.

The declining cost of next-generation sequencing brings genetic testing to a mainstream approach for rare disease diagnosis. However, the interpretation of genetic testing results is time-consuming, which takes about 3.5 hours for geneticists or doctors to study a case. InheriNext saves 70% of the time and extremely reduces the cost of such resources. InherNext will be your good partner in healthcare institutes and genetic test companies.



1 Leverage AI to inherited diseases

InheriNext is an AI-based variant prioritization system for NGS data (VCF or FASTQ files). InheriNext provides accurate candidates and diverse functionalities for inherited disease analysis in a few minutes and extremely reduces the burdens for geneticists. InheriNext brings two AI-based variant prioritization approaches. Users could use either HPO-based phenotypes or gene panels to obtain candidates for SNPs and INDELs (Figure 1A-D). A customized panel designer allows users to assemble frequently used phenotypes or genes into a customized panel (Figure 1E).

InheriNext dedicatedly designed a user-friendly interface. Users could quickly learn to use and interpret analysis results in a short time. The analysis parameters are shown at the top of the case page to bring a quick review of parameters (Figure 2A). The phenotype analysis results show in Figure 2B and provide several annotations, such as HGVS, conditions, phenotype network, ClinVar, automated ACMG and dbSNP. For the prediction scores and population databases, inheriNext provides up to 14 and 11 annotations, respectively (Figure 2C and D). The phenotype network shows the distance of phenotype similarity between sample phenotypes and disease-associated phenotypes. ClinVar plays a

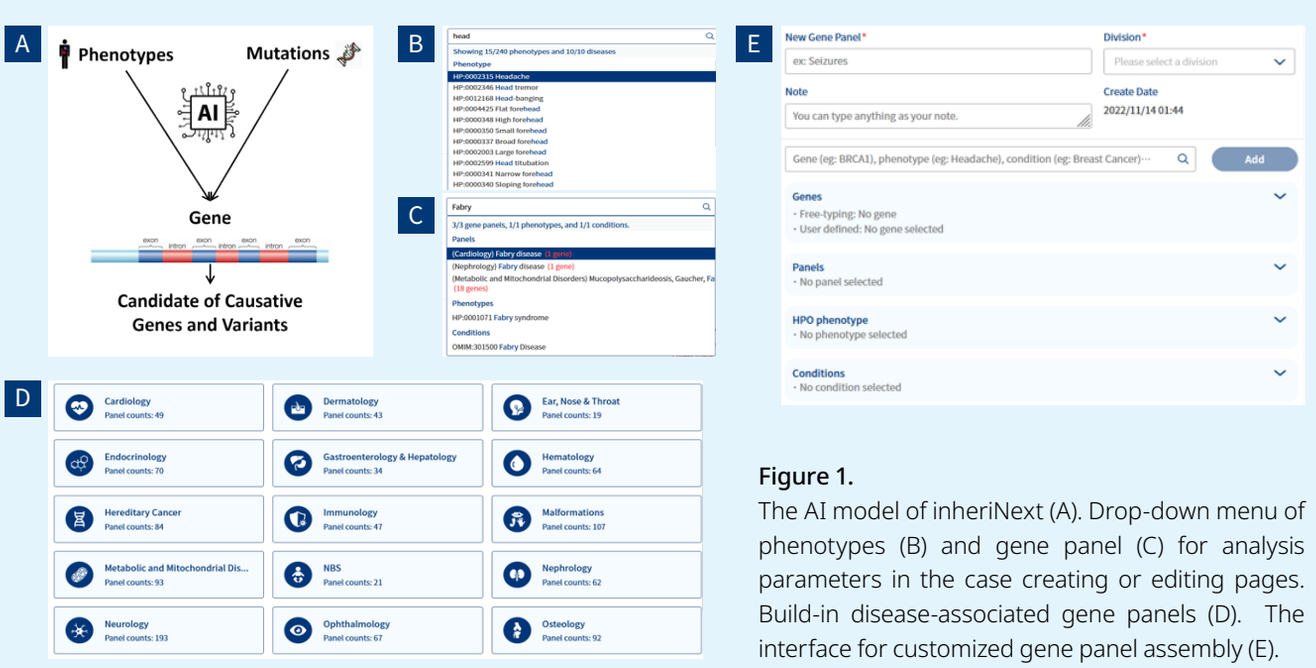


Figure 1.
The AI model of inheriNext (A). Drop-down menu of phenotypes (B) and gene panel (C) for analysis parameters in the case creating or editing pages. Build-in disease-associated gene panels (D). The interface for customized gene panel assembly (E).

critical reference during the analysis. InheriNext not only provides interpretation, review status, condition and conflict from ClinVar but also tabulates metrics between consequences and interpretation for each gene (Figure 2E). This could tremendously assist users to infer the pathogenicity according to consequences. The American College of Medical Genetics and Genomics (ACMG) published a guideline to classify variants based on 28 criteria (2015). However, the determination of these criteria is a labor burden. InheriNext implements up to 17 ACMG automated criteria and elaborates reasons for those criteria to save time for users (Figure 2F). In addition, the Proactive analysis is implemented and automatically filters pathogenic, null, carrier and ACMG secondary findings variants (Figure 2A). In summary, users could adopt these comprehensive annotations to determine the likely causative variants.

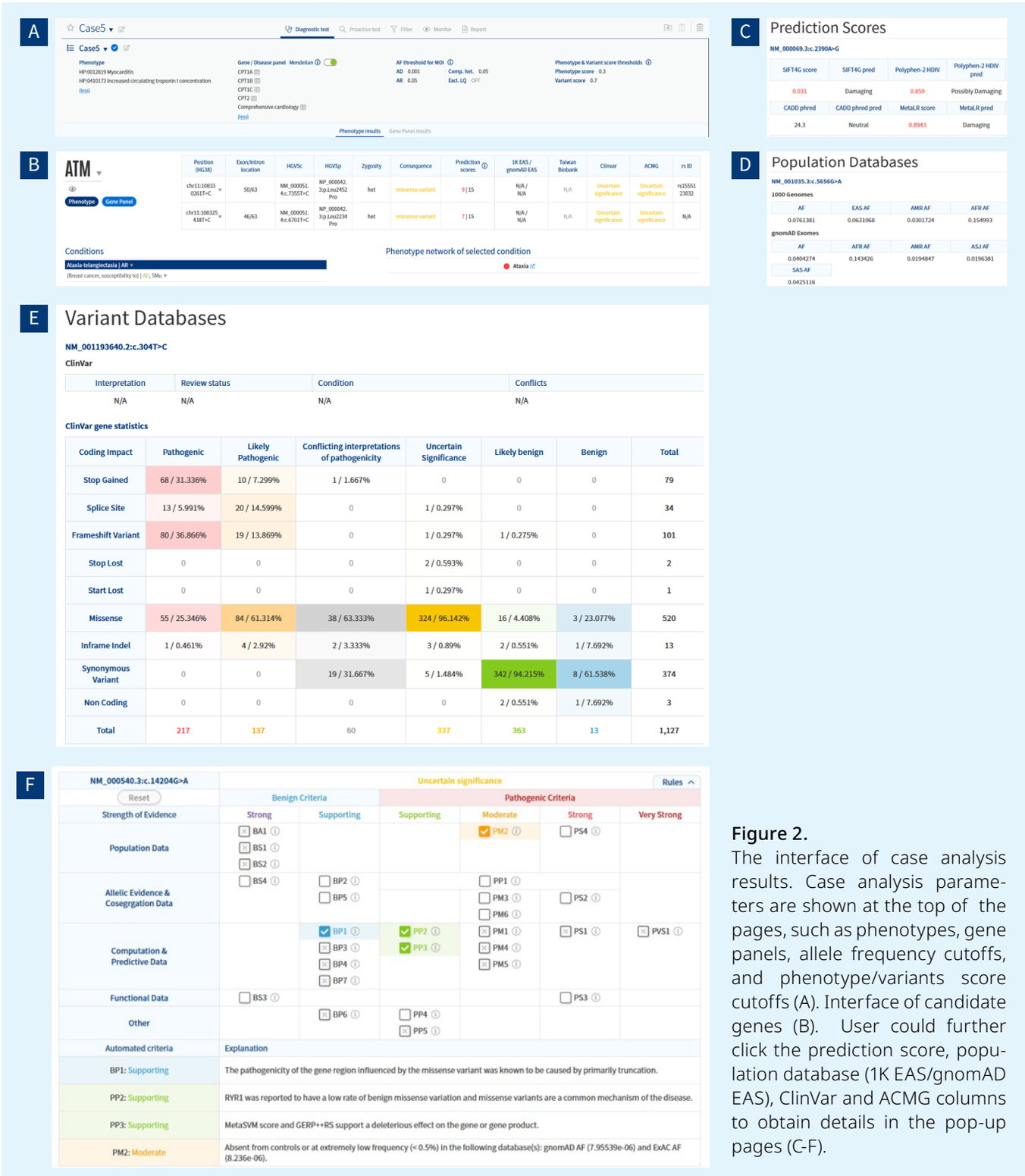


Figure 2.
The interface of case analysis results. Case analysis parameters are shown at the top of the pages, such as phenotypes, gene panels, allele frequency cutoffs, and phenotype/variants score cutoffs (A). Interface of candidate genes (B). User could further click the prediction score, population database (1K EAS/gnomAD EAS), ClinVar and ACMG columns to obtain details in the pop-up pages (C-F).

2 Explore the human genome

InheriNext brings a fully manual filtering function for seasoned users, the dynamic filter (Figure 3). Multiple filtering criteria can be applied, such as variants quality, allele frequency, consequence, chromosome coordinate, gene panel, ClinVar, dbSNP, ACMG classification and mode of inheritance. User-defined filter configurations can be saved, allowing quick adoption of filters for one other analysis.

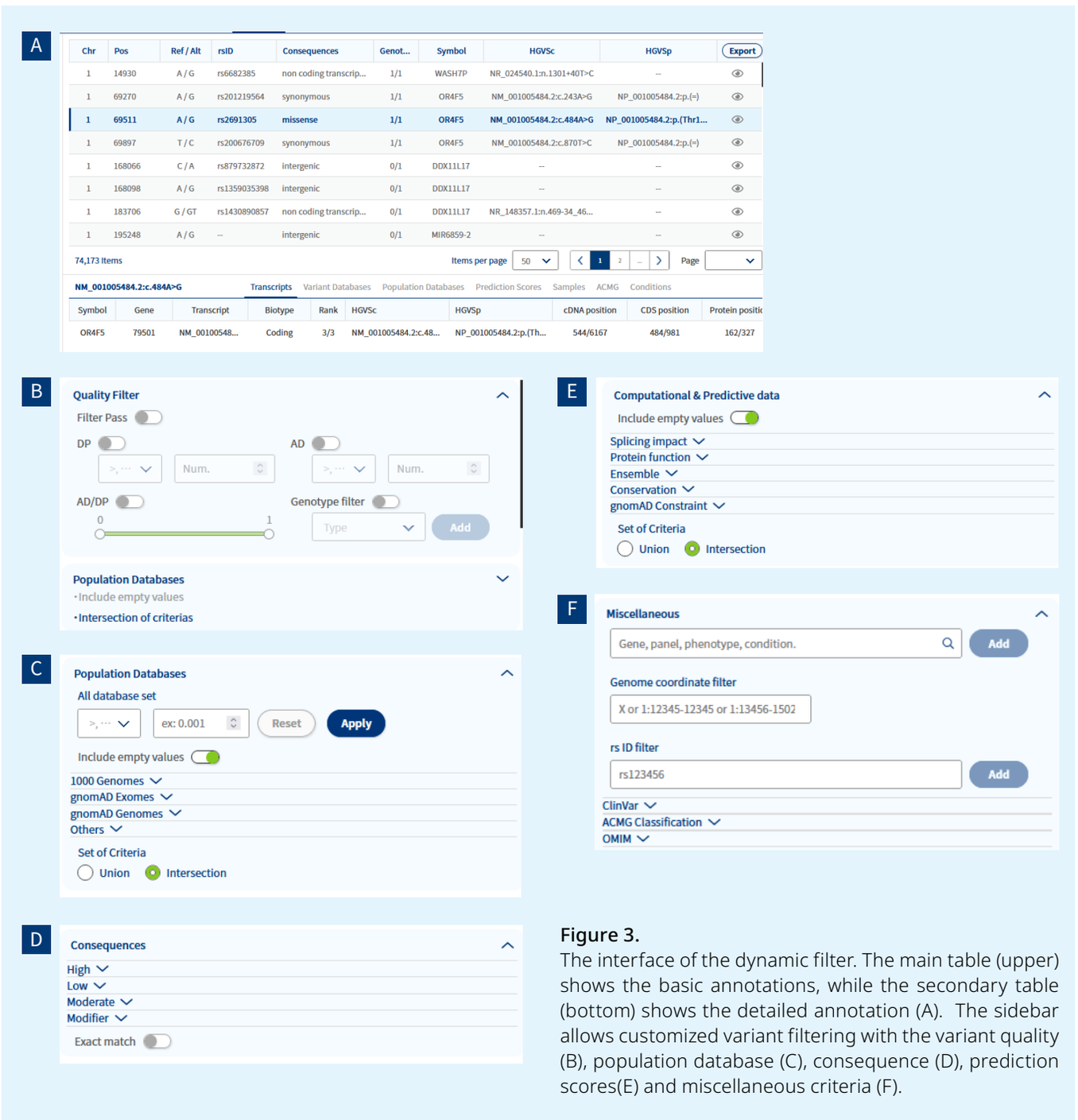


Figure 3.
The interface of the dynamic filter. The main table (upper) shows the basic annotations, while the secondary table (bottom) shows the detailed annotation (A). The sidebar allows customized variant filtering with the variant quality (B), population database (C), consequence (D), prediction scores(E) and miscellaneous criteria (F).

3 The report system

As long as the preliminary analysis is finished, users could add the likely-causative variants or any variants of interest to the Monitor (Figure 4A). The listed variants in the Monitor can be deleted or changed position with another variant accordingly. To report variants, users could add the "add to report" button in the Monitor. The added variants will be shown in the Report system. The Report system provides several chapters allowing users to choose which chapter is needed to be published (Figure 4B). These chapters include Diagnosis (phenotype and gene filter), Proactive (pathogenic, null, carrier, ACMG SF variants), Pharmacogenetics and Appendix, etc... The report can be reviewed and edited with its descriptions as well (Figure 4C).

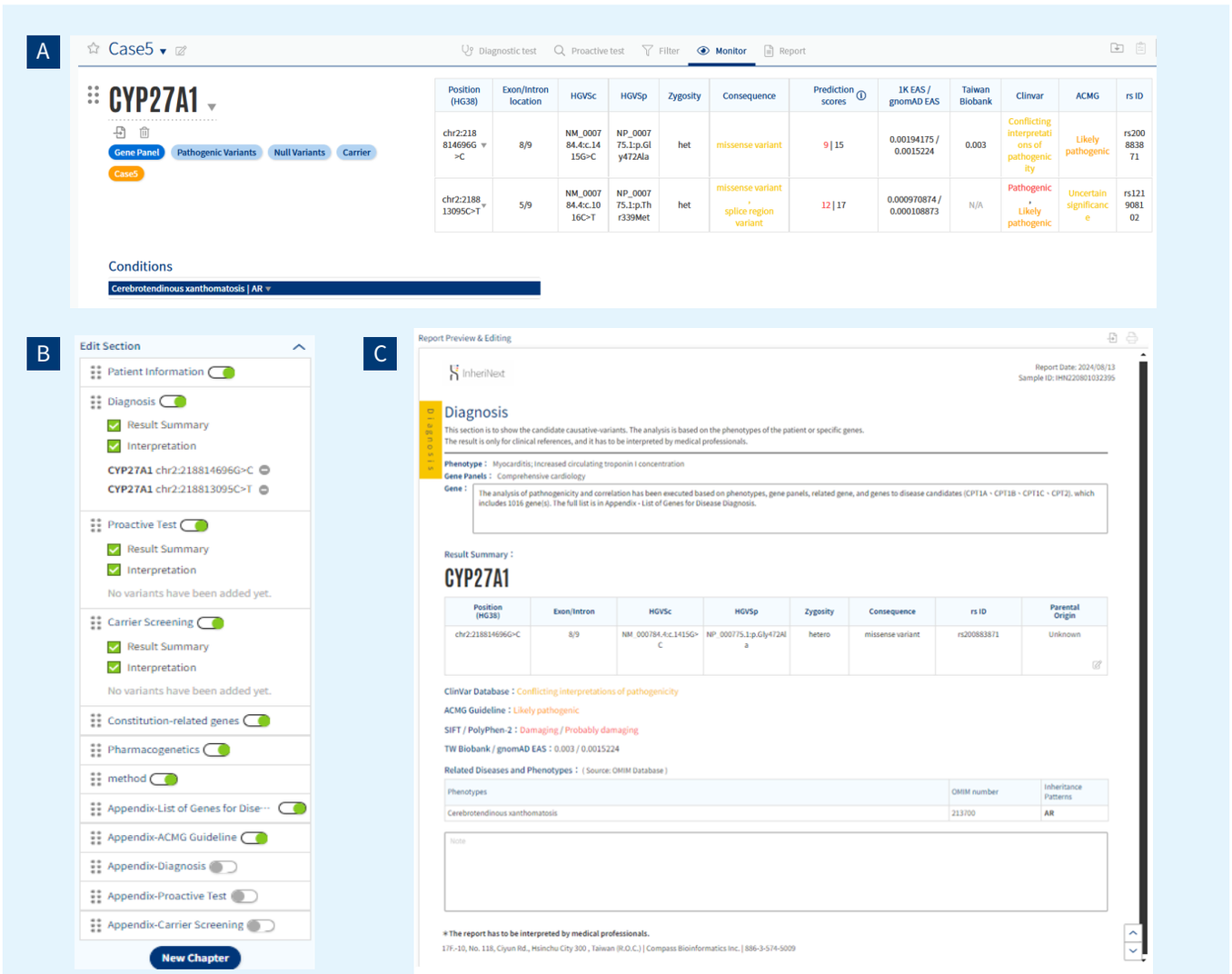


Figure 4.
The interface of the Monitor and the Report system. The Monitor section (A). The chapter selection tool in the Report system (B). Preview page of the Report system (C).

4 Build your own population database

The population databases play very important role regarding to the inherited disease analysis. Sometimes, we could find the highly recurrent variants in our dataset which exhibit extremely low AF in public databases. This may suggest that possible artifacts (sequencing or library preparation) or

population stratifications. Hence, a customized population database is essential. InheriNext implements the PhenoVarDB system to allow users to create their own population databases and annotate to their own cases (Figure 5A and B). The PhenoVarDB is also embedded with community and sharing system destined for creating a huge population database with other cooperators or community users.

B

PhenoVarDB Admin

Filter	Missing rate	AC	AN	AF	DP	AC in heterozygous genotypes
PASS	0.0679748	554	2660	0.208271	78169	428

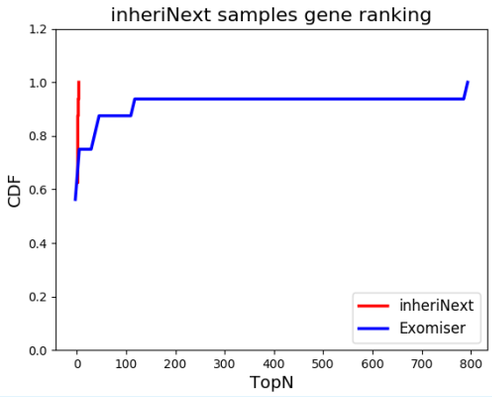
AC in homozygous genotypes	AC in hemizygous genotypes
126	0

Figure 5.
The statistic summary of customized population database (A). The statistic includes sex, variant count, affected case (%), affected gene(%), phenotype distributions, consequences proportion(%), variant counts per sample by consequences and AF distribution against TOPMED and gnomAD databases. A case was annotated by the customized population database which was shown in the pop-up page of population database annotation (B).

5 Performance

To evaluate the overall performance of InheriNext, we compared it with Exomiser using 16 samples. The data demonstrated that inheriNext pinpointed the positives within 10 top-ranked variants for each sample. However, Exomiser exhibited lower performance that some positives were found in hundreds-ranked variants (Figure 6A). This data suggested that InheriNext exhibits higher accuracy compare to the Exomiser. We further compared the performance and functionalities with a commercial product, competitor M. The data demonstrated that inheriNext outperforms competitor M in several functionalities, such as supported file format, independent analysis, gene panel prioritization, proactive analysis, automated ACMG annotations, number of allele frequency annotations, Dynamic filter, custom population database and genomic browser (Figure 6B). For analysis duration, even competitor M declares WES for 2-5 minutes and WGS for 10-15 minutes. However, it takes 1~2 hr for queuing. Therefore, it's hard to judge the analysis duration for inheriNext and competitor M. For the accuracy performance, inheriNext found the positives in the 4 and 1 samples with 1st and 2nd rank, respectively. While competitor M found the positives in the 4 samples with 1st rank but left 1 positive sample without answering. In conclusion, it's hard to judge which one outperforms one other owing to the sample size.

A



B

	InheriNext	Competitor M
Input file format	**Fastq, VCF**	VCF
Analysis duration (VCF)	WES : 5~10 min WGS : 15~20 min	**WES : 2~5 min** **WGS : 10~15 min**
Variant prioritization performance	**1st:4; 2nd:1**	1st:4; No result:1
Trio analysis	No (Developing)	**Yes**
Independent analysis results	**3**	1
Gene panel prioritization	**Yes**	No
Proactive analysis	**Yes (Pathogenic, Null, Carrier, ACMG IF)**	Yes (Carrier, ACMG IF)
Automated ACMG annotation	**Yes**	No (Manual)
Prediction score annotations	14	**17**
Allele frequency annotations	**11**	1
Report	Yes	Yes
Dynamic Filter	**Yes**	Yes (AF ≥ 2% variants were not showed)
Custom population database	**Yes (Summary charts, Community sharing)**	Yes
CNV	TXT file download(fastq)	**CNV interface**
QC metrics	TXT file download(fastq)	**QC interface**
Genomic browser	**8 tracks**	5 tracks

Figure 6.
The performance comparison of InheriNext. Cumulative distribution of top rank for inheriNext and Exsomiser (A). Comparison table of functionalities and performance between inherNext and competitor M (B).

+886 3-5745009 13400 Briarwick Drive, Unit 103, Austin, TX 78728
www.compassbioinfo.com service@compassbioinfo.com