Evaluating a Standard Benchmark for Gene Prioritization: The InheriNext® Algorithm's Integration of Genomic and Phenotypic Information

JY Chang^, KT Li^, M Kubal, YS Tsai, A Hamby, N Thomson, J Sheridan, S Barfield, R Rutz, FS Ong, R Felciano, S Kahn, SM Wu*

^co-first authors
*corresponding author

Abstract

Accurate and efficient variant prioritization is crucial for the interpretation of genomic data in both research and clinical settings. This study presents a comprehensive benchmarking analysis of InheriNext®, a novel tool designed for phenotype-driven pathogenic variant prioritization. We evaluated InheriNext's performance across diverse testing scenarios, comparing it against four established tools across various variant consequences with varying degrees of phenotype complexity and four major disease groups. For this study 7,244 whole exome test cases were generated using phenotype and genotype data from Phenopackets, along with pools of variants from healthy individuals to serve as genomic backgrounds. Sharing these benchmarking results and samples is intended to drive progress by assisting clinicians and researchers in evaluating interpretation tools and identifying areas of improvements.

Introduction

Rare diseases, defined as conditions affecting fewer than 200,000 individuals in the United States or fewer than 1 in 2,000 people in the European Union, still collectively impact over 400 million people worldwide, highlighting a significant global health challenge [1]. Many of these diseases, often caused by subtle genetic mutations, go undiagnosed for extended periods. Accurate identification of genetic variants is essential; however, with over 10 million genetic variations present in an average human genome, effective prioritization becomes crucial. Advanced sequencing technologies, such as Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS), provide efficient methods for profiling genetic data. Through these sequencing methods, a patient's specific genetic variants are identified, followed by prioritization based on the pathogenicity of the variants and the phenotypes relevant to the patient [2]. Previous research has highlighted the benefits of enhancing diagnostic accuracy by integrating phenotype data into variant prioritization algorithms [3, 4]. However, there remains significant room for improvement. For example, existing algorithms may not fully account for the complexity or diversity of phenotypic features, leading to incorrect prioritization of variants in certain cases.

Methods

About InheriNext®
InheriNext® is a cutting-edge AI-powered tool designed for the prioritization of genetic

variants in next-generation sequencing (NGS) data, handling both VCF and FASTQ formats. It addresses the existing limitations in achieving accurate and rapid diagnosis of complex inherited diseases by providing a comprehensive and precise analysis. The system significantly alleviates the workload of geneticists by delivering accurate candidate variants and versatile functionalities within minutes. InheriNext® distinguishes itself with two approaches in its variant prioritization algorithms: one utilizes Human Phenotype Ontology (HPO)-based phenotyping [5] and the other focuses on gene panels to identify causative SNPs and INDELs. Its user-friendly interface and customizable configurations enable users to effortlessly construct personalized presets with commonly used phenotypes or genes panels, enhancing analytical flexibility. Overall, InheriNext®'s robust capabilities and ease of use contribute to its potential as a valuable tool in genetic variant analysis for inherited diseases.

Benchmark Sources

The Global Alliance for Genomics and Health (GA4GH) has established the Phenopacket Schema as an ISO-approved standard in 2022 for sharing comprehensive clinical and genomic information at an individual level. A Phenopacket connects phenotype descriptions with disease, patient, and genetic data, allowing researchers to develop optimized accurate disease models [6]. The GA4GH Phenopacket Store version 0.1.21 emerges as an essential benchmarking resource, with a collection of 7,830 Phenopacket Samples representing a wide range of inherited diseases and disorders. The collection encapsulates 489 Mendelian and chromosomal diseases linked to 432 genes and 4,263 unique pathogenic alleles. The Phenopacket Samples are considered to be a robust large-scale repository, providing standardized, case-level phenotypic information and comprehensive clinical data sourced from the literature. Previous studies have also used the GA4GH Phenopacket as a key benchmark for evaluating ranking performance [7]. Consequently, it serves as an important standard for the development of InheriNext® diagnostic software.

K-means Clustering Diseases Using Term Frequency-Inverse Document Frequency (TF-IDF) and Principal Component Analysis (PCA)

The benchmark samples encompass nearly five hundred diseases. Grouping similar diseases helps reveal their distribution for the Phenopacket schema, facilitating a better understanding and allowing for performance analysis across different disease groups. The four (4) disease groupings acquired are clustered using the steps described below:

(1) Building Ontology Structure:

MONDO ontologies are used to standardize disease terminology within a unified framework to establish initial relationships and processes specific MONDO for analysis [8]. The diagram below shows a simplified version of the MONDO ontology (Fig1). At the root is "Disease" highlighted in the red box. Following it is "human disease", which branches into several second-layer categories, such as "cardiovascular disorder", "acute disease", "cancer or benign tumor", and "disorder of development or morphogenesis," each enclosed in a green box.

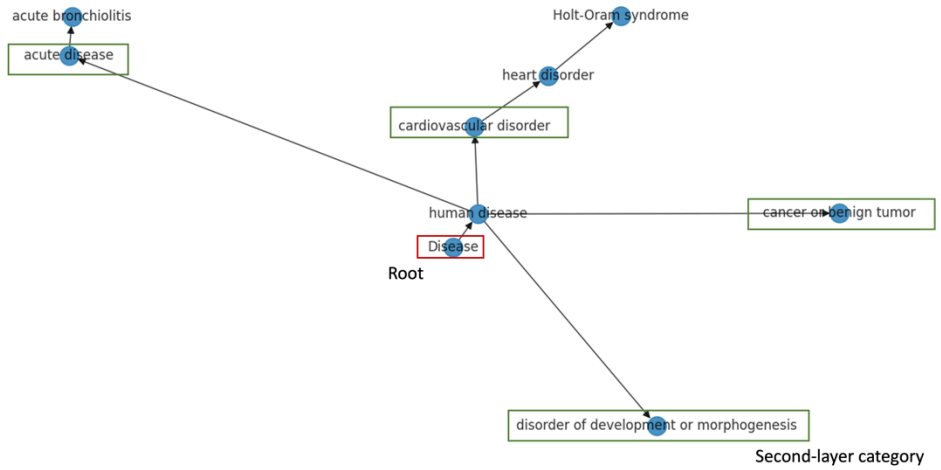


Fig1. The simplified ontology hierarchical structure of diseases in the MONDO System. This diagram represents a portion of the MONDO disease ontology, illustrating the hierarchical relationships among various diseases. The "Root" node at the center is labeled "Disease," from which several branches extend to different disease categories and sub-categories.

(2) Text Processing and Term Frequency-Inverse Document Frequency (TF-IDF) Calculation: In the Phenopacket-annotated diseases, there are 28 unique second-layer categories from MONDO ontologies. These second-layer categories composition terms are broken down into individual words, with stopwords such as "the", "is", "that" removed, and assembled into a word matrix. Each Phenopacket disease's corresponding second-layer categories are then represented by a set of word vectors, which are converted into TF-IDF vectors for subsequent similarity analysis [9].

(3) Principal Component Analysis (PCA) and K-means Clustering:
To present the results more intuitively, PCA is applied to reduce dimensionality and capture the most significant differences in the TF-IDF vectors of Phenopacket diseases [10]. This reduction process enables the visualization of results in a 2D plot. After applying PCA, k-means clustering is used to group the diseases based on these reduced dimensions. K-means clustering is an unsupervised learning algorithm used for grouping unlabeled data points into distinct clusters [11]. To effectively apply k-means, the elbow method [12] is used to determine the optimal number of clusters. Testing K values from one to seven shows that K=4 provides a balance between cluster compactness and flexibility, representing the most effective clustering.

Benchmark Study Design: Generate Simulated Benchmark Samples
To benchmark our variant prioritization approach, we used the GA4GH Phenopacket Store to impute a benchmark dataset of simulated genetic samples. First, genomes from 600 healthy individuals were obtained from the 1,000 Genomes Project, a comprehensive dataset representing genetic diversity across various populations. Considering the population bottlenecks experienced by European, Asian, and American populations, which drastically reduced genetic diversity in these groups [13], we have developed a method to generate synthetic Whole Exome Sequencing (WES) samples. By pooling variants from a diverse selection of healthy individuals, we create a rich reservoir from which 50,000 variants are randomly drawn to craft each simulated case of a “healthy” (normal genetic variability) exome. This pooling strategy ensures that the simulated exomes capture a broad spectrum of

genetic variability for analysis. Next, data from 7,244 of the 7,830 Phenopacket samples that meet the analysis criteria (e.g., annotated with phenotypic features) are used. Each known disease-causing variant is added into a synthetic healthy exome along with patient's phenotypic features to simulate a patient with that genetically-driven disease or disorder (Fig2).

This process results in 7,244 test cases representing disease patients, with each synthesized sample containing annotated phenotypic features, one pathogenic variant from Phenopacket, and 50,000 background variants. These samples are finalized for further analysis and used to evaluate the ranking performance of InheriNext® against four (4) other commonly used variant prioritization tools (A-D).

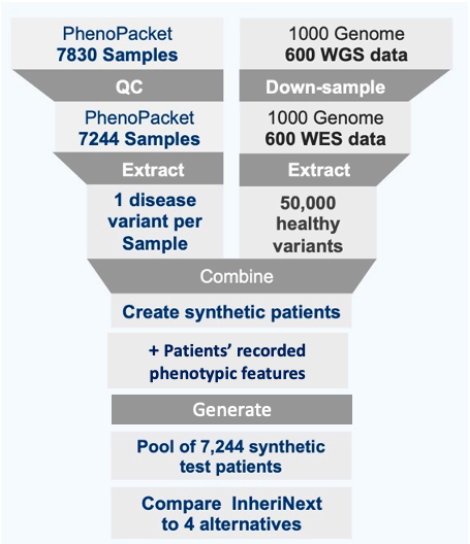


Fig2. Workflow for generating simulated samples. The diagram outlines the process for generating synthetic patients. We have prepared 600 healthy individuals' background genome sourced from the 1000 Genome Project, and the causative variants sourced from Phenopacket. After filtering, 7,244 synthetic samples, along with their respective phenotypic features, are created for the following benchmark study.

Results & Discussions

Ranking Performance in Benchmark Samples

InheriNext® is benchmarked against four (4) different software tools (A-D) that also utilize phenotype-driven gene prioritization methods to rank candidate pathogenic genes in the 7,244 samples. The tools are: Tool A (Exomiser [14]), Tool B (Lirical [15]), Tool C (Xrare [16]), and Tool D (Amelie [17]). The "Top-10 Rate" is a practical benchmark for evaluating tool performance, ensuring that the causative gene is captured within the top ranks. It is often visualized using the Cumulative Distribution Function (CDF) plot, which displays the proportion of genes that fall below any given rank. According to previous literature, this method offers an intuitive way to assess and compare the effectiveness of different tools in ranking causative genes [18]. Results show that InheriNext®, in 98.6% of the cases, correctly identifies the causative genes within the Top-10 rankings compared to tool A with 95.0%, tool B with 86.2%, tool C with 87.0%, and tool D with 90.9% (Fig3). The results indicate that InheriNext® is the most *sensitive* tool in this benchmarking study, having achieved the

highest Top-10 rate of 98.63%, paired with the lowest outside Top-10 rate of 1.37%, outperforming the other tools in both critical indexes.

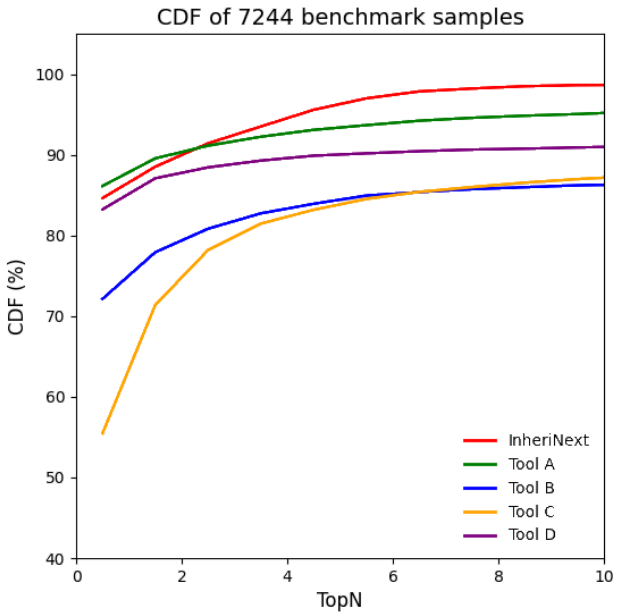


Fig3. Performance evaluation of InheriNext®® with other tools (A-D). The cumulative distribution function (CDF) shows the ranking distribution of causative genes across all five (5) tools. The CDF plots illustrate the percentage of the samples with causative genes ranked within the top*N* by each tool. *N* could be any integer between 1 and 10. Each tool is represented by a different color.

Ranking Performance in Different Variant Consequences

Variant consequences are pivotal in identifying pathogenic variants, which play a crucial role in diagnosing genetic disorders. This knowledge enables clinicians to pinpoint the genetic basis of a disease accurately, facilitating more precise diagnostics. By assessing the functional impact of these variants, one can evaluate the causative genes more precisely in algorithmic calculations. An analysis of the distribution of variant consequences in the benchmarked samples was conducted, revealing 24 distinct types of variant consequences. In some cases, a variant exhibits multiple types of consequences, resulting in duplicated counts. Table1A indicates that the three (3) most common types of consequences are missense, comprising 49.05% of the total, stop gained at 12.34%, and frameshift truncation accounting for 9.10%. The subsequent evaluation uses the Top-10 rate to assess each tool's performances in identifying causative genes across different variant consequences (refer to Table1B). InheriNext® excels in the three most common, therefore relevant, consequences (missense variant, stop gained, and frameshift truncation), outperforming the other tools. However, InheriNext® falls short in the synonymous variant category, achieving only a 30.43% in the Top-10 rankings.

(A)

Consequences	Proportion % (count)
Missense variant	49.05 (4345)
Stop gained	12.34 (1093)
Frameshift truncation	9.10 (806)
Coding transcript intron variant	6.48 (574)
Frameshift variant	4.96 (439)
Splice region variant	4.00 (354)
Splice donor variant	3.04 (269)
Frameshift elongation	2.34 (207)
Splice acceptor variant	1.96 (174)
Disruptive inframe deletion	1.51 (134)
Complex substitution	1.12 (99)
Inframe deletion	0.98 (87)
5 prime UTR exon variant	0.72 (64)
Feature truncation	0.58 (51)
Start lost	0.52 (46)
mnv	0.33 (29)
Direct tandem duplication	0.26 (23)
Synonymous variant	0.26 (23)
Disruptive inframe insertion	0.25 (22)
3 prime UTR intron variant	0.07 (6)
5 prime UTR intron variant	0.06 (5)
Stop lost	0.03 (3)
Inframe insertion	0.03 (3)
Exon loss variant	0.03 (3)

(B)

Consequences	InheriNext®	Tool A	Tool B	Tool C	Tool D
Missense variant	**98.64**	93.9	87.59	89.07	94.38
Stop gained	**99.27**	97.8	86.55	86.18	91.49
Frameshift truncation	**99.38**	98.39	91.44	89.33	93.05
Coding transcript intron variant	**98.08**	93.55	74.22	81.88	75.44
Frameshift variant	**98.86**	96.58	81.09	83.83	87.24
Splice region variant	**94.63**	94.35	76.27	85.59	66.67
Splice donor variant	**99.63**	98.51	86.99	86.25	95.54
Frameshift elongation	98.55	**100**	92.27	92.27	96.62
Splice acceptor variant	**99.43**	94.25	86.21	83.33	91.38
Disruptive inframe deletion	**99.25**	98.51	94.03	96.27	96.27
Complex substitution	**100**	**100**	93.94	94.95	97.98
Inframe deletion	**98.85**	96.55	83.91	83.91	81.61
5 prime UTR exon variant	**85.94**	82.81	1.56	3.12	3.12
Feature truncation	**100**	**100**	**100**	98.04	98.04
Start lost	**100**	84.78	97.83	91.3	**100**
Mnv	**100**	86.21	68.97	82.76	82.76
Direct tandem duplication	**100**	86.96	73.91	95.65	86.96

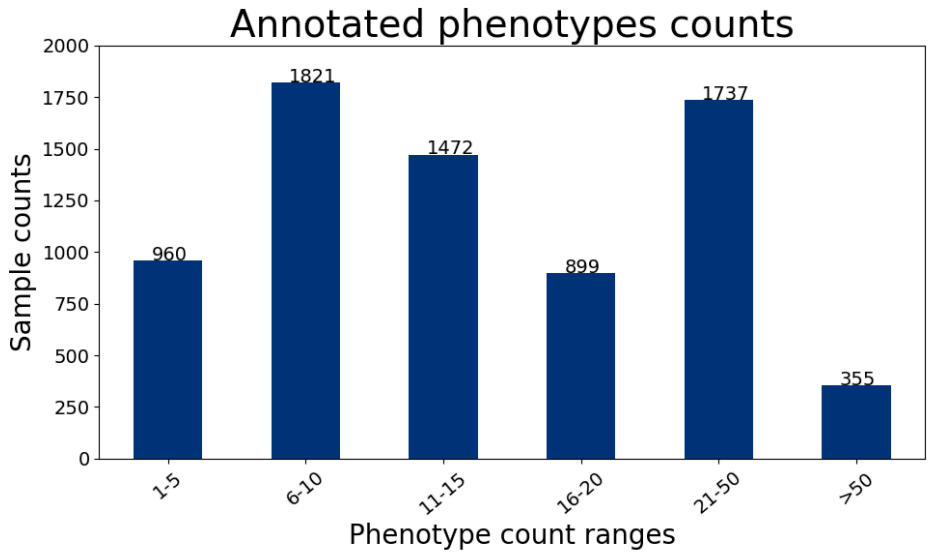
Synonymous variant	30.43	60.87	43.48	**78.26**	21.74
Disruptive inframe insertion	**100**	86.36	68.18	95.45	86.36
3 prime UTR intron variant	**100**	**100**	**100**	**100**	**100**
5 prime UTR intron variant	**100**	60	**100**	**100**	**100**
Stop lost	**100**	**100**	33.33	**100**	**100**
Inframe insertion	**100**	**100**	**100**	**100**	**100**
Exon loss variant	**100**	**100**	**100**	**100**	**100**

Table1. Distribution of Variant Consequences in Benchmark Samples. (A) The descending proportion and counts of 24 variant consequences identified in benchmark samples. (B) The breakdown of variant consequences by the five (5) tools. The percentages reflect the ability of each tool to rank causative genes within their Top-10s for each consequence.

Assessment of Diverse Annotated Phenotype Counts

Phenotypic features, as annotated in the Phenopacket schema, describe clinical symptoms in patients and are used to illustrate the similarity between disease-associated features and those present in a patient. InheriNext® integrates the HPO project [3], which provides an ontology of medically relevant phenotypic features and disease-phenotype annotations, to calculate the likelihood of diseases based on the phenotypic features observed in patients. For example, "arachnodactyly" is a relevant feature for "Marfan syndrome" in HPO database, so if a patient exhibits "arachnodactyly," they are more likely to have the disease. InheriNext® and other tools use phenotype-driven methods to rank potential pathogenic genes based on phenotypic features in addition to patient genotype data. In the benchmark samples, diverse annotated phenotype counts are inputted for each sample based on phenotypic features in the Phenopacket record used to create that particular simulated patient. However, a large count might decrease the performance in ranking causative genes because it may include more unrelated (noise) features, which are not annotated in the disease. For phenotype counts distribution, the results show that: the range of 6-10 phenotypes is most common, seen in 1821 samples, followed closely by the 21-50 range with 1737 samples (Fig4 A). The number of phenotype inputs across each tool's Top-10s were evaluated to assess performance. The results show that InheriNext® demonstrates consistent performance with percentages ranging from 98% to 100%, indicating strong capability across all phenotype input ranges. Tool A maintains high percentages, reaching 97% in the 6-10 and 11-15 ranges, but drops to 82% in the >50 category. Tool B, Tool C, and Tool D remain steady with over 82% in most ranges; however, they experience a more significant drop in the >50 category. These charts highlight InheriNext®'s non-discriminatory capabilities in managing various phenotype input ranges and implies its robustness in handling of noise features (Fig4 B).

(A)



(B)

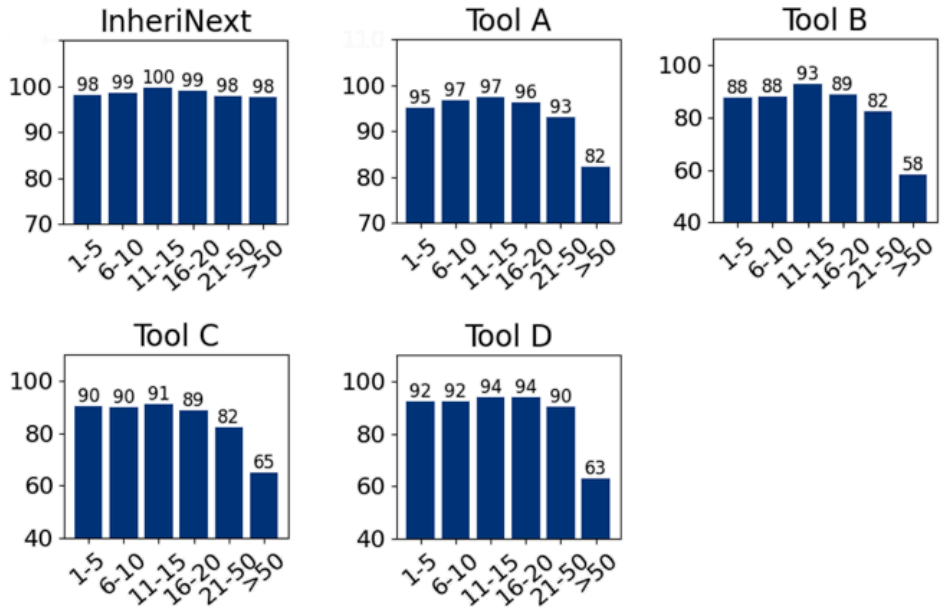


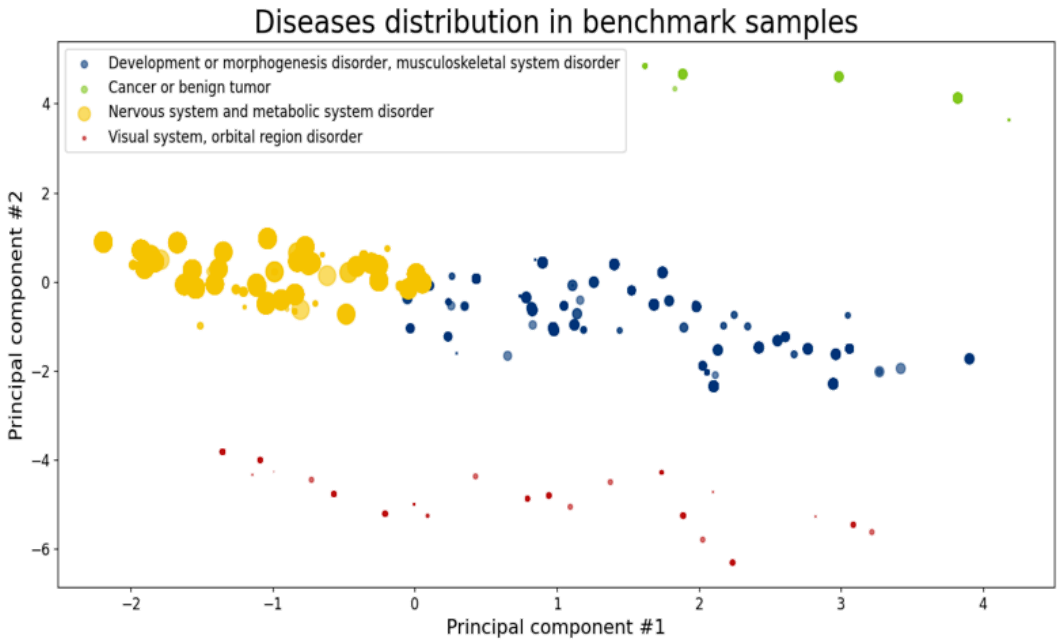
Fig4. Distribution of Phenotypes Counts in Benchmark Samples. (A) The bar graph illustrates the distribution of samples with different ranges of annotated phenotypic feature counts. Each bar's height represents the sample count within a specific range. (B) The five bar charts represent the performance of different tools—InheriNext®, Tool A, Tool B, Tool C, and Tool D—across various feature count ranges. Each chart shows the percentage of samples' causative genes successfully ranked within the Top-10 for each phenotypic feature count range.

Evaluation of Ranking Performance in Four Disease Groups

Each benchmark sample is annotated with a specific disease. However, some diseases do not match a specific MONDO ID and is therefore removed, leaving 7215 samples for analysis. The Phenopacket-annotated diseases can be classified into four (4) major groups using the previously described elbow method in k-means, which indicates the distribution of diseases across all Phenopacket samples.

The scatter plot below displays the distribution of different diseases across four major disease groups, with each data point symbolizing a distinct disease. The size of each point reflects the number of samples for that disease. As reflected by the dense cluster of orange points, this plot clearly shows that the Nervous System and Metabolic System Disorder group has the highest number of samples (Fig5 A). A table is presented that lists examples of actual diseases found in the benchmark samples grouped in four major disease categories (Fig5 B). To evaluate ranking performance, the Top-10 rate was used as an indicator across four disease groups. The results reveal that InheriNext® consistently ranks over 98% of samples' causative genes within the Top-10 across all groups, demonstrating strong non-discriminatory performance. In comparison, other tools show variability based on the disease type. For example, Tool B performs well in the Cancer or Benign Tumor category but less effectively in the Nervous or Metabolic Disorder group. Overall, InheriNext® demonstrates consistent effectiveness across 4 disease groups (Fig5 C), indicating that InheriNext® is well-suited for evaluating clinical genetic profiles in patients with a wide range of clinical indications.

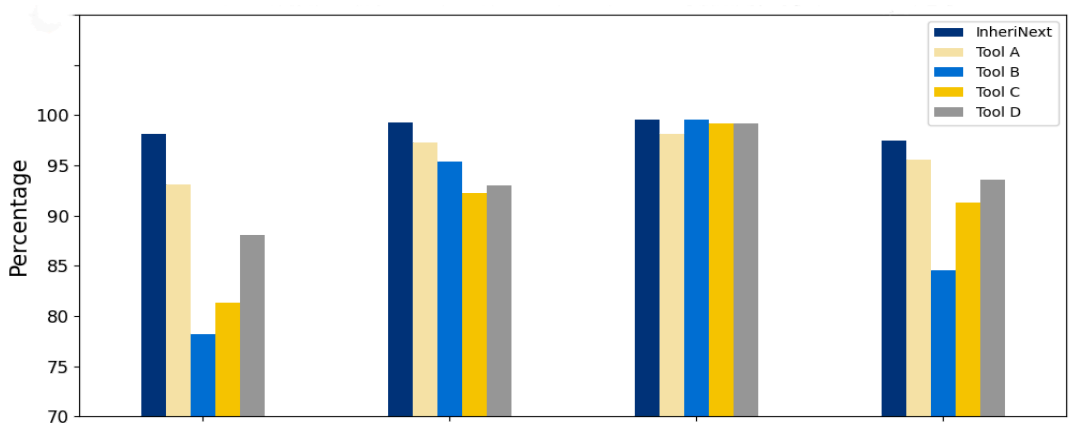
(A)



(B)

Disease Groups	Examples	
Development or morphogenesis disorder, musculoskeletal system disorder	KBG syndrome	Holt-Oram syndrome
Nervous system & metabolic system disorder	Cockayne syndrome	Developmental and epileptic encephalopathy 4
Cancer or benign tumor	Multiple endocrine neoplasia type 2A	Neurofibromatosis type 1
Visual system, orbital region disorder	Leber congenital amaurosis 6	Ectopia lentis, familial’

(C)



For the purpose of plotting charts, we abbreviated the names of disease groups.
*Development or morphogenesis disorder, musculoskeletal system disorder = Development or musculoskeletal disorder
* Nervous system and metabolic system disorder = Nervous or metabolic disorder
* Cancer or benign tumor = Cancer or benign tumor
* Visual system, orbital region disorder = Visual, orbital region disorder

Fig5. Disease Distribution in Benchmark. (A) The scatter plot illustrates the distribution of diseases by condensing complex data into two principal components (Principal Component #1 and Principal Component #2), highlighting similarities and clustering in four disease groups among the samples. Each point represents a unique disease, and the size indicates the number of samples for that disease; larger points signify a greater number of samples. (B) Example of actual diseases observed in the benchmark samples across four disease groups.

(C) The performance comparison across the five tools shows the percentage of samples' causative genes in Top-10 distributed among the four groups of diseases.

Strengths of the InheriNext®

InheriNext® benchmarking tests revealed several key performance advantages over existing approaches, which we attribute to the following factors:

(1) Integration of Disease Similarity Score Enhances InheriNext® Ranking Performance:
InheriNext® engages in discussions with clinicians to continually improve the ranking of pathogenic genes. Notably, a clinician's feedback seeking to better assess the similarity between disease symptoms and patient phenotypes led to InheriNext®'s integration of a disease similarity score. The significance of this enhancement is evident in benchmark results, which demonstrate strong and stable performance in ranking various diseases (Fig5). This integration ensures that the system is more aligned with clinical realities, ultimately leading to more accurate predictions and diagnostics.

(2) Delay Filtering and Removal of Candidates Until After Prioritization:
InheriNext® evaluates all potential causative variants and genes, even those unlikely to be pathogenic, acknowledging their possible impact on disease. While other tools may apply variant consequence filters, such as excluding 5 prime UTR exon variants to narrow down candidate genes, this approach can lead to the exclusion of important pathogenic genes, potentially missing critical genetic insights (Table1B: 5 prime UTR exon variant, Tool B-D). However, InheriNext® retains all causative genes on the ranking list, which ensures consistent ranking performance across diverse variant consequences. By comprehensively including all potential variants, InheriNext® facilitates a more thorough exploration of genetic data, accommodating potential exceptions and rare pathogenic occurrences.

Limitations and Future Work

Our benchmark study highlighted suboptimal performance by InheriNext®'s in certain samples. Further investigation has identified two (2) primary causes, providing key areas for future optimization.

(1) Variants Annotated With Synonymous Consequence Did Not Rank Well (Table1B):
InheriNext® initially excluded synonymous variants from its algorithm due to its general likelihood of being neutral compared to non-synonymous variants. This "rule of thumb" criterion was aimed at streamlining the analysis by focusing on variants with a higher probability of pathogenicity. However, this exclusion led to important variant genes being omitted from the ranking lists in the InheriNext®, which our strategy seeks to avoid (see "Delay Filtering and Removal of Candidates Until After Prioritization"). Some research shows many synonymous variants, previously considered neutral, can indeed have significant pathogenic effects: they can impact splicing motifs and cryptic splice sites, alter micro-RNA binding sites [19], and influence mRNA structure and protein levels by reducing codon optimality [20, 21, 22]. Based on these findings, the strategy involves incorporating synonymous variants, acknowledging their potential pathogenic value and improving the overall analysis.

(2) High Allele Frequency Of The Variants Affect Ranking:
The ClinGen Sequence Variant Interpretation Working Group has refined the ACMG/AMP variant pathogenicity guidelines for rule BA1, which designates variants with a minor allele frequency (MAF) > 0.05 as benign. However, they identified nine (9) variants with MAF > 0.05 that may exhibit pathogenicity. These findings suggest that MAF and pathogenicity do not always correlate. As a result, InheriNext® algorithm may inadvertently penalize genes

based on the high allele frequency and affect the ranking accuracy. While high-frequency variants are typically deemed benign, some may still possess pathogenic potential, emphasizing the need to evaluate additional factors when assessing variant pathogenicity. Future optimizations will incorporate additional criteria for adjusting gene scores. The intention is to retain the variant frequency filter as recommended by ACMG/AMP guidelines [23]. However, specific exceptions are systematically reviewed and curated as they appear in the clinical genetics literature, allowing potentially pathogenic variants to be considered despite their relatively high frequency.

Conclusion

The benchmarking analysis demonstrates InheriNext®'s highly performant and non-discriminatory capability in prioritizing candidate pathogenic genes across various testing scenarios. When using phenotype-driven gene prioritization methods, InheriNext® consistently outperformed four (4) different tools (A-D), demonstrating highest sensitivity and achieving the highest Top-10 capture rate and lowest missed rate (outside of Top-10). This exceptional performance is underscored by InheriNext®'s robust handling of diverse variant consequences, where it excelled in three most common consequences (missense variants, stop gains, and frameshift truncations), although it did show some limitations in synonymous variants.

Our assessment of performance across diverse phenotype count ranges further highlights InheriNext®'s strength, maintaining high performance levels even as phenotype complexity increases. InheriNext® maintains consistent Top-10 performance across varying phenotype counts, in contrast to the other tools which showed performance drops, suggesting robust adaptability and reliability. Moreover, when evaluated within four (4) major disease groups among the Phenopacket samples, InheriNext® maintained its leading position, showcasing its versatility and effectiveness across different disease contexts. Overall, these results establish InheriNext® as a cutting-edge tool in the field of gene prioritization, demonstrating outstanding ranking performance. Its ability to effectively integrate genomic and phenotypic data positions it as a critical resource for diagnostics in genetic research and clinical applications.

Acknowledgements

We acknowledge Melissa Haendel and the Monarch Initiative (https://monarchinitiative.org/) in developing a standardized framework for integrating genomic, environmental, and phenotypic data [6]. Phenopackets provide a critical foundation for ensuring that phenotype profiles are FAIR++ for disease diagnosis and discovery [24]. Their efforts in standardizing data formats, including gene data (GFFs, VCFs, BED files), environmental data, and the novel PXF format, have been instrumental in advancing interoperability and data sharing in rare disease research.

Data Availability

The benchmark simulated samples from the GA4GH Phenopacket are available from the corresponding author upon request.

References

1. Groft, S. C., Posada, M., & Taruscio, D. (2021). Progress, challenges and global approaches to rare diseases. *Acta paediatrica*, *110*(10), 2711-2716.

2. Basel-Salmon, L. (2024). Phenotypic compatibility and specificity in genomic variant classification. *European Journal of Human Genetics*, *32*(5), 471-473.

3. Robinson, P. N., Köhler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., ... & Sanger Mouse Genetics Project. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome research*, *24*(2), 340-348.

4. Jacobsen, J. O., Kelly, C., Cipriani, V., Research Consortium, G. E., Mungall, C. J., Reese, J., ... & Smedley, D. (2022). Phenotype‐driven approaches to enhance variant prioritization and diagnosis of rare disease. *Human mutation*, *43*(8), 1071-1081.

5. Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, *83*(5), 610-615.

6. Danis, D., Bamshad, M. J., Bridges, Y., Caballero-Oteyza, A., Cacheiro, P., Carmody, L. C., ... & Robinson, P. N. (2025). A corpus of GA4GH Phenopackets: case-level phenotyping for genomic diagnostics and discovery. *Human Genetics and Genomics Advances*, *6*(1).

7. Bridges, Y., de Souza, V., Cortes, K. G., Haendel, M., Harris, N. L., Korn, D. R., ... & Jacobsen, J. O. (2024). Towards a standard benchmark for variant and gene prioritisation algorithms: PhEval-Phenotypic inference Evaluation framework. *bioRxiv*.

8. Vasilevsky, N. A., Matentzoglu, N. A., Toro, S., Flack IV, J. E., Hegde, H., Unni, D. R., ... & Haendel, M. A. (2022). Mondo: unifying diseases for the world, by the world. *MedRxiv*, 2022-04.

9. Arivarasan, A., & Karthikeyan, M. (2019). Data mining K-means document clustering using tfidf and word frequency count. *International Journal of Recent Technology and Engineering*, *8*(2), 2542-2549.

10. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

11. Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning* (p. 29).

12. Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering* (Vol. 336, p. 012017). IOP Publishing.

13. Zheng-Bradley, X., & Flicek, P. (2017). Applications of the 1000 Genomes Project resources. *Briefings in functional genomics*, *16*(3), 163-170.

14. Smedley, D., Jacobsen, J. O., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., ... & Robinson, P. N. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature protocols*, *10*(12), 2004-2015.

15. Robinson, P. N., Ravanmehr, V., Jacobsen, J. O., Danis, D., Zhang, X. A., Carmody, L. C., ... & Smedley, D. (2020). Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, *107*(3), 403-417.

16. Li, Q., Zhao, K., Bustamante, C. D., Ma, X., & Wong, W. H. (2019). Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genetics in Medicine*, *21*(9), 2126-2134.

17. Birgmeier, J., Haeussler, M., Deisseroth, C. A., Steinberg, E. H., Jagadeesh, K. A., Ratner, A. J., ... & Bejerano, G. (2020). AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Science translational medicine*, *12*(544), eaau9113.

18. Masino, A. J., Dechene, E. T., Dulik, M. C., Wilkens, A., Spinner, N. B., Krantz, I. D., ... & White, P. S. (2014). Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. *BMC bioinformatics*, *15*, 1-11.

19. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., & Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, *156*(6), 1324-1335.

20. Brest, P., Lapaquette, P., Souidi, M., Lebrigand, K., Cesaro, A., Vouret-Craviari, V., ... & Hofman, P. (2011). A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nature genetics*, *43*(3), 242-245.

21. Bartoszewski, R. A., Jablonsky, M., Bartoszewska, S., Stevenson, L., Dai, Q., Kappes, J., ... & Bebok, Z. (2010). A synonymous single nucleotide polymorphism in ΔF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *Journal of Biological Chemistry*, *285*(37), 28741-28748.

22. Kim, A., Le Douce, J., Diab, F., Ferovova, M., Dubourg, C., Odent, S., ... & de Tayrac, M. (2020). Synonymous variants in holoprosencephaly alter codon usage and impact the Sonic Hedgehog protein. *Brain*, *143*(7), 2027-2038.

23. Kim, S. Y., Kim, B. J., Oh, D. Y., Han, J. H., Yi, N., Kim, N. J., ... & Choi, B. Y. (2022). Improving genetic diagnosis by disease-specific, ACMG/AMP variant interpretation guidelines for hearing loss. *Scientific Reports*, *12*(1), 12457.

24. Haendel, M. (2016). Phenopackets: Making phenotype profiles FAIR++ for disease diagnosis and discovery. figshare. Presentation. https://doi.org/10.6084/m9.figshare.3180898.v1