BMC Genomics

SOFTWARE

Open Access

Single-cell data combined with phenotypes improves variant interpretation

Timothy Chapman[1,2] and Timo Lassmann[1,2*]

Abstract

Background Whole genome sequencing offers significant potential to improve the diagnosis and treatment of rare diseases by enabling the identification of thousands of rare, potentially pathogenic variants. Existing variant prioritisation tools can be complemented by approaches that incorporate phenotype specificity and provide contextual biological information, such as tissue or cell-type specificity. We hypothesised that integrating single-cell gene expression data into phenotype-specific models would improve the accuracy and interpretability of pathogenic variant prioritisation.

Methods To test this hypothesis, we developed IMPPROVE, a new tool that constructs phenotype-specific ensemble models integrating CADD scores with bulk and single-cell gene expression data. We constructed a total of 1,866 Random Forest models for individual HPO terms, incorporating both bulk and single cell expression data.

Results Our phenotype-specific models utilising expression data can better predict pathogenic variants in 90% of the phenotypes (HPO terms) considered. Using single-cell expression data instead of bulk benefited the models, significantly shifting the proportion of pathogenic variants that were correctly identified at a fixed false positive rate ($p < 10^{-30}$, using an approximate Wilcoxon signed rank test). We found 57 phenotypes' models exhibited a large performance difference, depending on the dataset used. Further analysis revealed biological links between the pathology and the tissues or cell-types used by these 57 models.

Conclusions Phenotype-specific models that integrate gene expression data with CADD scores show great promise in improving variant prioritisation. In addition to improving diagnostic accuracy, these models offer insights into the underlying biological mechanisms of rare diseases. Enriching existing pathogenicity-related scores with gene expression datasets has the potential to advance personalised medicine through more accurate and interpretable variant prioritisation.

Keywords Rare disease, Variant prioritisation, Machine learning, Random forest, Interpretable models, Whole Genome sequencing

Background

Rare disease diagnosis remains an immense challenge. While specific rare diseases are singularly uncommon, collectively rare diseases are prevalent, with 250–450 million people estimated to be affected globally [1]. In addition to the burden on patients and their families [2], rare diseases create a large cost on the healthcare system, around a trillion dollars in America alone [3]. Diagnosis is complicated by the asymmetry between the total number of people affected overall and the rarity of individual

*Correspondence:
Timo Lassmann
timo.lassmann@thekids.org.au
[1] The Kids Research Institute Australia, 15 Hospital Ave, Nedlands, WA 6009, Australia
[2] UWA Centre for Child Health Research, The University of Western Australia, 35 Stirling Hwy, Crawley, Western Autralia 6009, Australia

diseases. Furthermore, the symptoms of rare diseases can overlap with more common conditions, increasing the risk of misdiagnosis.

Whole Genome Sequencing (WGS) has emerged as a valuable tool for rare disease diagnosis due to its comprehensive coverage of the genome. WGS allows for the detection of a wider range of potential disease-causing variants, including those in regulatory regions or involving complex genomic rearrangements. As WGS becomes more cost-effective, it is increasingly preferred as the first line of testing in rare disease cases [4, 5]. With 72% of rare diseases thought to be genetic [1], genetic testing is a critical avenue for rare disease diagnosis. When paired with well-annotated clinical datasets, WGS can provide key insights into the underlying genetic basis of rare diseases, facilitating more accurate diagnoses and personalised treatments. However, applying these insights in practice remains challenging. Despite substantial and ongoing work, individual rare disease diagnosis is hampered by a paucity of clinical and genetic data, making genotype–phenotype correlations tenuous.

Databases such as ClinVar and the Human Phenotype Ontology (HPO) serve as valuable resources for linking genetic variants to disease. ClinVar [6] catalogues a growing number of clinically verified pathogenic and benign variants, while the HPO offers a hierarchy of phenotypes, and associations between phenotypic features and genes. Together with efforts like Online Mendelian Inheritance in Man (OMIM), Orphanet, and the DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER), these resources provide an essential diagnostic reference for clinicians and researchers. These datasets provide a valuable foundation for variant interpretation, and our approach takes advantage of the potential from combining these resources to enhance variant prioritisation.

With burgeoning biomedical datasets, many variant prioritisation tools have been developed over the last decade to help identify potentially pathogenic variants. Various omics datasets form the core of different prioritisation tools. Some approaches focus on functional genomics [7], or proteomics specifically [8], with each approach providing unique benefits to variant prioritisation. The value of phenotype information is well-established [9], and has been incorporated into tools such as eXtasy [10], PhenGen [11], Exomiser [12], and AMELIE [13]. In these tools, phenotype data is typically used at a late stage of the variant prioritisation process to re-rank or filter candidate variants. Our previous work on VARPP [14] provided an early demonstration that phenotype-specific models can be derived from general-purpose in-silico pathogenicity scores by linking gene expression data with HPO terms.

Alongside the exploration of data sources, significant effort has been directed towards the development of more sophisticated models to improve prediction accuracy [15, 16]. Some models combine outputs from multiple predictors to optimise predictive performance [17, 18]. While these more complex meta-predictors can provide high-quality predictions, these methods are difficult to interpret, complicating the model inspection and explainability. Beyond identifying the pathogenic variant, interpretable models that can explain predictions are able to support clinical decision-making and guide research. Interpretability is particularly important when trying to identify key features of a rare disease (relevant tissues and cell-types in this study), and in the development of targeted therapeutic interventions [19, 20]. There remains an ongoing need for simple, interpretable tools to prioritise rare disease variants effectively.

Recent advances in single-cell RNA sequencing offer new opportunities to improve the biological relevance of features used in variant prioritisation models. Single-cell data provides unprecedented resolution by capturing gene expression profiles of individual cells. This granularity enables the identification of cell-type-specific expression patterns and rare cell populations. These patterns, lost in the blending of cell populations in bulk RNA-seq data, provide a deeper understanding of cellular heterogeneity and its role in disease. We also know that some heritable differences in gene function are captured in the stochasticity of gene expression, which is observable on a cell-to-cell basis but lost by averaging [21]. Integrating single-cell data with variant prioritisation approaches could improve the accuracy of genotype–phenotype correlations and enhance interpretability, giving clinicians insights into disease mechanisms and providing diagnostic guidance.

Given these advances, we hypothesise that incorporating single-cell gene expression data into phenotype-specific models will improve both the interpretability and accuracy of pathogenic variant prioritisation. To test this hypothesis and address the need for interpretable phenotype-specific tools integrating expression data, we have built on VARPP [14] to develop IMPPROVE: an Interpretable Modular Platform for the Prediction of Relevant Obscure Variants by Expression. IMPPROVE utilises Random Forest models, a machine learning method known for effectively balancing predictive accuracy and interpretability. We evaluate the predictive performance and interpretability of models using bulk and single-cell RNA-seq data. This combination of biological context with interpretable models

provides researchers and clinicians with a tool to identify pathogenic variants and understand the biological rationale behind predictions, uncovering novel disease mechanisms and improving diagnostic outcomes.

Methods

Experimental design

IMPPROVE serves as a framework for building ensemble model-based pathogenicity predictors, and generating variant reports. The framework allows for flexible selection of a baseline pathogenicity score, gene expression dataset, and ensemble model. It is implemented within a containerised computational pipeline, providing flexibility and reproducibility across computational environments. To test our hypothesis effectively, we selected Combined Annotation Dependent Depletion (CADD) phred scores (version 1.6) as the baseline pathogenicity score, and Random Forests as the ensemble model. These components were chosen for their well-established use, allowing us to demonstrate the value of bulk and single-cell expression data. This experimental setup enables rigorous evaluation of how gene expression data contributes to variant prioritisation.

CADD scores [22] predict the deleteriousness of Single Nucleotide Variants (SNVs) by combining genetic conservation scores, epigenetic modifications, functional predictions, and genetic context. Moreover, CADD's independence from ClinVar is crucial to avoid potential circularity. Circularity can occur if model performance is evaluated using data that overlaps with training data, biasing results. Since we use pathogenic variants from ClinVar to label pathogenic variants in the training and testing data, it is important to use a score like CADD that is independent of ClinVar.

Random Forests are a widely known ensemble machine learning model. They were chosen in our study for their balance of accuracy and interpretability. A Random Forest is comprised of many individual decision trees, which make predictions through a series of yes/no questions about the data (e.g., "is the number of transcripts per million of this gene in heart tissue greater than this threshold?"). The cumulative nature of these decisions enables us to identify the key features contributing to pathogenicity predictions, making them particularly suitable for investigating phenotype-specific effects of variants. Random Forests are highly interpretable, not prone to overfitting, and straightforward to train.

Data sources

IMPPROVE requires per-phenotype tabular datasets to learn rules that predict variant pathogenicity using gene expression. This section provides a high-level overview of how we construct phenotype-specific expression datasets from public resources. The preparation workflow is outlined in Fig. 1, with more detail on the individual steps given in subsections.

Starting with a list of phenotypes, given as HPO terms, we associate each phenotype with a list of genes using Phen2Gene [23]. For each phenotype-specific gene set, we select high-confidence pathogenic variants from ClinVar and high-frequency benign variants from the combined Genome Aggregation Database (gnomAD). We match gene symbols from Phen2Gene with ClinVar variants using the HGNC database [24]. Pathogenic and benign variants are then filtered and assigned to genes by intersecting them with exon coordinates from GENCODE (version 26). Each variant is annotated with its CADD phred score and the gene's expression profile, either using builk RNA-seq data from Genotype-Tissue Expression (GTEx) [25] or single-cell data from Tabula Sapiens [26].

This process produced 14,928 combined phenotype-specific datasets with paired gene expression data. These datasets form the foundation for our phenotype-specific models and enable systematic evaluation of how gene expression data impacts variant prioritisation.

Seed gene selection by Phen2Gene

Building phenotype-specific models requires lists of genes associated with target phenotypes. To enable direct comparison with VARPP [14], we began with the same list of 1,879 HPO terms used in that study. These terms were originally selected using Phenolyzer to map HPO terms to genes, in combination with ClinVar (2019 release) and dbNSFP (version 3.4a), retaining only terms associated with at least 25 genes with a ClinVar pathogenic variant (with a minor allele frequency less than 1% in five cohorts). In this study we re-evaluated those terms using the successor to Phenolyzer, Phen2Gene [23], to find genes associated with HPO terms. Phen2Gene is preferred over Phenolyzer due to substantial improvements in the gene-phenotype databases used in its construction, expanded use of gene–gene interaction databases, and methodological improvements, such as the incorporation of skewness-based weighting of HPO terms. Using Phen2Gene and updated databases, we identified 1,866 HPO terms from the original list that could still be used with the February 2022 release of HPO. We did not apply any additional filtering based on ontology depth, subtree, or redundancy. Phen2Gene associated an average of 297 genes with each HPO term. These genes allow us to select high-confidence pathogenic variants for each phenotype.

Pathogenic variants

Using the ClinVar database of clinically verified pathogenic and benign variants, we constructed a high-quality

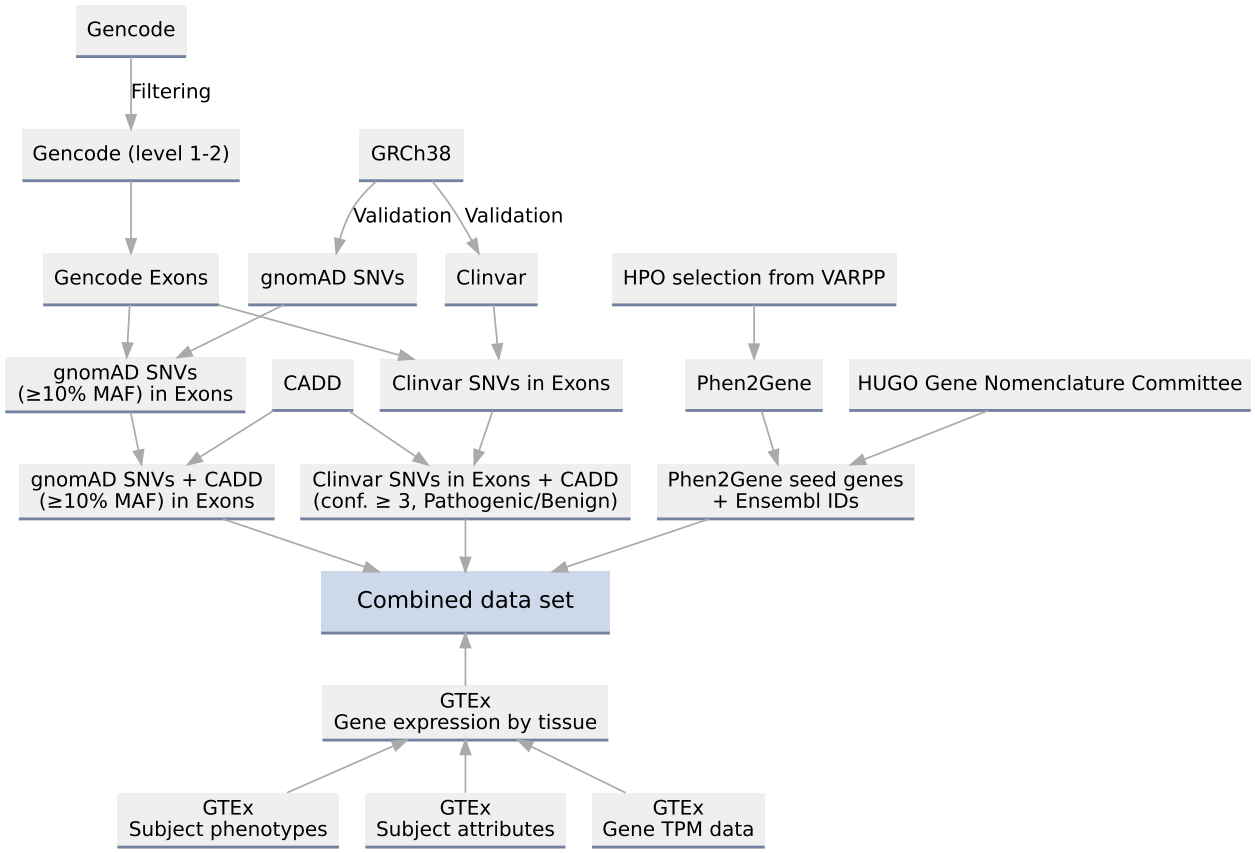

Fig. 1 Integration of data sources. The workflow for creating the training data for IMPPROVE models, when using GTEx expression data

set of known pathogenic variants. We downloaded the February 2022 release of ClinVar, and filtered it to SNVs with "Pathogenic" clinical significance and a 3–5 star review confidence of either "practice guideline", "reviewed by expert panel", or "criteria provided, multiple submitters, no conflicts". After filtering, we were left with 9479 high-confidence pathogenic variants across 1450 genes. This corresponds to a median of 547 variants, or 59 genes, per HPO term. Each pathogenic variant was then paired with CADD (version 1.6) phred scores, as a general indicator of pathogenicity. This process yielded a set of high-quality pathogenic variants linked to specific phenotypes identified by Phen2Gene.

Benign variants

While ClinVar is an excellent source of pathogenic variants, we were concerned that it would not provide a representative selection of benign variants. We thus considered three sources for high-quality benign variants: ClinVar [6] high-confidence benign variants, The Single Nucleotide Polymorphism Database (dbSNP) [27], and gnomAD [28]. From ClinVar, we selected high-confidence benign variants (at least three stars). With dbSNP (build 153) and gnomAD (version 3.1.1), we selected SNVs that occur in at least 10% of one population as we can be reasonably sure they are not responsible for a rare disease. To establish whether these sources of benign variants were comparable as control data for pathogenicity predicting models, we trained and tested preliminary Random Forest models using each of the three benign variant sources. We found that the choice of benign variant source impacted model predictions, evaluated according to the $\pi\alpha^{-1}5{,}000$ model criteria (see 2.3.2.). We observed stark differences in CADD score distributions among ClinVar, dbSNP, and gnomAD benign variants (see Supplementary Fig. 2). On this basis, we selected gnomAD as the source for benign variants in this study. For the subsequent stage of data integration, we need each variant to be associated with a gene. This is already the case for ClinVar, but not for gnomAD. We assigned each gnomAD variant to a GENCODE gene by intersecting the variant coordinates with GENCODE v26 exons of confidence one or two, yielding 301,373 benign variants. As with the pathogenic variants, each benign variant was paired with its CADD phred score.

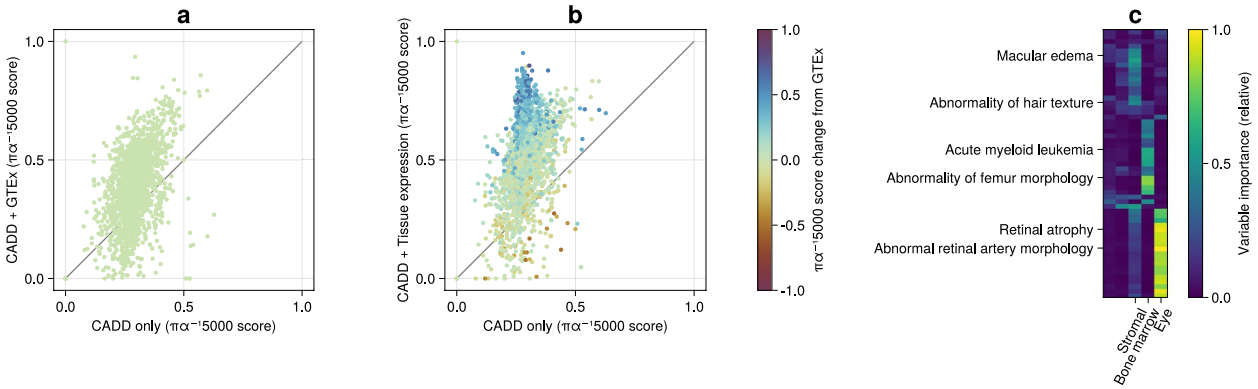

Fig. 2 Single-cell data enhances model predictive power. Comparison of pathogenic variant prediction performance using bulk and single-cell expression data against raw CADD scores, and the most important variables of selected phenotypes. **a** Comparison of raw CADD predictive performance to IMPPROVE models using GTEx bulk expression data. Points above the diagonal (78%) represent phenotypes with pathogenic variants where predictions are more accurately made by GTEx models as measured by the $\pi\alpha^{-1}5{,}000$ metric. **b** Comparison of raw CADD predictive performance to IMPPROVE models using single-cell tissue pseudobulk (from the Tabula Sapiens dataset). Phenotypes above the diagonal (90%) have pathogenic variants better predicted by single-cell models. Moving from a to b, some phenotypes exhibit a particularly large improvement compared to the GTEx model scores. In particular, there are 180 phenotypes whose model's $\pi\alpha^{-1}5{,}000$ score more than doubles to at least 0.5. These phenotypes constitute the "blue horn" seen in the upper left quadrant. There are also 28 phenotypes whose model's $\pi\alpha^{-1}5{,}000$ score drops by at least 0.2, which can be identified as the brown dots around the bottom of the plot. **c** A heatmap of the variable importance for 57 selected phenotypes of the 180 "blue horn" models identified in b as much improved, split into variables in common with GTEx, and those unique to Tabula Sapiens. The five variables with the greatest mean importance are shown, with the top three labeled by their tissue. Six particular phenotypes are also picked out, as examples

Gene expression data

The GTEx version 8 dataset [25] provides a large catalogue of bulk gene expression data from 838 donors across 49 tissue types, with a total of 17,382 samples. We downloaded the Transcripts Per Million (TPM) frequency data, and aggregated expression data by tissue type according to the sample ID. The tissue-specific expression was aggregated by calculating the median expression and standard deviation for each gene across the tissue samples, providing measures of both central tendency and variability. Each benign and pathogenic variant was then associated with the expression profile of its assigned gene. The expression data introduces a biological context, locations where a gene is actively transcribed, which we anticipated would improve the quality of pathogenicity predictions.

Tabula Sapiens provides a second expression dataset, an atlas of scRNA-seq reads. We downloaded 29 tissue-specific and one combined single-cell dataset as h5ad files, and used Muon.jl [29] to extract the DecontX [30] processed 10X count data. We normalised the counts for each cell by dividing by the number of Unique Molecular Identifiers (UMIs) recorded. Central tendency was estimated by calculating the mean expression as the sum of normalised counts divided by the number of cells. Variability was assessed using the standard deviation calculated directly from the counts, and dispersion was estimated via the marginal MLE of a Negative Binomial [31]. We repeated this process by tissue rather than cell-type to produce an analogue to the GTEx bulk RNA-seq dataset.

With these gene expression datasets, we then create per-phenotype training datasets by combining the CADD score of the selected pathogenic and benign variants with the corresponding gene expression data from the selected gene expression dataset.

Model training

(Re)sampling scheme for model training

When training and testing models, there are some key statistical assumptions implicit in the methodology that must not be violated. We avoided potential circularity with ClinVar by selecting CADD scores, but we also need to consider the statistical independence of observations when constructing Random Forests with bootstrapping. This is required for the evaluation of the out-of-bag sample to actually be unbiased. Since multiple variants can occur in the same gene, variants can share the same expression data and are thus not independent. More specifically, a gene's expression data can be mirrored across the testing and training groups. This necessitates a custom sampling scheme to avoid splitting variants of the same gene across the in-bag and out-of-bag samples. Additionally, the training dataset is highly imbalanced, with a median benign to pathogenic variant ratio of

600:1. Both of these complications can be addressed by modifying the sampling method used.

We developed a flexible stratified bootstrapping approach to resolve these issues. Stratifying the data by gene category, then imposing a maximum ratio of 50:1 of genes containing only benign and pathogenic variants, preserves the statistical independence of observations. After selecting the in-bag sample of genes with this capped imbalance, a random variant from each gene is selected. The out-of-bag sample is then constructed as all remaining variants from all other genes. After each iteration of this custom sampling scheme, a single model (here, a decision tree) is trained. We can then fairly assess how well the model predicts pathogenic variants.

Performance evaluation of models, the $\pi\alpha^{-1}5{,}000$ score

Appropriate evaluation of model performance is crucial for effectively judging IMPPROVE models. Common metrics like area under the ROC (AUROC) and area under the PRC (AUPRC) assume balanced data or representative class imbalance, which is not the case in our study. Previous work [14] addressed this by counting the number of true positives in the top hundred results. This approach simulates a clinical scenario where a geneticist examines a prioritised list of variants and assesses top-ranked predictions. However, as datasets evolve with varying numbers of true positive entries in each dataset, we sought a metric that adjusts to these differences and enables more consistent comparison across models and phenotypes.

Variant prioritisation in clinical scenarios often involves identifying a single pathogenic in a pool of a few thousand benign variants [32]. To reflect this, we consider the probability of correctly identifying a pathogenic gene when the threshold is set to limit the false positive rate to no more than one in n variants. This can be framed as finding the maximum statistical power (π) on the Receiver Operator Characteristic (ROC) curve such that the false positive rate (α) is no more than $1/n$. For convenience, this is denoted as $\pi\alpha^{-1}n$, with $\pi\alpha^{-1}5{,}000$ used throughout this paper.

Random forest hyperparameter tuning

To optimise the performance of Random Forests, we tuned several key hyperparameters: the bootstrap ratio, maximum feature count, tree depth, and tree count. Using a randomly selected subset of 200 HPO terms, of which 186 were viable (see Supplementary Table 2), hundreds of permutations of hyperparameters were tested and compared. We took a sequential approach, initially considering a wide range of values for each parameter, and then revisiting the parameter later and testing a few values to make sure it was still optimal.

Bootstrap ratio

To construct decision trees effectively under the extreme class imbalance, we make use of the custom sampling scheme discussed above. To determine an appropriate maximum ratio between benign and pathogenic variants in the training data, we tried a range of values (1, 2, 3, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, unbounded). Larger ratios linearly increased training time but yielded diminishing returns in tree accuracy. A bootstrap maximum class size ratio of 50:1 was chosen because it produced comparable model performance to larger ratios and did so with a much faster training time.

Maximum number of features

Typically, selecting $\sqrt{\text{number of features}}$ is recommended [33]. However, in our case, many variables were expected to be uninformative, making a larger value more appropriate. We tested the full range of possible selected features (or "mtry") values, and calculated the five-fold train and test error rates. Based on these results (Supplementary Fig. 6), we selected $\frac{2}{3}$number of features, as by this point the $\pi\alpha^{-1}5{,}000$ score increased by half compared to the $\sqrt{n}$ value. Beyond this point, the score increases only marginally.

Tree depth

In Random Forests, the maximum depth of decision trees can be tuned to limit their complexity. We examined both the $\pi\alpha^{-1}5{,}000$ and AUPRC out-of-bag (OOB) scores as tree depth increased to identify an effective tree depth (Supplementary Fig. 7). The $\pi\alpha^{-1}5{,}000$ score increased with depth, but beyond a depth of four the improvement was marginal. The AUPRC peaked with a depth of three or four. We thus selected a tree depth of four for this study.

Number of trees

Random forests do not overfit as the number of trees increases [33]. Therefore, we aimed to determine the minimum number of trees needed to produce stable results. To estimate this, we examined both the $\pi\alpha^{-1}5{,}000$ (Supplementary Fig. 8) and AUPRC scores and noted that by 500 trees, the variance in the $\pi\alpha{-}15{,}000$ score was less than 1% for 98% of the HPO terms (the AUPRC varied by less than 1% for 99% of the terms). Based on these results, 500 trees are sufficient, but to be conservative we used 1000 trees in our analysis.

Performance evaluation

Comparison to the baseline

To determine the contribution of gene expression information to model performance, we compared our models against the baseline predictive score, CADD. This

baseline is the single non-expression variable provided. It represents the predictive power of variant prioritisation without incorporating phenotype-specific gene expression information. For each selected model, we assess its performance using $\pi\alpha^{-1}5{,}000$ and directly compare this to the $\pi\alpha^{-1}5{,}000$ score achieved when CADD is used as the sole prioritisation criterion. This allows us to quantify the improvement from incorporating expression data into our models.

Variable importance

The impact of particular variables can be evaluated using permutation feature importance and SHAP (SHapley Additive exPlanations). Permutation importance is simpler but is susceptible to correlated features, feature interactions, and may produce biased estimates due to the number of feature splits. SHAP values offer a unified measure of feature importance and provide insights into how each feature impacts the model's predictions [34]. By breaking down the contribution of each feature to the prediction, SHAP values aid in understanding the model's behaviour and the underlying relationships between features and the outcome. This ability is particularly useful in our study to evaluate the impact of biological features used by the models. We first selected the top ten variables using permutation importance, then calculated their SHAP values based on the training data.

Spike-in predictions

In addition to considering the $\pi\alpha^{-1}5{,}000$ score of models, we also conducted a series of spike-in experiments. As part of these experiments, we selected a subset of models to compare against CADD. These models were selected by setting a minimum $\pi\alpha^{-1}5{,}000$ score, and applied to related variants according to the phenotype-gene associations identified by Phen2Gene. This simulates a best-case clinical scenario, where the patient has been precisely and accurately phenotyped.

Results

Single-cell data improves rare disease variant prioritisation

To determine how much bulk expression data enhances variant prioritisation, and whether single-cell RNA-seq data enhances performance, we assessed the performance of phenotype-specific models using both kinds of expression data. The value of bulk gene expression data in phenotype-specific models has been previously established [14], but since publication the breadth and quality of relevant datasets has markedly improved. ClinVar's August 2018 release (used by VARPP) contained 31,672 known pathogenic variants, and by February 2022 this had expanded to 46,330 variants. Concurrently, the GTEx bulk-tissue expression dataset [25] has undergone significant enhancements, doubling the number of available RNA-Seq samples and adding an additional tissue. We also used CADD release 1.6 instead of 1.4, which fixed GERP annotations and improved splice predictions. To assess the benefit of a particular data source for variant prioritisation, we scored the predictions of its models for each phenotype and then counted the number of phenotypes for which the models using expression data scored higher than those using CADD alone.

We first compare our results against VARPP [14], which uses earlier versions of our bulk data sources, by employing their "pp100" metric introduced therein (the proportion of true pathogenic variants in the top 100 predictions). Across the 1,866 phenotypes, we observed an improvement in 87% of the models using the updated GTEx gene expression data, a 37% increase compared to prior results in VARPP. Consistent with these pp100 results, we found that 78% of the models scored higher (Fig. 2a) according to the $\pi\alpha^{-1}5{,}000$ metric (the number of true positives when the false positive rate is no more than 1 in 5000). This reaffirms the value of gene expression data to phenotype-specific variant prioritisation models, and demonstrates the substantial benefit of newer and more comprehensive datasets.

We evaluated the predictive performance of models using both bulk and single-cell gene expression data. To facilitate a direct comparison with bulk RNA-seq data, we averaged single-cell gene expression values across cell types and tissues, as described in the Methods section. When using Tabula Sapiens single-cell expression data aggregated by tissue, models achieved a higher $\pi\alpha^{-1}5{,}000$ score than CADD across 90% of the phenotypes (Fig. 2b). This is a small improvement over the 87% of GTEx predictions that outperformed CADD; directly comparing Tabula Sapiens tissue pseudobulk models to GTEx we found that 76% of the models scored higher using Tabula Sapiens. Single-cell expression data generally lead to more accurate predictions.

Using mean expression by cell-type instead of tissue, we found that 86% of the models outperformed CADD, and 67% outperformed GTEx. Within the 76% of Tabula Sapiens tissue models that outperformed GTEx and the 24% that did not, there were two smaller groups that stood out. We found 28 phenotypes where the $\pi\alpha^{-1}5{,}000$ score dropped by at least 0.2. These phenotypes had significantly fewer associated genes ($p < 10^{-8}$ using a Mann–Whitney U test, with 65 associated on average compared to an average of 297 overall), but no other commonalities were identified. We also found 180 models whose $\pi\alpha^{-1}5{,}000$ score more than doubled to at least 0.5 with the use of single-cell data. Inspecting these models, we saw biologically plausible associations with the most important variables (ranked by permutation

Table 1 Tabula Sapiens model performance, comparing variant prediction performance to CADD using the $\pi\alpha^{-1}5000$ metric. These are the same four groups of models seen in Fig. 3

Aggregation	Statistic	Proportion that beat CADD
Tissue	Mean	89%
Tissue	Dispersion	89%
Cell-type	Mean	86%
Cell-type	Dispersion	84%

importance). In 57 of these models, the most important tissue is Eye, stromal, or bone marrow, all of which are absent in GTEx (Fig. 2c). Even when excluding these models from the comparison, the predictions using single-cell data are significantly improved compared to models using GTEx ($p < 10^{-30}$, using an approximate Wilcoxon signed rank test). In cases where a relevant tissue not part of the bulk expression data is included, the presence of the tissue drives the performance improvement. There is no overall advantage to single cell in this regard though, Tabula Sapiens has 15 tissues that GTEx does not, but it is also missing 16 tissues present in GTEx (Supplementary Table 1). Together with the performance comparisons, this shows that the incorporation of single-cell data significantly enhances model performance for most phenotypes.

Aggregation scale and statistics affect single-cell model performance

Having established that single-cell data benefits variant predictions, we wanted to better understand how prediction quality is affected by the choice of method and tissue or cell-type aggregation of single-cell data. To assess this, we conducted four experiments aggregating single-cell data either by tissue or cell-type, using either mean expression or dispersion. We first compare models using Tabula Sapiens tissue pseudobulk against Tabula Sapiens cell-type pseudobulk using mean expression, then using dispersion, then Tabula Sapiens tissue pseudobulk using mean expression vs. dispersion, and finally Tabula Sapiens cell-type pseudobulk using mean expression vs. dispersion. We once again examined the proportion of phenotypes whose model scored higher than CADD alone according to the $\pi\alpha^{-1}5{,}000$ score. We observed comparable overall performance across the four groups of models (Table 1). Directly comparing the models, we found 83 models (57 unique models, covering 55 phenotypes, with 26 duplicated across comparisons) which lay within the outer third of the score comparison plots, representing a difference of at least 0.423 (Fig. 3; purple dots). Of the 57 models, 28 performed much better using cell-type aggregation, with no clear commonalities, summarised by either mean expression or dispersion (see Supplementary Table 4 for a list of the 55 phenotypes, and their model's performance). This reveals that the choice of aggregation method can affect specific phenotypes to a much greater extent than the overall performance would indicate, with the average score differing by no more than 0.05.

We find that model performance can be highly dependent on the choice of tissue or cell-type aggregation. When comparing the models using mean expression by aggregation scale, the tissue and cell-type models had a mean $\pi\alpha^{-1}5{,}000$ score of 0.5 and 0.48 respectively, with a Pearson correlation coefficient of 0.83. There are 19 that lie in the outer third of the comparison plot (Fig. 3a). Four of them performed much better with cell-type aggregation, and 15 that performed much better with tissue aggregation. For models using dispersion, both tissue and cell-type models had a mean $\pi\alpha^{-1}5{,}000$ score of 0.46, with a Pearson correlation of 0.76. In the comparison plot (Fig. 3d), 21 models fell in the outer third. Of these, 14 performed much better with cell-type aggregation, and seven performed much better with tissue aggregation. This contrasts with the mean expression results, where twice as many models performed better using cell-type aggregation, underscoring the importance of cell-type variability in certain phenotypes. These comparisons indicate that neither tissue not cell-type aggregation is consistently better, and the best kind of model varies across phenotypes. We use this information to select the model that is the best pathogenic variant predictor on a per-phenotype basis.

Comparing the use of summary statistic: mean expression or dispersion, with models aggregated by both tissue and cell-type (Fig. 3b and Fig. 3c), we observe a similar pattern to our comparison of tissue and cell-type (Fig. 3a). Tissue models had a mean $\pi\alpha^{-1}5{,}000$ score of 0.50 using expression and 0.46 using dispersion respectively, with a Pearson correlation coefficient of 0.72. The cell-type models had a mean $\pi\alpha^{-1}5{,}000$ score of 0.56 and 0.48 using mean expression and dispersion respectively, with a Pearson correlation coefficient of 0.88. Both comparisons revealed particular phenotypes that are dramatically affected by the choice of summary statistic, lying in the outer third of the plot (a difference in $\pi\alpha^{-1}5{,}000$ score of at least 0.423). For tissue models, 36 phenotypes fell into this group, with 29 favouring mean expression and seven favouring dispersion. In cell-type models, only seven phenotypes were strongly affected, with three performing better with mean expression and four with dispersion. These results suggest that while mean expression generally results in superior model performance, in cases

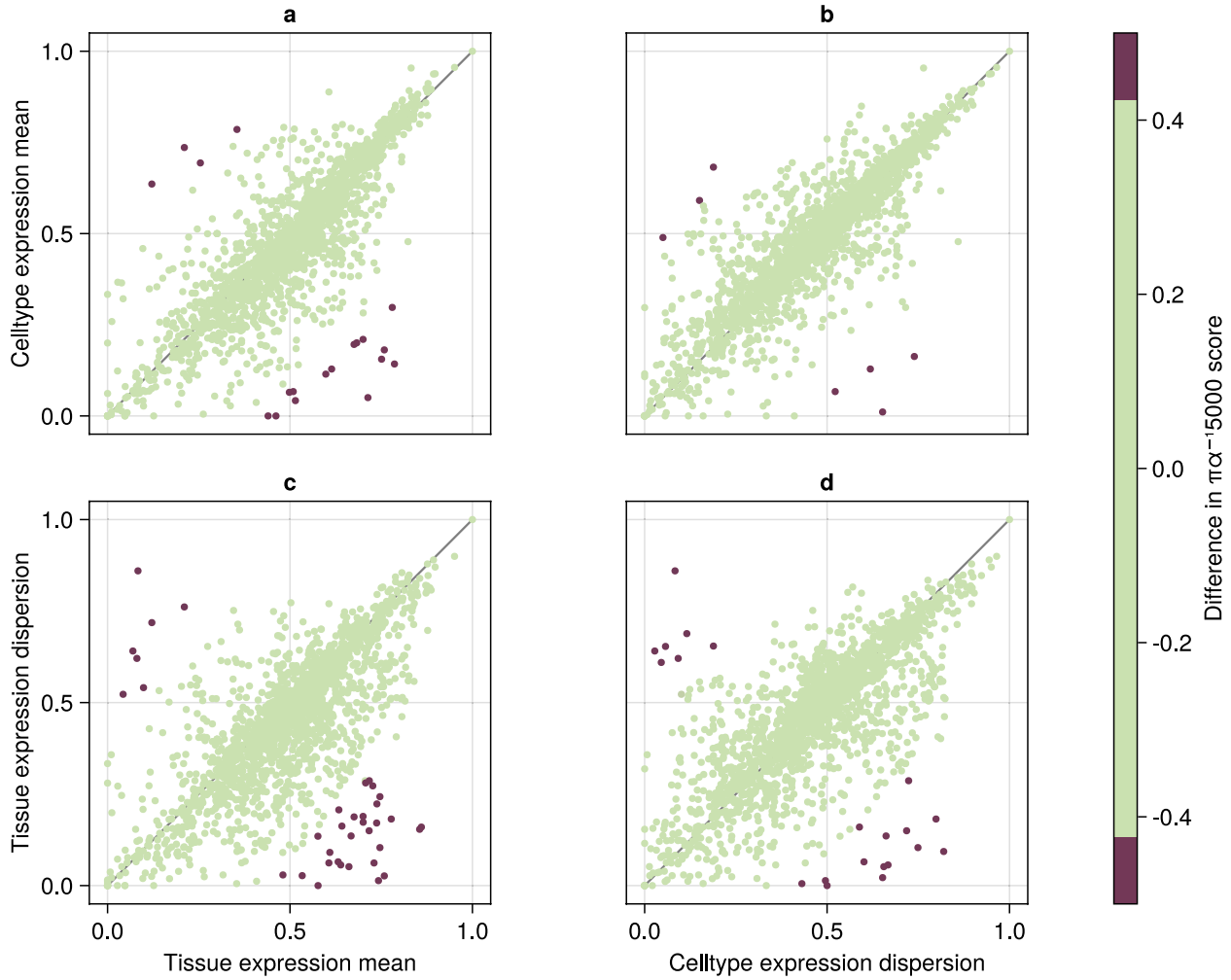

Fig. 3 Predictive performance can vary greatly by expression source on a per-model basis. Scatter plots of model scores using different scales of single-cell result aggregation, and different summary statistics. The performance of some phenotypes' models differ enough to place them in the outer third of their comparison plot (in the form of a triangle in the top left and bottom right corner), with a $\pi\alpha^{-1}5{,}000$ score that changes by at least 0.42. We consider these phenotypes' models to be much higher scoring with one dataset than the other, and highlight them in purple. **a**-**d** Comparisons of model performance across expression datasets, illustrating data-dependent variability in phenotype scoring

where variation in expression is informative, dispersion can offer an advantage.

Known and novel disease–cell type associations uncovered by IMPPROVE

We will use the following acronyms when discussing the four Tabula Sapiens model and expression datasets: Tabula Sapiens tissue (pseudobulk) mean expression (TS-TME), Tabula Sapiens tissue dispersion of expression (TS-TDE), Tabula Sapiens cell-type (pseudobulk) mean expression (TS-CME), and Tabula Sapiens cell-type dispersion of expression (TS-CDE).

Having identified 83 models whose performance strongly depends on the particular summary statistic and choice of tissue or cell-type aggregation, we sought a deeper understanding of the factors driving this variation in performance. For these models, we hypothesised that the variables with the greatest impact on the model's predictions could be connected to the underlying biology of the model's associated phenotype. Across the models we examined, CADD (the only direct measure of variant deleteriousness) consistently featured as the most important variable. Examining the top ten variables of the 83 models, we made qualitative assessments of the strength of the association between the top variables and known disease pathology using our knowledge and basic literature searches. Models that passed this qualitative assessment were investigated in greater depth, and considered

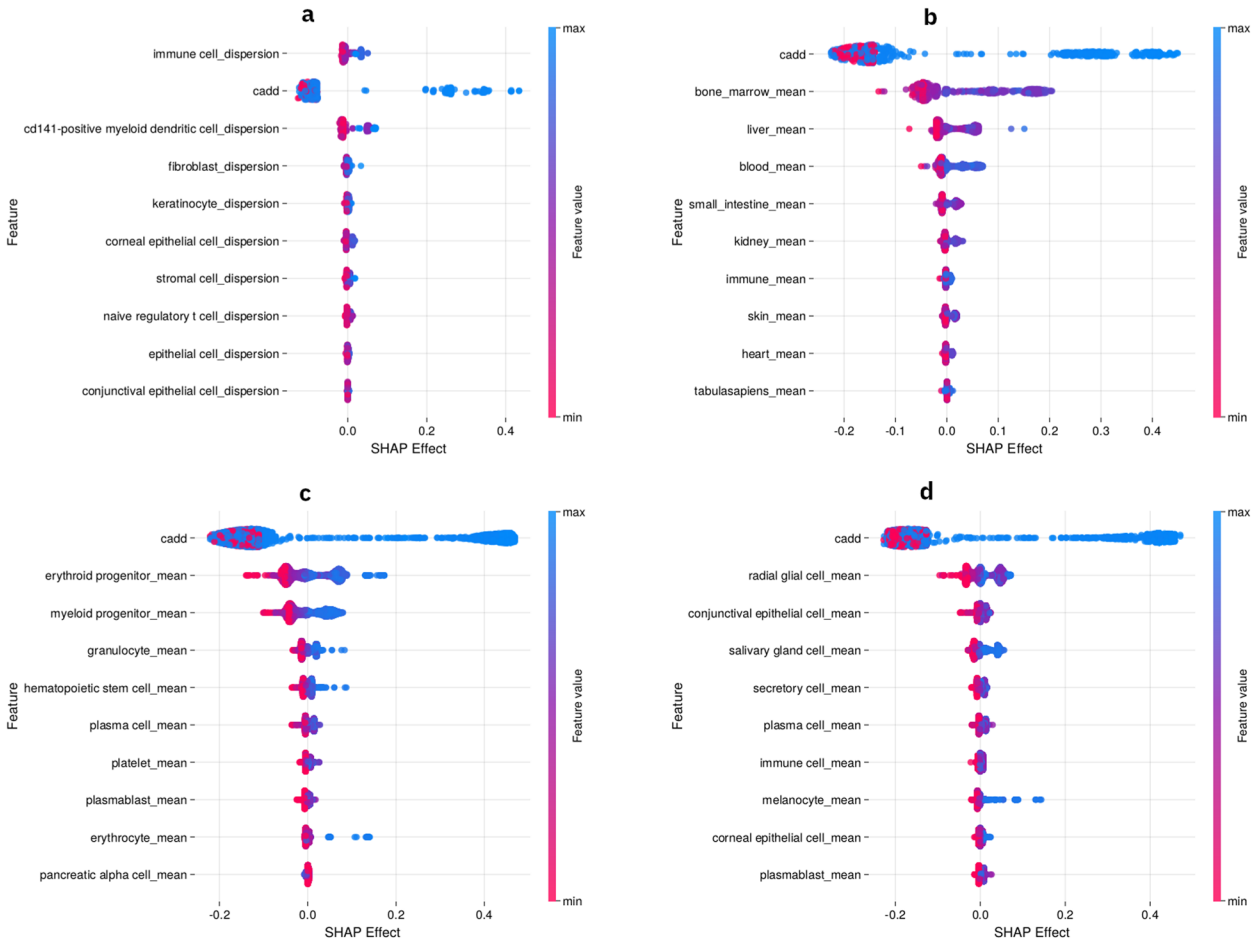

Fig. 4 Effect of individual variables on model predictions. Effects of the top ten features (by permutation importance) of the four example models chosen. **a** Psoriasiform dermatitis (HP:0003765) **b** Abnormality of the common coagulation pathway (HP:0010990) c Myelodysplasia (HP:0002863) d Squamous cell carcinoma (HP:0002860)

for presentation based on the expression dataset used and recorded points of interest.

We selected four illustrative cases: Psoriasiform dermatitis (HP:0003765), Abnormality of the common coagulation pathway (HP:0010990), Squamous cell carcinoma (HP:0002860), and Myelodysplasia (HP:0002863). Each model represents a unique aspect of the Tabula Sapiens expression datasets, showcasing the framework's ability to integrate different types of gene expression data effectively. The SHAP effects of these models (Fig. 4) demonstrate how an interpretable model using gene expression data, such as IMPPROVE, can uncover biologically meaningful associations between disease pathology and model variables.

Psoriasiform dermatitis is characterised by inflammation and redness of the skin, along with thickened red skin that has flaky, silver-white patches [35]. The TS-CDE model identified key cell types that align with known immune and skin-related pathologies. With a relatively high $\pi\alpha^{-1}$5,000 score of 0.65, the TS-CDE model scored more than thirty times higher than all other models in Table 2. Examining its SHAP effect plot (Fig. 4a), we note that the top three cell-types are immune cells, CD141 +myeloid dendritic cells, and keratinocytes. These cell-types align with the known pathology of psoriasis, an autoimmune condition involving abnormal

Table 2 The $\pi\alpha^{-1}$5000 model scores of each Tabula Sapiens based model for the four example terms selected. The particular models discussed further are marked set in bold

HPO term	TS-TME	TS-TDE	TS-CME	TS-CDE
Psoriasiform dermatitis	0.011	0.022	0.011	**0.652**
Abnormality of the common coagulation pathway	**0.769**	0.026	0.181	0.197
Squamous cell carcinoma	0.778	0.182	**0.830**	0.799
Myelodysplasia	0.355	0.093	**0.786**	0.820

differentiation of keratinocytes [36]. Moreover, CD141 + cells are specifically involved in skin homeostasis and affect skin inflammation [37]. Notably, dispersion of expression in naïve regulatory T-cells emerged as the eighth most important variable in the model. Previous studies have found T-cells to be heavily involved in psoriasis, although average expression was not considered informative [38]. This finding suggests that for naïve regulatory T-cells, the variability or pattern of gene expression is more relevant to psoriasiform dermatitis than the mean expression level. This exemplifies the value of using dispersion as a metric, particularly in identifying more complex associations between expression and disease. These observations, consistent with the known disease mechanisms of psoriasis, underscore the ability of our models to identify biological variables associated with disease aetiology.

Abnormality of the common coagulation pathway is a disruption to the series of protein activations that contribute to blood coagulation [35]. The TS-TME model identified both obvious and non-obvious tissues related to coagulation. Selected for its high $\pi\alpha^{-1}5{,}000$ score of 0.77, the TS-TME model scored around four times higher than the other three models in Table 2. The SHAP effect plot (Fig. 4b), lists bone marrow, liver, and blood as the top three tissues, all of which are associated with aspects of coagulation. Furthermore, the small intestine, ranked as the fourth most important tissue, plays a crucial role in the absorption of vitamin K [39], and vitamin K deficiency prevents normal coagulation. The fifth and sixth top tissues, kidney and skin, are also associated with the coagulation pathway in certain kidney and skin diseases [40–42]. These associations again demonstrate the connection between model variables and disease aetiology.

Squamous cell carcinoma is a malignant tumor of the skin's squamous epithelium [35]. The TS-CME model identified melanocytes and skin-related tissues, well-established in the disease's pathology, alongside less common tissues such as salivary glands and ocular cells. While all models except TS-TDE scored highly (Table 2), the TS-CME model performed best, with a high $\pi\alpha^{-1}5{,}000$ score of 0.83. In the SHAP effect plot (Fig. 4d), despite ranking eighth in the ordering, melanocytes had the largest SHAP effect (other than the CADD score). The influence of melanocytes is unsurprising given the basic pathology of squamous cell carcinoma (SCC). Interestingly, some of the other cell-types have indirect relations to SCC pathologies. For example, although rare, SCC can occur in the salivary glands [43] (the third most important tissue) and in the eye [44] (relating to the second and ninth tissues: conjunctival and corneal epithelial cells). While not all top variables are directly linked to the disease itself, the presence of tissues like ocular and salivary cells may indicate underlying biological associations yet to be fully understood. Alternatively, this finding could reflect the model's sensitivity to cell-type variability in tissues where SCC might develop in rare cases, highlighting both the potential and limitations of machine learning predictions in identifying known and potential associations. To distinguish between true biological associations and artefacts of the model, further experiments such as transcriptomic profiling of SCC in the identified tissues would be needed.

Myelodysplasia is characterised by ineffective production within one or more hematopoietic cell lineages, resulting in anemia and cytopenia [35]. The TS-CME model prioritised blood and bone marrow progenitor cells, crucial in the pathology of the myelodysplasia, while highlighting a less well-studied relationship with pancreatic cells. In terms of performance, both the cell-type based motels (TS-CME and TS-CDE) substantially outperform the tissue-based models (TS-TME and TS-TDE, see Table 2). Among the tissue models, the mean expression model outperformed the dispersion model by 0.26, while the two cell-type models only differed by 0.03. Given these results, we selected the TS-CME model, balancing the superior performance of the cell-type models with the better performance of mean expression models overall. In the TS-CME model's SHAP effect plot (Fig. 4c), the top cell-types are erythroid progenitors and myeloid progenitors, which are highly relevant to myelodysplasia. Myelodysplasia is a cancerous clonal stem cell disorder characterised by ineffective hematopoiesis, disrupting the maturation of blood cells. It is also associated with the function of the spleen and lymph nodes [45]. An analysis of erythroid dysplasia in myelodysplastic syndromes found the percentage of CD117 erythroid progenitor cells, together with two other factors, provided the best discrimination between myelodysplastic syndromes and non-clonal cytopenia [46]. Low-risk myelodysplastic subtypes are reportedly frequently characterised by an expansion of common myeloid progenitors [47]. The next six cell-types are all found in peripheral blood and affected by myelodysplasia [48]. The tenth cell type, pancreatic alpha cells, was unexpected as they are not typically associated with the pathology of myelodysplasia. However, there is evidence suggesting a potential link between myelodysplastic syndrome and both acute and autoimmune pancreatitis, which could explain this association [49, 50]. The TS-CME model's identification of relevant cell types demonstrates the benefit of incorporating gene expression data at the cell-type scale. In this case, while the TS-TME model was able to identify related tissues (bone marrow, spleen, lymph node, and blood), it lacked the specificity of the TS-CME model. The capability of our models to capture less obvious

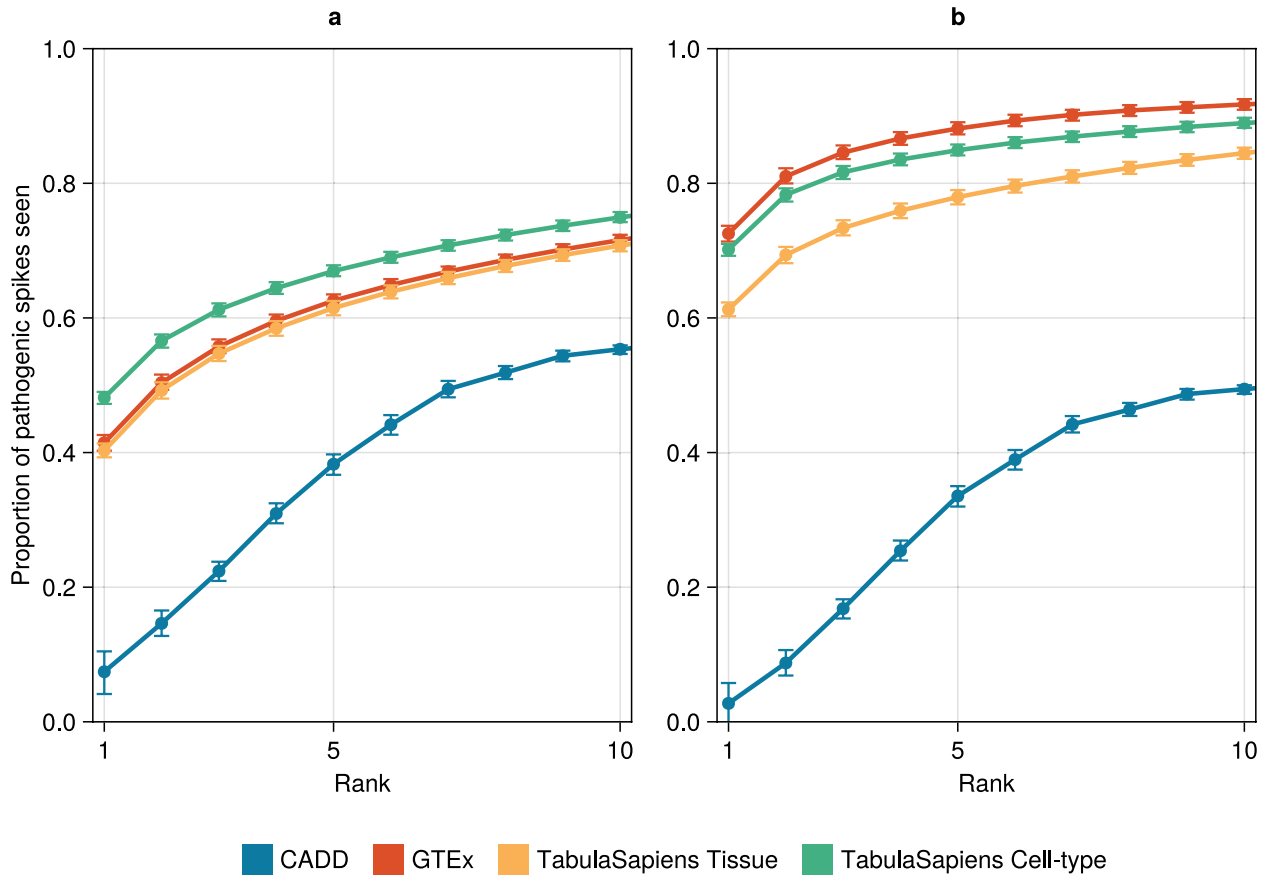

Fig. 5 Comparing spike-in ranks to CADD (cumulatively). Cumulative rank of pathogenic variants introduced to an otherwise healthy genome, using new pathogenic variants in the March 2023 ClinVar release (compared against the February 2022 release). **a** Predictions for all 1866 phenotypes on the new variants, regardless of their models' scores on the training data, with a 95% confidence interval indicated by error bars. **b** Predictions of models that had a $\pi\alpha^{-1}5{,}000$ score above 0.8, with a 95% confidence interval indicated by error bars. This includes 6 GTEx models, 58 TS-TME models, 67 TS-CME models, and the CADD-only predictions for all included variants

associations suggests that our models could serve as hypothesis generators, identifying potential biological connections that warrant further investigation.

Performance validation with spike-in variants

To assess the predictive performance of the models in a clinical context, we conducted a series of spike-in tests, simulating real-world variant prioritisation by introducing known pathogenic variants into healthy genomes. Healthy genomes were sourced from the 1000 Genomes project [51], with 100 individuals selected at random. Pathogenic variants were sourced from the March 2023 release of ClinVar, excluding variants present in the February 2022 release that had been used to train the models. We used the Phen2Gene data as a gene-phenotype association list, and assigned each variant to all phenotypes that listed the variant's gene. This approach allowed us to test the models' ability to accurately prioritise previously unseen variants, providing a more comprehensive evaluation of their clinical utility than the training-based $\pi\alpha^{-1}5{,}000$ statistic alone.

In total, we used 2,945 new high-confidence pathogenic variants, across 1,206 genes. For all HPO terms associated with each new variant, we applied the relevant model and recorded the final rank of the pathogenic variant. In total, 308 million spike-in experiments were run. This gave us sufficient data to investigate how well our models generalise to unseen variants, using the self-evaluation scores to select high-performing models.

In an unfiltered evaluation of all models on new ClinVar variants, we observe a 5.5–8.2 fold increase in the top-two rank prediction rate compared to CADD alone, with cell-type pseudobulk models performing the best overall (Fig. 5a). While this unfiltered analysis demonstrates the general utility of incorporating expression data and value of single-cell information, a key advantage of our approach is that we can leverage the per-model training performance score ($\pi\alpha^{-1}5{,}000$) as an a priori

indicator of model reliability. This enables the identification of models that we expect to perform well or poorly for a given phenotype, helping guide practical usage.

To examine how a priori information can aid model selection, we considered $\pi\alpha^{-1}5{,}000$ score thresholds from zero to one, in steps of 0.05, and found a score of 0.8 to be a stringent threshold for selecting high-performing models. All filtered phenotype-specific models using gene expression substantially outperform CADD across the top ten predictions (Fig. 5). The pathogenic variant ranked in the top five predictions in 88% of the GTEx predictions, 85% of the TS-CME predictions, and 78% of the TS-TME predictions. Across all the predicted variants, the CADD score alone only placed the pathogenic in the top five 34% of the time. More impressive is the frequency with which the pathogenic variant was ranked in the top two predictions: 73%, 70%, and 61% of the GTEx, TS-CME, and TS-TME predictions respectively, compared to only 9% of the CADD predictions. While GTEx models performed the best in this experiment, it is important to note that only six GTEx models actually passed the $\pi\alpha^{-1}5{,}000$ score threshold. By comparison, approximately ten times as many single-cell models passed this threshold, with 58 TS-TME models and 67 TS-CME models. This translated into a three-fold increase in the number of variants that single-cell models could be applied to.

These results should be interpreted with two key limitations in mind. Firstly, this analysis assumes precise and accurate phenotyping, a scenario that is rare in clinical settings where phenotypes are often broader or ambiguous. Secondly, by restricting our analysis to high-performing models, we only cover 689 (11%) of the new ClinVar variants. Despite these limitations, the results demonstrate the utility of per-phenotype model self-evaluation scores in ensuring that only models likely to generalise well are applied to unseen variants. This trade-off reflects a realistic clinical consideration: clinicians are likely to trust and rely on models only when they are known to perform well. In this context, single-cell models are far more frequently usable, and all expression-informed models markedly outperform CADD.

Discussion

Enhanced variant prioritisation with single-cell gene expression

Our results support the hypothesis that integrating single-cell gene expression data into phenotype-specific models enhances both the accuracy and interpretability of pathogenic variant prioritisation. While previous work has shown that phenotype-specific models and bulk gene expression data improve variant prioritisation [7, 9, 14, 52], the benefit of single-cell data, capturing variation within and across tissues and cell-types, has not been explored.

The improvement in variant prioritisation using single-cell expression data was particularly pronounced for phenotypes that exhibited tissue-specific expression patterns. While bulk sequencing has proven valuable in rare disease diagnosis [53], the value of single-cell data remains relatively uncharted [5]. Our study reveals that single-cell data provides an additional layer of biological context that enhances predictive power. Specifically, a clear majority of models using single-cell expression aggregated by cell-type or tissue outperformed bulk gene expression data (GTEx). Reassessing the performance of models using more recent versions of the datasets used in [14], we observed improvements in just over a third of the models. This suggests that single-cell data is a valuable biological context and indicates that model performance is likely to advance further with the continued expansion and refinement of relevant datasets.

While single-cell-derived expression data generally outperforms bulk gene expression, the results also revealed considerable variation between datasets and across phenotypes. No single dataset consistently outperformed others; model performance was strongly affected by the phenotype and dataset used. For instance, phenotypes associated with ocular tissues showed reduced accuracy when using Tabula Sapiens data, which lacks eye samples. The absence of key tissues or cell types relevant to a disease in the data diminished the model's ability to accurately prioritise variants based on expression patterns was diminished. Conversely, certain cases benefitted greatly from information on specific tissues or features. Specifically considering the Tabula Sapiens dataset, 28 phenotypes were identified that performed much better when using cell-type information, unattainable through bulk sequencing alone. This highlights the importance of comprehensive coverage when using expression data for accurate predictions. Ongoing initiatives, such as the Human Cell Atlas (HCA) [54], are critical for filling these gaps and enhancing the utility of single-cell RNA-seq in pathogenicity prediction. Models with insufficient training data, which perform poorly, can be identified and excluded from clinical use, improving the reliability of predictions. The effectiveness of this approach was seen in the spike-in test, where a stringent threshold was applied and the models that passed outperformed CADD in predicting the pathogenic variant, ranking it first or second around seven times as often.

Moreover, our findings suggest that not only mean expression levels but highly variable gene expression patterns within a tissue or cell-type are important for identifying causative variants in certain phenotypes. This was seen in our comparison of summarisation statistic,

between mean expression and dispersion. While mean expression generally provided robust predictions, we found that for certain phenotypes, such as psoriasis dermatitis, dispersion-based metrics offered unique advantages. Dispersion captures the variability in gene expression within a population of cells, something that cannot be measured in bulk expression data, but that can be particularly valuable for conditions involving high levels of cellular heterogeneity, such as immune-related conditions [55]. This provides an opportunity to pick the model most appropriate for each phenotype, exploring which summary statistics best capture the relevant biology. These insights can guide future refinement and application of variant prioritisation models, ensuring that they are specific to the biology of each condition.

Uncovering biological connections with interpretable models

Using the performance metrics generated during training, such as the (OOB) $\pi\alpha^{-1}5{,}000$ score, we can select highly accurate models that have likely inferred relevant biology. This is possible thanks to the combination of biologically relevant features and an interpretable model design. To assess the practical implications of this, we considered both an unfiltered scenario and a filtered scenario, the latter only using models passing a conservative score threshold ($\pi\alpha^{-1}500 \geq 0.8$). This filtering substantially increased pathogenic ranking precision, at the cost of reduced coverage. Single-cell models more often met the performance threshold than bulk models (3% of models compared to 0.3%), making them more broadly useful in this setting. We expect the number of high-quality models to grow as more and higher quality expression and variant data becomes available.

Beyond established aetiology, our investigation of four high-performing models in "Known and novel disease–cell type associations uncovered by IMPPROVE" found novel associations between particular phenotypes and cell types. Our SHAP-based analysis suggested that variants in highly dynamically regulated genes in naive regulatory T-cells are associated with psoriasiform dermatitis, and that pancreatic alpha cells are linked to myleodysplasia. These results demonstrate that IMPPROVE can identify unexpected yet biologically plausible connections that warrant further investigation, thereby contributing to research on rare disease pathology. Such capabilities highlight the potential of interpretable models to advance our understanding of disease mechanisms and guide future research.

Limitations and future directions

While encouraging, our study has several limitations that should be addressed in future research. The bulk and single-cell RNA-seq datasets examined here cover 30 and 28 different tissues respectively. We investigated the difference in tissues covered and identified phenotypes where the presence of specific tissue is important to model performance. Critically, we found associations between these specific tissues and underlying biology of diseases related to the phenotypes. Emerging large single-cell datasets offer more comprehensive coverage of the human body. We anticipate that repeating our work with these new datasets will offer further insights, improved model performance, and better characterisation of the contribution of single-cell data to variant pathogenicity investigations.

Our models also rely on the quality of the ClinVar database. The database relies on submissions from clinicians and researchers, with a wide range in the confidence and evidence behind particular variants. We account for this by stringently selecting high-confidence ClinVar variants for our work. As the quality and coverage of the ClinVar database improves, our work will benefit from improved accuracy and generalisability.

We selected Random Forests as the machine learning algorithm in this work, for their balance of interpretability and performance. This has been an effective choice in our analysis, but other per-phenotype models using expression data could be explored in future work.

Single-cell data provides insight into expression dynamics at the cell-type level. Moreover, gene dispersion is known to be heritable and associated with gene function and patient phenotypes [21]. These factors suggest that more sophisticated approaches leveraging single-cell data could uncover additional phenotypes that benefit from this level of detail.

By identifying pathogenic variants more effectively than CADD alone and enabling biological insights, IMPPROVE-like methods could enhance the diagnostic process, leading to more accurate and timely treatments for a wide range of conditions. The per-model scores let us assess whether cell-type or tissue expression information is more helpful, and revert to standard approaches if neither score is satisfactory.

Conclusions

Our study demonstrates that the IMPPROVE framework improves predictive accuracy for rare disease prioritisation from a baseline pathogenicity score, by incorporating single-cell and bulk gene expression data, achieving a 52–64% increase in top-two rank prediction rates in best case scenarios. Compared to existing methods, it offers enhanced precision and interpretability, particularly for phenotypes with tissue-specific expression patterns.

Furthermore, the integration of multiple expression datasets, including single-cell RNA-seq data, reveals biological patterns related to the pathology of certain diseases. This capacity for biological insight benefits both the prioritisation of pathogenic variants and provides a foundation for further research into the underlying biological mechanisms of rare diseases.

Abbreviations

AUROC	Area under the ROC
AUPRC	Area under the PRC
CADD	Combined Annotation Dependent Depletion
DECIPHER	DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
DbSNP	The Single Nucleotide Polymorphism Database
GTEx	Genotype-Tissue Expression
GnomAD	Genome Aggregation Database
HPO	Human Phenotype Ontology
HCA	Human Cell Atlas
OOB	Out-of-bag
OMIM	Online Mendelian Inheritance in Man
ROC	Receiver Operator Characteristic
SNV	Single Nucleotide Variant
TS-TME	Tabula Sapiens tissue (pseudobulk) mean expression
TS-TDE	Tabula Sapiens tissue dispersion of expression
TS-CME	Tabula Sapiens cell-type (pseudobulk) mean expression
TS-CDE	Tabula Sapiens cell-type dispersion of expression
TPM	Transcripts Per Million
UMI	Unique Molecular Identifier
WGS	Whole Genome Sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11711-w.

Supplementary Material 1.

Acknowledgements
We would like to thank Kate Farley and Kevin Chen for providing critical feedback.

Authors' contributions
Conceptualization: TL. Methodology: TC, TL. Software: TC. Validation: TC. Formal analysis: TC. Investigation: TC. Resources: TC, TL. Data curation: TC. Writing original draft: TC, TL. Writing, review and editing: all authors. Visualization: TC. Supervision: TL. Project Administration: TL. Funding acquisition: TL.

Funding
TL was supported by fellowships from the Feilman Foundation and the Stan Perron Foundation. The work was funded in part by the Australian Government's Medical Research Future Fund (2007567).

Data availability
The data and models used in this study, as well as the container that runs them and a tool to do so can be downloaded as an 84 GiB bundle from https://doi.org/https://doi.org/10.26182/37wm-bz23. The IMPPROVE project container source code can be found at https://github.com/tecosaur/impprove_container.
Project home page: https://github.com/tecosaur/impprove_container
Operating system(s): POSIX (Linux, MacOS, BSD).
Programming languages: Julia, Python, Shell.
Other requirements: Downloaded data, Docker/Podman/Apptainer container runtime.
License: MIT.Known and novel disease–cell type associations uncovered by IMPPROVE
The IMPPROVE container together with the required data, the models used in this study, and a tool to apply them to new VCFs can be downloaded as an 84 GiB bundle from https://doi.org/https://doi.org/10.26182/37wm-bz23.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 January 2025 Accepted: 14 May 2025
Published online: 28 May 2025

References

1. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A. Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. Eur J Hum Genet. 2020;28:165–73.
2. Zurynski Y, Deverell M, Dalkeith T, Johnson S, Christodoulou J, Leonard H, Elliott EJ, APSU Rare Diseases Impacts on Families Study group. Australian children living with rare diseases: Experiences of diagnosis and perceived consequences of diagnostic delays. Orphanet J Rare Dis. 2017;12:68.
3. Yang G, Cintina I, Pariser A, Oehrlein E, Sullivan J, Kennedy A. The national economic burden of rare disease in the United States in 2019. Orphanet J Rare Dis. 2022;17:163.
4. Nisar H, Wajid B, Shahid S, Anwar F, Wajid I, Khatoon A, Sattar MU, Sadaf S. Whole-genome sequencing as a first-tier diagnostic framework for rare genetic diseases. Exp Biol Med. 2021;246:2610.
5. Marwaha S, Knowles JW, Ashley EA. A guide for the diagnosis of rare and undiagnosed disease: Beyond the exome. Genome medicine. 2022;14:23.
6. Landrum MJ, Lee JM, Benson M, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic acids res. 2018;46:D1062–7.
7. Kumar AA, Van Laer L, Alaerts M, Ardeshirdavani A, Moreau Y, Laukens K, Loeys B, Vandeweyer G. pBRIT: Gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. Bioinformatics. 2018;34:2254–62.
8. Hu J, Lepore R, Dobson RJB, Al-Chalabi AM, Bean D, Iacoangeli A. DGLinker: Flexible knowledge-graph prediction of disease–gene associations. Nucleic Acids Res. 2021;49:W153–61.
9. Jacobsen JOB, Kelly C, Cipriani V, Research Consortium GE, Mungall CJ, Reese J, Danis D, Robinson PN, Smedley D Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease. Human mutation. https://doi.org/10.1002/humu.24380
10. Sifrim A, Popovic D, Tranchevent L-C, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y. eXtasy: Variant prioritization by genomic data fusion. Nat methods. 2013;10:1083–4.
11. Javed A, Agrawal S, Ng PC. Phen-Gen: Combining phenotype and genotype to analyze rare disorders. Nat methods. 2014;11:935–7.
12. Smedley D, Jacobsen JOB, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat protoc. 2015;10:2004–15.
13. Birgmeier J, Haeussler M, Deisseroth CA, et al (2020) AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Science translational medicine. 10/gphh28
14. Anderson D, Baynam G, Blackwell JM, Lassmann T. Personalised analytics for rare disease diagnostics. Nat commun. 2019;10:5274.
15. Boudellioua I, Kulmanov M, Schofield PN, Gkoutos GV, Hoehndorf R. DeepPVP: Phenotype-based prioritization of causative variants using deep learning. BMC Bioinformatics. 2019;20:65.
16. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction

from sequence by integrating long-range interactions. Nat methods. 2021;18:1196–203.
17. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. Am J Human Genet. 2016;99:877–85.
18. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med. 2021;13:153.
19. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat rev drug discov. 2019;18:463–77.
20. Visibelli A, Roncaglia B, Spiga O, Santucci A. The impact of artificial intelligence in the odyssey of rare diseases. Biomedicines. 2023;11:887.
21. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, Holmes C. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat Biotechnol. 2013;31:748–52.
22. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
23. Zhao M, Havrilla JM, Fang L, et al. Phen2Gene: Rapid phenotype-driven gene prioritization for rare diseases. Nar Genomics Bioinformatics. 2020;2:lqaa032.
24. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, Bruford EA. Genenames.org: The HGNC and VGNC resources in 2021. Nucleic Acids Res. 2021;49:D939–46.
25. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
26. The Tabula Sapiens Consortium, Jones RC, Karkanias J, et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science. 2022. https://doi.org/10.1126/science.abl4896.
27. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: The NCBI database of genetic variation. Nucleic acids res. 2001;29:308–11.
28. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, Consortium GAD, Rehm HL, MacArthur DG, O'Donnell-Luria A. Variant interpretation using population databases: Lessons from gnomAD. Human Mutation. 2022;43:1012–30.
29. Bredikhin D, Kats I, Stegle O. MUON: Multimodal omics analysis framework. Genome Biol. 2022;23:42.
30. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, Campbell JD. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21:57.
31. León-Novelo L, Fuentes C, Emerson S. Marginal likelihood estimation of negative binomial parameters with applications to RNA-seq data. Biostatistics. 2017;18:637–50.
32. Pedersen BS, Brown JM, Dashnow H, et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. NPJ Genom Med. 2021;6:1–8.
33. Breiman L. Random Forests. Machine learning. 2001;45:5–32.
34. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp 4768–4777
35. Köhler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021. Nucleic Acids Res. 2021;49:D1207–17.
36. Kamata M, Tada Y. Dendritic cells and macrophages in the pathogenesis of Psoriasis. Front immunol. 2022;13:941071.
37. Chu C-C, Ali N, Karagiannis P, et al. Resident CD141 (BDCA3)+ dendritic cells in human skin produce IL-10 and induce regulatory T cells that suppress skin inflammation. J exp med. 2012;209:935–45.
38. Owczarczyk-Saczonek A, Czerwińska J, Placek W (2018) The role of regulatory T cells and anti-inflammatory cytokines in psoriasis. Acta dermatovenerologica alpina pannonica et adriatica. https://doi.org/10.15570/actaapa.2018.4
39. Yamanashi Y, Takada T, Kurauchi R, Tanaka Y, Komine T, Suzuki H. Transporters for the intestinal absorption of cholesterol, vitamin e, and vitamin k. J atheroscler thromb. 2017;24:347–59.
40. Madhusudhan T, Kerlin BA, Isermann B. The emerging role of coagulation proteases in kidney disease. Nat rev nephrol. 2016;12:94–109.
41. Qiu Z, Pang X, Xiang Q, Cui Y. The crosstalk between nephropathy and coagulation disorder: pathogenesis, treatment, and dilemmas. J Am Soc Nephrol. 2023;34:1793.
42. Cugno M, Tedeschi A, Crosti C, Marzano AV. Activation of blood coagulation in autoimmune skin disorders. Expert Rev Clin Immunol. 2009;5:605–13.
43. Young A, Okuyemi OT (2023) Malignant Salivary Gland Tumors. StatPearls
44. Gichuhi S, Sagoo MS. Squamous cell carcinoma of the conjunctiva. Community eye health. 2016;29:52–3.
45. Kraus MD, Bartlett NL, Fleming MD, Dorfman DM. Splenic pathology in myelodysplasia: A report of 13 cases with clinical correlation. Am j surg pathol. 1998;22:1255–66.
46. Westers TM, Cremers EMP, Oelschlaegel U, et al. Immunophenotypic analysis of erythroid dysplasia in myelodysplastic syndromes. A report from the IMDSFlow working group. Haematologica. 2017;102:308–19.
47. Will B, Zhou L, Vogler TO, et al. Stem and progenitor cells in myelodysplastic syndromes show aberrant stage-specific expansion and harbor genetic and epigenetic alterations. Blood. 2012;120:2076–86.
48. Shaver AC, Seegmiller AC. Nuances of morphology in myelodysplastic diseases in the age of molecular diagnostics. Curr hematol malig rep. 2017;12:448–54.
49. Tanvetyanon T, Stiff P. Recurrent steroid-responsive pancreatitis associated with myelodysplastic syndrome and transformations. Leuk Lymphoma. 2005;46:151–4.
50. Tabata R, Tabata C, Okamoto T, Omori K, Terada M, Nagai T. Autoimmune pancreatitis associated with myelodysplastic syndrome. Int arch allergy immunol. 2010;151:168–72.
51. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
52. Godard P, Page M. PCAN: Phenotype consensus analysis to support disease-gene association. BMC Bioinformatics. 2016;17:518.
53. Turro E, Astle WJ, Megy K, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583:96–102.
54. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, Lein E, Teichmann SA. Cell type ontologies of the Human Cell Atlas. Nat cell biol. 2021;23:1129–35.
55. Satija R, Shalek AK. Heterogeneity in immune responses: From populations to single cells. Trends Immunol. 2014;35:219–29.

Publisher's Note