Cell type-specific contextualisation of the phenomic landscape: a comprehensive and scalable approach towards the diagnosis, prognosis and treatment of all rare diseases

Brian M. Schilder[aff-1*], Kitty B. Murphy[aff-1], Robert Gordon-Smith[aff-1], Jai Chapman[aff-1], Momoko Otani[aff-1], Nathan G. Skene[aff-1*]

[aff-1], Imperial College London.

*Corresponding author(s). E-mail(s): brian_schilder@alumni.brown.edu; n.skene@imperial.ac.uk;

Keywords: rare disease, phenotype, single-cell, gene therapy

Abstract

Rare diseases (RDs) are an extremely heterogeneous and underserved category of medical conditions. While the majority of RDs are strongly genetic, it remains largely unknown via which physiological mechanisms genetics cause RD. Therefore, we sought to systematically characterise the cell type-specific mechanisms underlying all RD phenotypes with a known genetic cause by leveraging the Human Phenotype Ontology and transcriptomic single-cell atlases of the entire human body from embryonic, foetal, and adult samples. In total we identified significant associations between 201 cell types and 9,575/11,028 (86.7%) unique phenotypes across 8,628 RDs. We estimate that this represents an over 500-fold increase in the collective knowledge of RD phenotype-cell type mechanisms.

Next, we demonstrated how these results may be used for personalised patient diagnosis and prognosis, as well as the development of novel therapeutics. Finally, we take a data-driven approach to highlight several of the most promising gene/cell therapy candidates with the highest probability of animal model-to-human patient translation. Furthermore, we have made these results entirely reproducible and freely accessible to the global community to maximise their impact. To summarise, this work represents a significant step forward in the mission to treat patients across an extremely diverse spectrum of serious RDs.

Introduction

While rare diseases (RDs) are individually uncommon, they collectively account for an enormous global disease burden with over 10,000 recognised RDs affecting at least 300-400 million people globally[1] (1 in 10-20 people)[2] . Over 75% of RDs primarily affect children with a 30% mortality rate by 5 years of age[3]. Despite the prevalence and severity of RDs, patients suffering from these conditions are vastly underserved due to several contributing factors. First, diagnosis is extremely challenging due to the highly variable clinical presentations of many of these diseases. The diagnostic odyssey can take patients and their families decades, with an average time to diagnosis of 5 years[4]. Of those, ~46% receive at least one incorrect diagnosis and over 75% of all patients never receive any diagnosis[5]. Second, prognosis is also made difficult by high variability in disease course and outcomes which makes matching patients with effective and timely treatment plans even more challenging. Finally, even for patients who receive an accurate diagnosis/prognosis, treatments are currently only available for less than 5% of all RDs[6]. In addition to the scientific challenges of understanding RDs, there are strong financial disincentives for pharmaceutical and biotechnology companies to develop expensive therapeutics for exceedingly small RD patient populations with little or no return on investment[7,8]. Those that have been produced are amongst the world's most expensive drugs, greatly limiting patients' ability to access it[9,10], The provision of timely, effective and affordable care for RD patients will require substantive transformations to our existing scientific, clinical, and regulatory frameworks.

A major challenge in both healthcare and scientific research is the scalable exchange of information. Even in the age of electronic healthcare records (EHR) much of the information about an individual's history is currently fractured across healthcare providers, often with differing nomenclatures for the same conditions. The Human Phenotype Ontology (HPO) is a hierarchically organised set of controlled clinical terms that provides a much needed common framework by which clinicians and researchers can precisely communicate patient conditions[14]. The HPO spans all domains of human physiology and currently describes 18082 phenotypes across 10,300 RDs. Each phenotype and disease is assigned its own unique identifier and organised as a hierarchical graph, such that higher-level terms describe broad phenotypic categories or *branches* (e.g. *HP:0033127*: 'Abnormality of the musculoskeletal system' which contains 4495 unique phenotypes) and lower-level terms describe increasingly precise phenotypes (e.g. *HP:0030675*: "Contracture of proximal interphalangeal joints of 2nd-5th fingers").

It has already been integrated into healthcare systems and clinical diagnostic tools around the world, with increasing adoption over time[11]. Common ontology-controlled frameworks like the HPO open a wealth of new opportunities, especially when addressing RDs. Services such as the Matchmaker Exchange[15,16] have enabled the discovery of hundreds of underlying genetic etiologies, and led to the diagnosis of many patients. This also opens the possibility of gathering cohorts of geographically dispersed patients to run clinical trials, the only viable option for treatment in many individuals. To further increase the number of individuals who qualify for these treatments, as well as the trial sample size, proposals have been made deviate from the traditional single-disease clinical trial model and instead perform basket trials on groups of RDs with shared molecular etiologies (SaME)[17]. However this approach, and indeed much of RD patient care, hinges upon first characterising the molecular mechanisms underlying each RD.

Over 80% of RDs have a known genetic cause[18,19]. Despite this our knowledge of the physiological mechanisms via which genetics cause pathogenesis is lacking for most RDs, severely hindering our ability to effectively diagnose, prognose and treat RD patients. The availability of standardised, ontology-controlled databases presents opportunities to systematically investigate RDs at scale. Since 2008, the HPO has been continuously updated using knowledge from the medical literature, as well as by integrating databases of expert validated gene-phenotype relationships, such as OMIM[20–22], Orphanet[23,24], and DECIPHER[25]. A subset of the HPO contains gene annotations for 11,047 phenotypes across 8,631 diseases. Yet genes alone do not tell the full story of how RDs come to be, as their expression and functional relevance varies drastically across the multitude of tissues and cell types contained within the human body.

Our knowledge of single-cell-resolution biology has exploded over the course of the last decade and a half, with numerous applications in both scientific and clinical practices[26–28]. More recently, comprehensive single-cell transcriptomic atlases across tissues have also emerged[29,30]. In particular, the Descartes Human[31] and Human Cell Landscape[32] projects provide comprehensive multi-system single-cell RNA-seq (scRNA-seq) atlases in embryonic, foetal, and adult human samples from across the human body. These datasets provide data-driven gene signatures for hundreds of cell subtypes. They also allow us to investigate disease mechanisms in the context of specific life stages.

Here, we combine and extend several of the most comprehensive genomic and transcriptomic resources currently available to systematically uncover the cell types underlying granular phenotypes across 8,628 diseases. We then go on to highlight thousands of novel phenotype-cell type associations which collectively expand our knowledge of cell type-resolved phenotypes by an estimated 567-fold. Next, we present several potential avenues for real world applications of these results in the context of RD patient diagnosis, prognosis, treatment, and therapeutics development.

Results

Phenotype-cell type associations

In this study we systematically investigated the cell types underlying phenotypes across the HPO. A summary of the phenome-wide results stratified by single-cell atlas can be found in **?@tbl-summary**. Within the results using the Descartes Human single-cell atlas, 19,929/ 848,078 (2.35%) tests across 77/ 77 (100%) cell types and 7,340/11,047 (66.4%) phenotypes revealed significant phenotype-cell type associations after multiple-testing correction ($FDR_{p,c} < 0.05$). Using the Human Cell Landscape single-cell atlas, 26,585/1,358,916 (1.96%) tests across 124/124 (100%) cell types and 9,049/11,047 (81.9%) phenotypes showed significant phenotype-cell type associations ($FDR_{p,c} < 0.05$). The median number of significantly associated phenotypes per cell type was 252 (Descartes Human) and 200 (Human Cell Landscape), respectively.

Across both single-cell references, the median number of significantly associated cell types per phenotype was 3, suggesting reasonable specificity of the testing strategy. 8,628/8,631 (~100%) of diseases within the HPO gene annotations showed significant cell type associations for at least one of their respective phenotypes.

Validation of expected phenotype-cell type relationships

Within each high-level branch in the HPO shown in Fig. 1b, we tested whether each cell type was more often associated with phenotypes in that branch relative to those in all other branches (including those not shown). We then checked whether each cell type was overrepresented (at $FDR_{b,c} < 0.05$) within its respective on-target HPO branch, where the number of phenotypes within that branch (N_p). Abnormality of the cardiovascular system: 5/6 types of 'cardiocyte' were overrepresented (N_p=673). Abnormality of the endocrine system: 3/4 types of 'endocrine cell' were overrepresented (N_p=291). Abnormality of the eye: 5/5 types of 'photoreceptor cell/retinal cell' were overrepresented (N_p=721). Abnormality of the immune system: 4/4 types of 'leukocyte' were overrepresented (N_p=255). Abnormality of the musculoskeletal system: 4/4 types of 'cell of skeletal muscle/chondrocyte' were overrepresented (N_p=2155). Abnormality of the nervous system: 19/23 types of 'neural cell' were overrepresented (N_p=1647). Abnormality of the respiratory system: 2/2 types of 'respiratory epithelial cell/epithelial cell of lung' were overrepresented (N_p=292)..

As an additional form of validation (Fig. 1d), we tested for a relationship between phenotype-cell type association significance ($-log_e(p_{p,c})$ where log_e denotes natural log and and $p_{p,c}$ denotes uncorrected phenotype-cell type association p-values) and the proportion of on-target cell types. The list of on-target cell types were determined by matching each high-level HPO branch to a corresponding CL branch. These cross-ontology mappings can be found in **?@tbl-celltypes**. For this analysis we used raw p-values ($p_{p,c}$) rather than multiple-testing corrected p-values ($FDR_{p,c}$) to provide a more dynamic range of values (as the latter can drive values to 1). All 7/7 high-level HPO branches showed a consistent upwards trend towards greater proportions of on-target cell types with increasing degrees of significance. Furthermore, all branches also

showed a proportion of on-target cell types above that expected by chance (baseline = on-target cell types / total cell types) at $-log_e(p_{p,c}) > 1$.

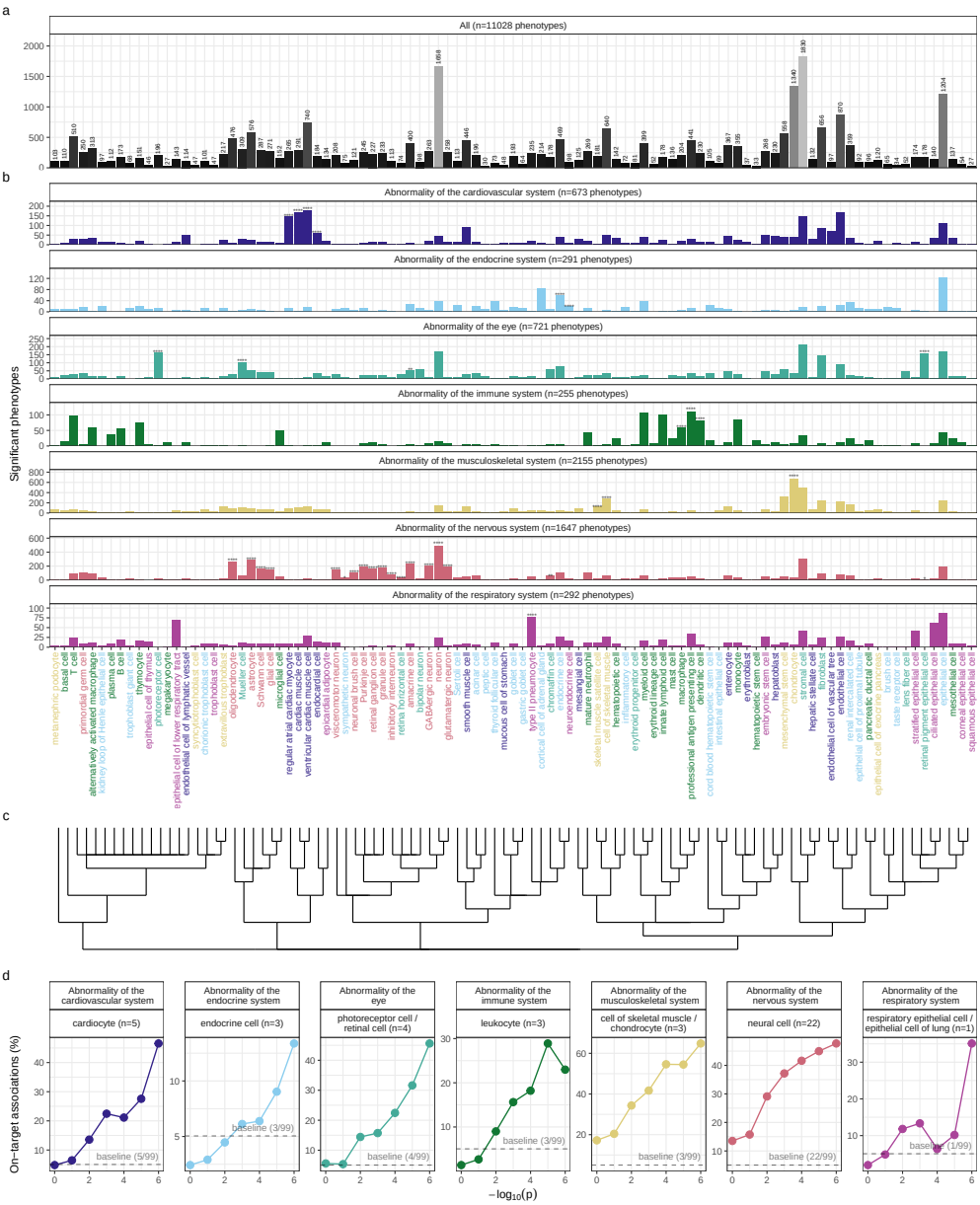


Figure 1: Summary of significant associations between phenotypes and cell types, aggregated by HPO branch. Here we show **a**, the total number of significant phenotype enrichments per cell type ($FDR_{p,c} < 0.05$) across all branches of the HPO. **b**, Number of phenotype association related to several high-level branches of the HPO. Asterisks above each bar indicate whether that cell type was significantly more often enriched in that branch relative to all other HPO branches, including those not shown here, as a proxy for how specifically that cell type is associated with that branch; $FDR_{b,c} < 1e-04$ (****), $FDR_{b,c} < 0.001$ (***), $FDR_{b,c} < 0.01$ (**), $FDR_{b,c} < 0.05$ (*). **c**, Dendrogram derived from the Cell Ontology (CL) showing the relatedness of all tested cell types to one another. For simplicity, cell type labels shown here are aligned to the CL[33] and can therefore encompass one or more cell types annotated by the original authors of scRNA-seq datasets[31,32]. **d**, Percentage of significant phenotype associations with on-target cell types (second row of facet labels), respective to the HPO branch. As significance increases ($-log_{10}(p)$ along the *x-axis*) the percentage of on-target enriched cell types also increases (*y-axis*).

Validation of inter- and intra-dataset consistency

Next, we sought to validate the consistency of our results across the two single-cell reference datasets (Descartes Human vs. Human Cell Landscape) across the subset of overlapping cell types Fig. 12. In total there were 142285 phenotype-cell type associations to compare across the two datasets (across 10945 phenotypes and 13 cell types annotated to the exact same CL term. We found that the correlation between p-values of the two datasets was high ($rho = 0.492, p = 1.08e{-}93$). Within the subset of results that were significant in both single-cell datasets ($FDR_{p,c} < 0.05$), we found that correlation of the association effect size were even stronger ($rho = 0.723, p = 1.08e-93$). We also checked for the intra-dataset consistency between the p-values of the foetal and adult samples in the Human Cell Landscape, showing a very similar degree of correlation as the inter-dataset comparison ($rho = 0.436, p = 2.36e-149$). Together, these results suggest that our approach to identifying phenotype-cell type associations is highly replicable and generalisable to new datasets.

More specific phenotypes are associated with fewer genes and cell types

First, we found that phenotype ontology showed a significant negative correlation with the number of genes annotated to that phenotype in the HPO data (Fig. 2a; $p = 2.23e-308, q = 2.23e-308, rho = -0.2634$). This is expected as broader phenotypes tend to have large gene set annotations. Next, we reasoned that lower HPO ontology levels representing more specific phenotypes were likely to be associated with fewer, more specific subsets of cell types. This was indeed the case, as we observed a strongly significant negative correlation between the two variables (Fig. 2b; $p = 2.23e{-}308, q = 2.23e{-}308, rho = -0.2927$). We also found that the effect size of significant phenotype-cell type associations ($FDR_{p,c} < 0.05$) increased with greater phenotype specificity, though the relationship was rather weak (Fig. 2c; $p = 7.30e{-}97, q = 7.30e{-}97, rho = 0.0966$). Finally, we found that the mean expression specificity of phenotype-associated genes (within the cell types significantly associated with those respective phenotypes at $FDR_{p,c} < 0.05$) was positively correlated phenotype ontology depth (Fig. 2d; $p = 2.71e-174, q = 3.61e-174, rho = 0.1398$).

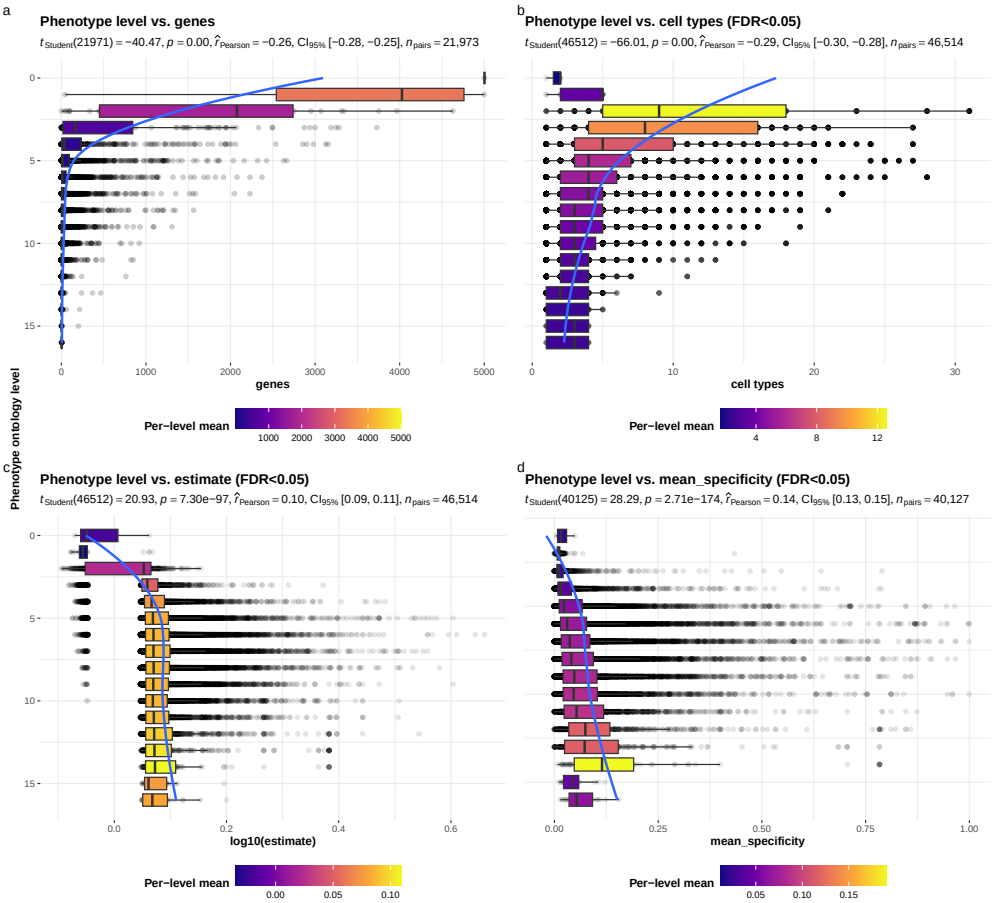


Figure 2: More specific phenotypes are associated with fewer, more specific genes and cell types. Box plots showing relationship between HPO phenotype level and **a**, the number of genes annotated to each phenotype, **b**, the number of significantly enriched cell types, **c**, the effect size of phenotype-cell type association tests at $FDR_{p,c} < 0.05$, and **d**, the mean expression specificity of phenotype-associated genes in the cell types significantly associated with those respective phenotypes ($FDR_{p,c} < 0.05$). Ontology level 0 represents the most inclusive HPO term 'All', while higher ontology levels (max=16) indicate progressively more specific HPO terms (e.g. 'Contracture of proximal interphalangeal joints of 2nd-5th fingers'). Boxes are coloured by the mean value (respective to the subplot) within each HPO level.

Hepatoblasts have a unique role in recurrent Neisserial infections

We selected the HPO term 'Recurrent bacterial infections' and all of its descendants (19 phenotypes) as an example of how investigations at the level of granular phenotypes can reveal different cell type-specific mechanisms (Fig. 3). As expected, these

phenotypes are primarily associated with immune cell types (e.g. macrophages, dendritic cells, T cells, monocytes, neutrophils). Some associations confirm relationships previously suggested in the literature, such as that between 'Recurrent staphylococcal infections' and myeloid cells[34–37]. Specifically, our results pinpoint monocytes as the most strongly associated cell subtypes ($FDR_{p,c} = 1.03e-30, B = 1.76e-01$).

In contrast to all other recurrent infection types, 'Recurrent Neisserial infections' highlighted a novel association with hepatoblasts (Descartes Human : $FDR_{p,c} = 1.13e-06, B = 8.24e-02$). Whilst unexpected, a convincing explanation involves the complement system, a key driver of innate immune response to Neisserial infections. Hepatocytes, which derive from hepatoblasts, produce the majority of complement proteins[38], and Kupffer cells express complement receptors[39]. In addition, individuals with deficits in complement are at high risk for Neisserial infections[40,41], and a genome-wide association study in those with a Neisserial infection identified risk variants within complement proteins[42] . While the potential of therapeutically targeting complement in RDs (including Neisserial infections) has been proposed previously[43,44], performing this in a gene- and cell type-specific manner may help to improve efficacy and reduce toxicity (e.g. due to off-target effects). Importantly, there are over 56 known genes within the complement system[45], highlighting the need for a systematic, evidence-based approach to identify effective gene targets.

Also of note, despite the fact that our datasets contain both hepatoblasts and their mature counterpart, hepatocytes, only the hepatoblasts showed this association. This suggests that the genetic factors that predispose individuals for risk of Neisserial infections are specifically affecting hepatoblasts before they become fully differentiated. It is also notable that these phenotypes were the only ones within the 'Recurrent bacterial infections' branch, or even the broader 'Recurrent infections' branch, perhaps indicating a unique role for hepatoblasts in recurrent infectious disease. The only phenotypes within the even broader 'Abnormality of the immune system' HPO branch that significantly associated with mature hepatocytes were 'Pancreatitis' ($FDR_{p,c} = 2.08e-02, B = 5.25e-02$) and 'Susceptibility to chickenpox' ($FDR_{p,c} = 1.20e-02, B = 5.49e-02$) both of which are well-known to involve the liver[46–48].

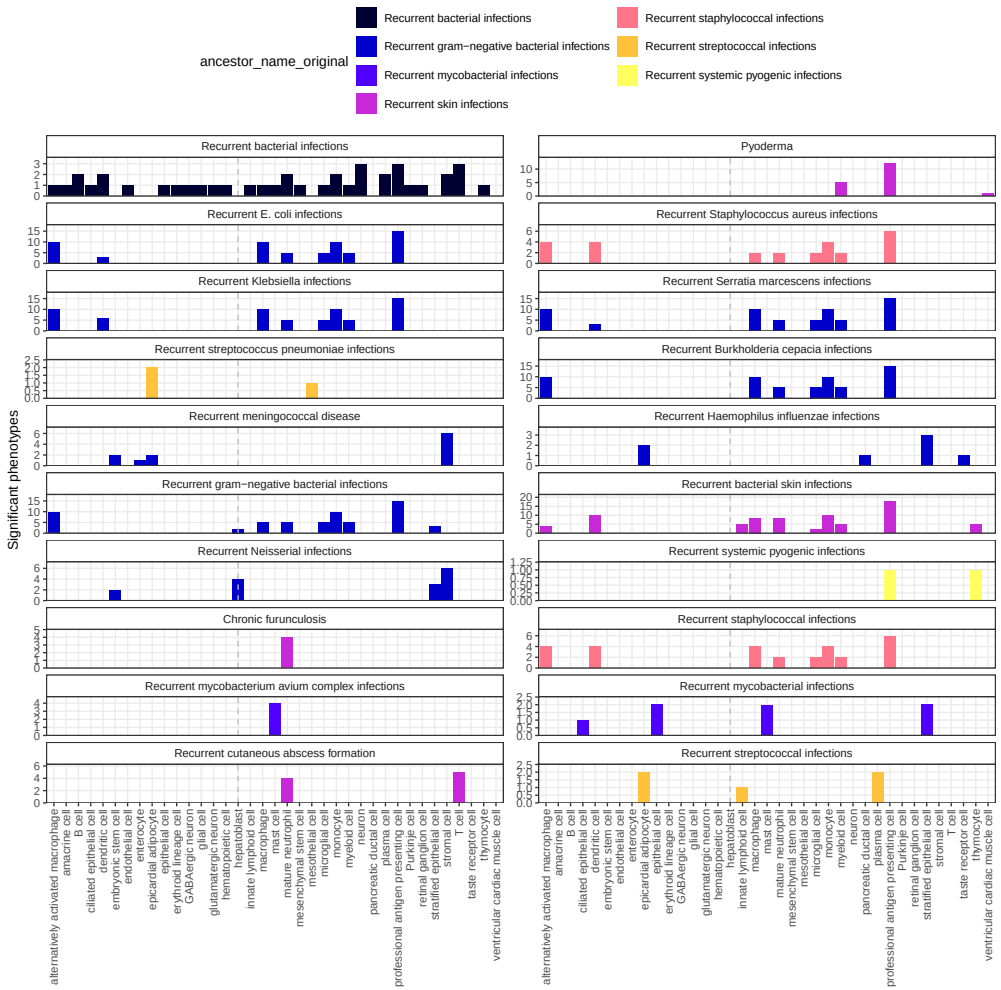


Figure 3: Hepatoblasts have a unique role in recurrent Neisserial infections. Significant phenotype-cell type tests for phenotypes within the branch 'Recurrent bacterial infections'. Amongst all different kinds of recurrent bacterial infections, hepatoblasts (highlighted by vertical dotted lines) are exclusively enriched in 'Recurrent gram−negative bacterial infections'. Note that terms from multiple levels of the same ontology branch are shown as separate facets (e.g. 'Recurrent bacterial infections' and 'Recurrent gram−negative bacterial infections').

Next, we sought to link multi-scale mechanisms at the levels of disease, phenotype, cell type, and gene and visualise these as a network (Fig. 4). This revealed that genetic deficiencies in different complement system genes (*C5*, *C8*, and *C7*) are primarily mediated by different cell types (hepatoblasts, stratified epithelial cells, and stromal cells, respectively). While genes of the complement system are expressed throughout

many different tissues and cell types, these results indicate that different subsets of these genes may mediate their effects through different cell types. This finding suggests that investigating (during diagnosis) and targeting (during treatment) different cell types may be critical for the diagnosis and treatment of these closely related, yet mechanistically distinct, diseases.

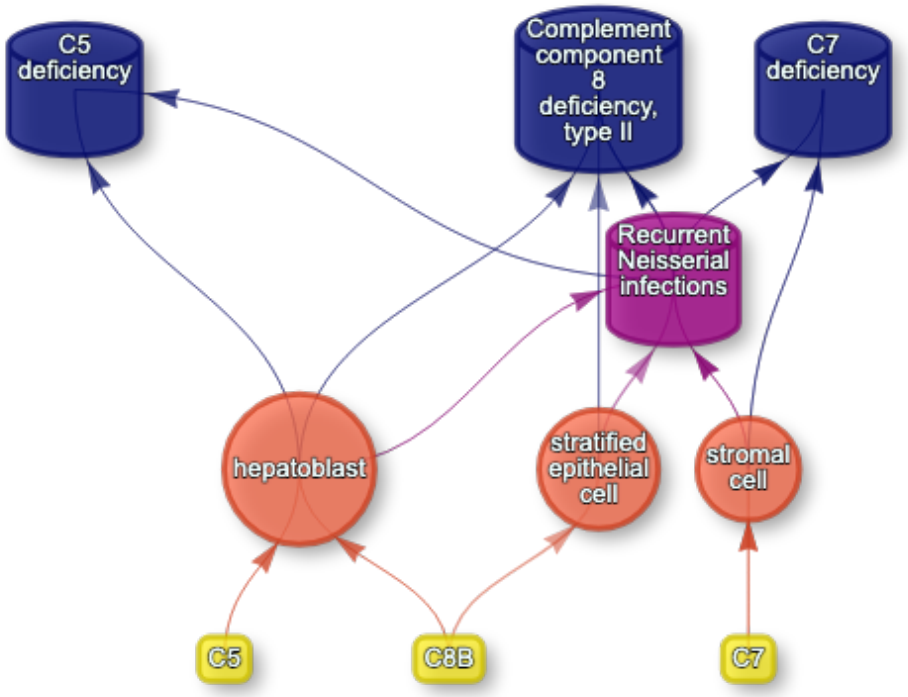


Figure 4: Multi-scale mechanisms of Recurrent Neisserial infections. Starting from the bottom of the plot, one can trace how causal genes (yellow boxes) mediate their effects through cell types (orange circles), phenotypes (pruple cylinders) and ultimately diseases (blue cylinders). Cell types are connected to phenotypes via association testing ($FDR_{p,c} < 0.05$), and to diseases when the symptom gene set overlap is >25%. Nodes were spatially arranged using the Sugiyama algorithm[49].

Monarch Knowledge Graph recall

Next, we used the Monarch Knowledge Graph (MKG) as a proxy for the field's current state of knowledge of phenotype-cell type associations. We evaluated the proportion of MKG associations that were recapitulation by our results. In total, our results contained at least one significant cell type associations for >90% of the phenotypes described in the MKG. Of these phenotypes, we captured >45% of the MKG

phenotype-cell associations when only considering exact overlap of CL-aligned cell type annotations. This proportion increased with greater flexibility in the matching of cell type annotations, reaching a maximum of **!!RECOMPUTE!!**% at a ontology graph distance of **!!RECOMPUTE!!** when considering the overlap of cell type annotations at the level of cell type ontology terms. This suggests that our results are in line with the current state of knowledge, and that our approach can be used to identify novel phenotype-cell type associations.

Annotation of phenotypes using generative large language models

Severity annotations were gathered from GPT-4 for 16982/18082 (93.9166%) HPO phenotypes. In our companion study, benchmarking tests of these results using ground-truth HPO branch annotations. For example, phenotypes within the 'Blindness' HPO branch (*HP:0000618*) were correctly annotated as causing blindness by GPT-4. Across all annotations, the recall rate of GPT-4 annotations was 91.26% (min=70.1%, max=100%, SD=11.84) with a mean consistency score of 91.21% (min=80.96%, max=97.48%, SD=5.739) for phenotypes whose annotation were collected more than once. This clearly demonstrates the ability of GPT-4 to accurately annotate phenotypes. This allowed us to begin using these annotations to compute systematically collected severity scores for all phenotypes in the HPO.

From these annotations we computed a weighted severity score metric for each phenotype ranging from 0-100 (100 being the theoretical maximum severity of a phenotype that always causes every annotation). Within our annotations, the most severe phenotype was 'Anencephaly' (*HP:0002323*) with a severity score of 58, followed by 'Atrophy/Degeneration affecting the central nervous system' (*HP:0007367*) with a severity score of 58. There were 677 phenotypes with a severity score of 0 (e.g. 'Thin toenail'). The mean severity score across all phenotypes was 14.89 (median=14, standard deviation=8.517).

Enrichment of foetal cell types in congenital phenotypes

The frequency of congenital onset with each phenotype (as determined by GPT-4 annotations) was strongly predictive with the proportion of significantly associated foetal cell types in our results ($p = 2e-203, \chi^2_{Pearson} = 940, \hat{V}_{Cramer} = 0.14$). Furthermore, increasing congenital frequency annotation (on an ordinal scale) corresponded to an increase in the proportion of foetal cell types: 'always'=24% (n=1636 associations), 'often'=20% (n=2979 associations), 'rarely'=12% (n=1956 associations), 'never'=10% (n=811 associations). This is consistent with the expected role of foetal cell types in development and the aetiology of congenital disorders.

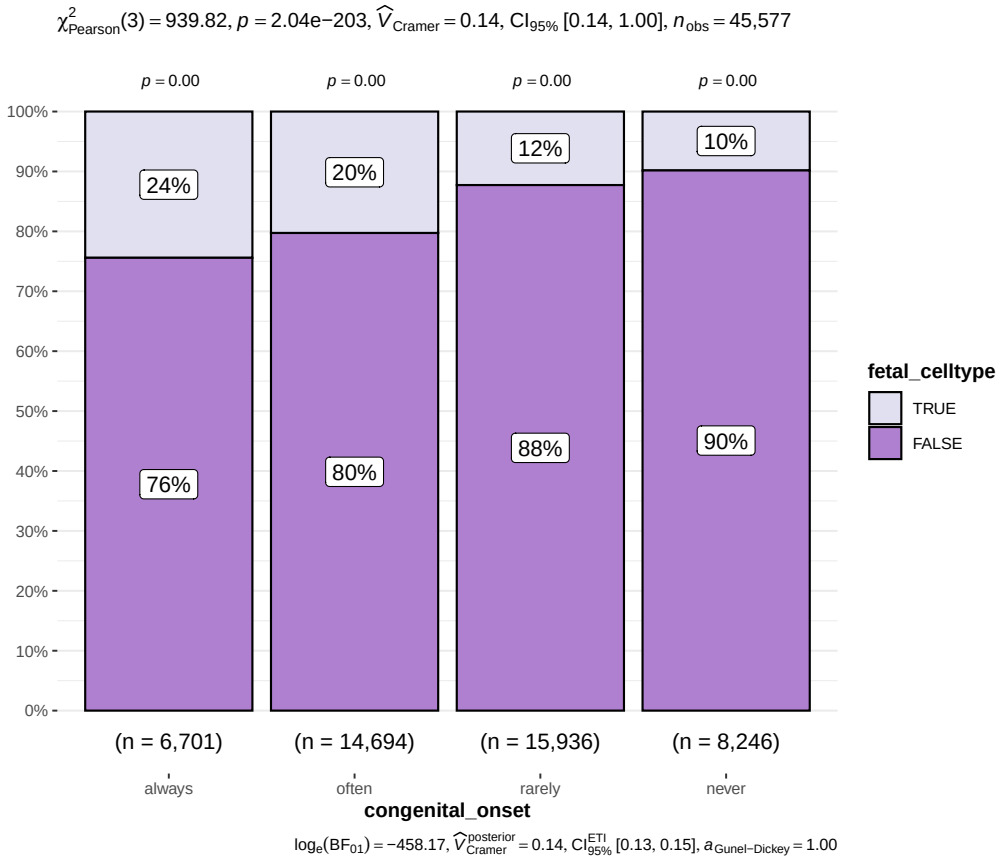


Figure 5: Congenital phenotypes are more often associated with foetal cell types. As a phenotype is more often congenital in nature, the greater proportion of foetal cell types are significantly asscoaited with it.

Diagnosis via cell type-specific disease prediction

Using the function `MSTExplorer::predict_celltypes` we input 3 inclusion phenotypes ('Generalized neonatal hypotonia' (*HP:0008935*), 'Scrotal hypospadias' (*HP:0012853*), 'Increased circulating progesterone' (*HP:0031216*)), 2 genes in which the patient is known to have deleterious mutations (*HSD3B2*, *HERC2*) and 1 gene in which the patient is known not to have any deleterious mutations (*SNORD115-1*). This predicted that cortical cell of adrenal gland (score sum=1.38, score mean=0.0256, score standard deviation=0.137) were the most probable cell types underlying this combination of phenotypes and genotypes (Fig. 6), which is highly consistent with existing evidence that adrenal insufficiency can cause both phenotypes via mutations in these genes[50,51]. This was the only cell type to receive a score two standard deviations from the mean score of all cell types (mean score: 0.000668).

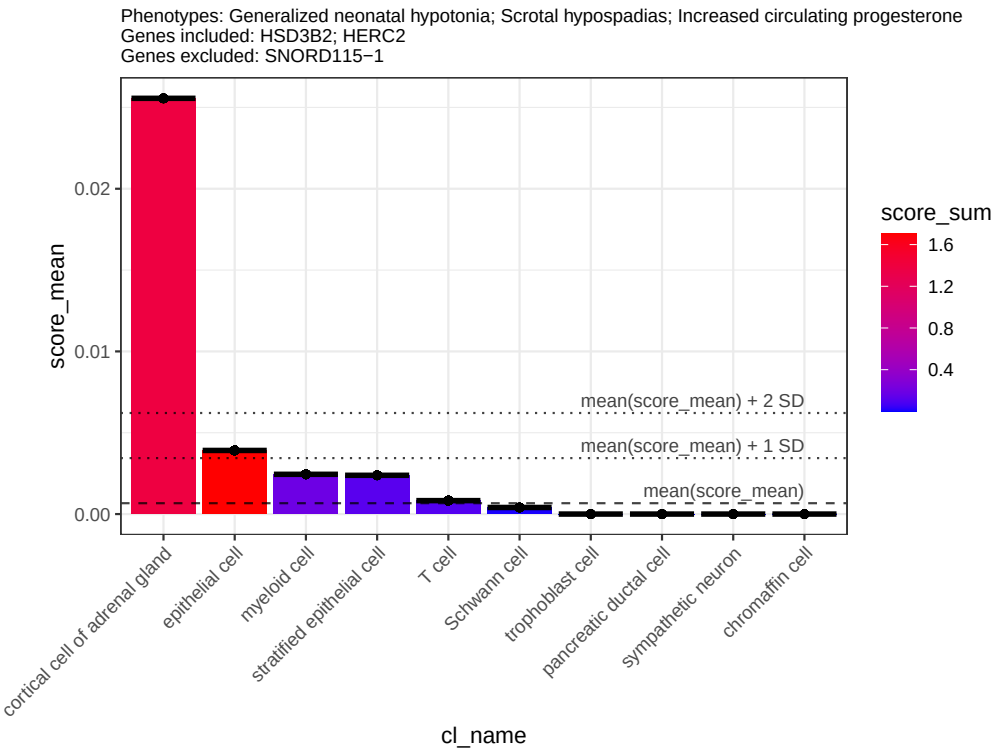


Figure 6: Diagnosis - Observed phenotypes/genotypes can be used to identify causal cell types in individuals. Our phenotype-cell type association results can be used to make predictions about which cell types are underlying a set of phenotypes observed in a given patient. Here we input three inclusion phenotypes, two inclusion genes, and one exclusion gene into the function `MSTExplorer::predict_celltypes`. The output is a ranked list of the top 10 most probable cell types (*x-axis*) underlying this combination of phenotypes/genotypes (highest to lowest rank from left to right). The score on the *y-axis* is computed by aggregating phenotype-celltype association summary statistics and evidence-weighted phenotype-gene associations. In this simple example, cortical cells of the adrenal gland were predicted as the most probable cell type. The mean of the score sum is shown as a dashed line, while one standard deviation (SD) above this is shown as a dotted line. Each bar is coloured by its mean.

Prognosis via cell type-mediated differential outcomes

Hypotonia (*HP:0001252*) is a very broad phenotype containing 13 subterms (e.g. "Generalised neonatal hypotonia") and is associated with 2569 unique diseases in the HPO gene annotations. Together, these hypotonia phenotypes were significantly associated with 29/99 (29.29%) unique CL-aligned cell types. This reflects the highly variable set of disease etiologies that can cause this broad-level phenotype. Across all diseases, hypotonia phenotypes tended to be most consistently severe (lower mean age of death

score) when associated with the cell type inhibitory interneuron. While other cell types were associated with lower mean age of death scores (e.g. stromal cell, astrocyte), the severity of the outcomes were more variable.

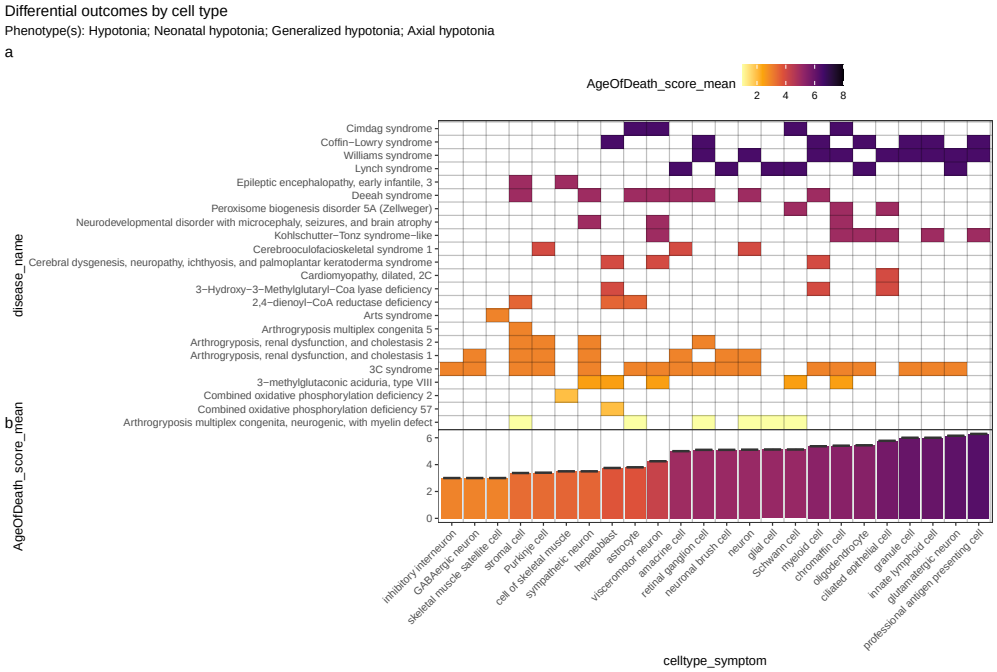


Figure 7: Prognosis - Cell types predict the probability of deadly diseases. The broad phenotype 'Hypotonia' and its descendants occur in many different diseases (1,832 diseases in the HPO annotations).Therefore, it can be difficult to prognose clinical outcomes of a newborn individual with hypotonia. With additional knowledge of the particular cell types underlying a patient's hypotonia phenotype, one can greatly narrow down the range of potential outcomes (e.g. age of death). **a**, Here, we show the various cell types by which hypotonia phenotypes confer disease risk. **b**, We also computed the mean age of death score for each cell type across hypotonia-associated diseases, revealing that disrupted inhibitory neurons confer the greatest risk of early death. Ordinal age of death categories from the HPO disease annotations were encoded numerically and averaged (**?@tbl-death**) to produce mean Age of Death scores for each disease (on a scale from 1-8). For example, a score of 1 corresponds to prenatal death, while a score of 8 corresponds to death in late adulthood.

Therapeutic target identification

Next, we identified putative cell type-specific gene targets for several severe disease phenotypes. This yielded putative therapeutic targets for 5287 phenotypes across 4850 diseases in 201 cell types and 3180 genes (Fig. 13). While this constitutes a

large number of genes in total, each phenotype was assigned a median of 2 gene targets (mean=3.29, min=1, max=10). Relative to the number of genes annotations per phenotype in the HPO overall (median=7, mean=61.95, min=1, max=5003) this represents a substantial decrease in the number of candidate target genes, even when excluding high-level phenotypes (HPO level>3). It is also important to note that the phenotypes in the prioritised targets list are ranked by their severity, allowing us to distinguish between phenotypes with a high medical urgency (e.g. 'Hydranencephaly') from those with lower medical urgency (e.g. 'Hyperplastic labia majora'). This can be useful for both clinicians, biomedical scientists, and pharmaceutical manufacturers who wish to focus their research efforts on phenotypes with the greatest need for intervention.

Across all phenotypes, epithelial cell were most commonly implicated (834 phenotypes), followed by stromal cell (627 phenotypes), stromal cell (627 phenotypes), neuron (478 phenotypes), chondrocyte (385 phenotypes), and endothelial cell (363 phenotypes). Grouped by higher-order ontology category, 'Abnormality of the musculoskeletal system' had the greatest number of enriched phenotypes (961 phenotypes, 863 genes), followed by 'Abnormality of the nervous system' (745 phenotypes, 1163 genes), 'Abnormality of head or neck' (545 phenotypes, 997 genes), 'Abnormality of the genitourinary system' (446 phenotypes, 710 genes), and 'Abnormality of the eye' (379 phenotypes, 572 genes).

Therapeutic target validation

To determine whether the genes prioritised by our therapeutic targets pipeline were plausible, we checked what percentage of gene therapy targets we recapitulated. Data on therapeutic approval status was gathered from the Therapeutic Target Database (TTD; release 2024-03-22)[52]. Overall, we prioritised 79% of all non-failed existing gene therapy targets. A hypergeometric test confirmed that our prioritised targets were significantly enriched for non-failed gene therapy targets ($p = 0.0104$). Importantly, we did not prioritise any of the failed therapeutics (0%), defined as having been terminated or withdrawn from the market. The hypergeometric test for depletion of failed targets did not reach significance ($p = 0.365$), but this is to be expected as there was only one failed gene therapy target in the TTD database.

Even when considering therapeutics of any kind (Fig. 14), not just gene therapies, we recapitulated 44% of the non-failed therapeutic targets and 0% of the terminated/withdrawn therapeutic targets (n=1255). Here we found that our prioritised targets were significantly enriched for non-failed therapeutics ($p = 3e{-}19$), and highly significantly depleted for failed therapeutics ($p = 3e - 199$). This suggests that our multi-scale evidence-based prioritisation pipeline is capable of selectively identifying genes that are likely to be effective therapeutic targets.

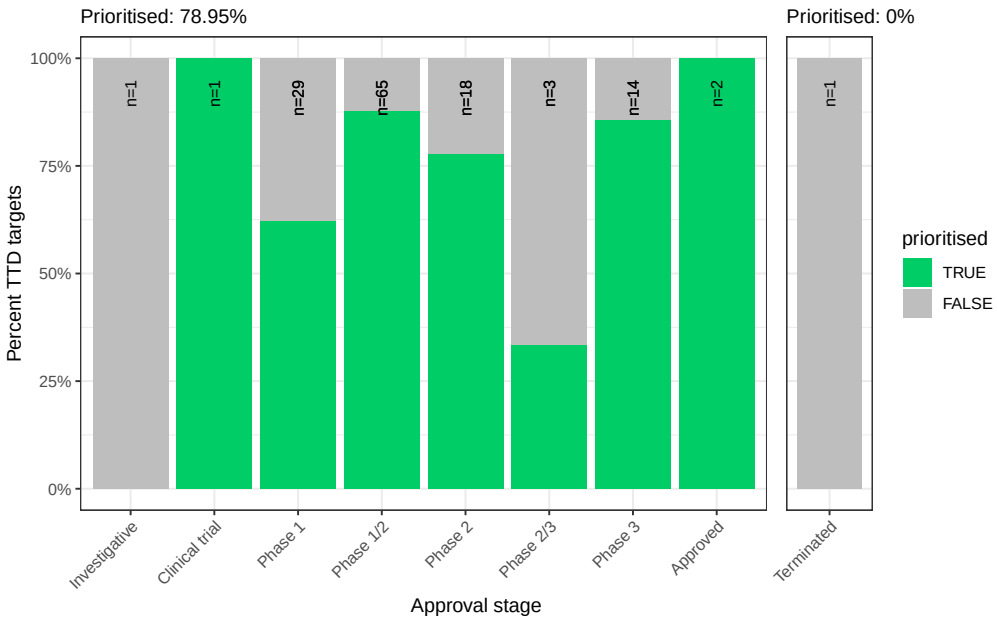


Figure 8: Therapeutics - Validation of prioritised therapeutic targets. The proportion of existing gene therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline. Therapetics are stratified by the stage of clinical development they were at during the time of writing.

Selected example targets

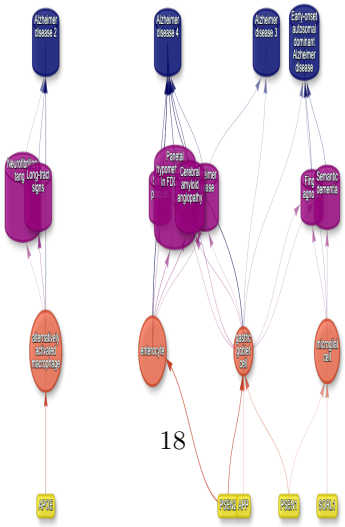
From our prioritised targets, we selected the following four sets of phenotypes or diseases as examples: 'Lethal skeletal dysplasia', 'GM2-ganglioside accumulation', 'Alzheimer disease', 'Parkinson disease'.

Skeletal dysplasia is a heterogeneous group of over 450 disorders that affect the growth and development of bone and cartilage. This phenotype can be lethal when deficient bone growth leads to the constriction of vital organs such as the lungs. Even after surgical interventions, these complications continue to arise as the child develops. Pharmacological interventions to treat this condition have largely been ineffective. While there are various cell types involved in skeletal system development, our pipeline nominated chondrocytes as the causal cell type underlying the lethal form of this condition. Assuringly, we found that the disease 'Achondrogenesis Type 1B' is caused by the genes *SLC26A2* and *COL2A1* via chondrocytes. We also found that 'Platyspondylic lethal skeletal dysplasia, Torrance type'. Thus, in cases where surgical intervention is insufficient, targeting these genes within chondrocytes may prove a viable long-term solution for children suffering from lethal skeletal dysplasia.

Tay-Sachs disease is a devastating disease in which children are born appearing healthy, which gradually degrades leading to death after 3-5 years. The underlying

(a) Lethal skeletal dysplasia

(b) GM2-ganglioside accumulation



cause is the toxic accumulation of gangliosides in the nervous system due to a loss of the enzyme produced by *HEXA*. While this could in theory be corrected with gene editing technologies, there remain some outstanding challenges. One of which is early detection and diagnosis, before irreversible damage has occurred. Our pipeline implicated extravillous trophoblasts of the placenta in 'GM2-ganglioside accumulation'. While not necessarily a target for gene therapy, checking these cells *in utero* for an absence of *HEXA* may serve as a viable biomarker as these cells normally express the gene at high levels. Early detection of Tay-Sachs disease may lengthen the window of opportunity for therapeutic intervention, especially when genetic sequencing is not available or variants of unknown significance are found within *HEXA*.

Alzheimer disease (AD) is the most common neurodegenerative condition. It is characterised by a set of variably penetrant phenotypes including memory loss, cognitive decline, cerebral proteinopathy. Interestingly, we found that different forms of early onset AD (which are defined by the presence of a specific disease gene) are each associated with different cell types via different phenotypes. For example, AD 3 and AD 4 are primarily associated with cells of the digestive system ('enterocyte', 'gastric goblet cell') and are implied to be responsible for the phenotypes 'Senile plaques', 'Alzheimer disease', 'Parietal hypometabolism in FDG PET', 'Cerebral amyloid angiopathy'. Meanwhile, early-onset autosomal dominant AD and AD 2 are primarily associated with immune cells ('alternatively activated macrophage', 'microglial cell') and are implied to be responsible for the phenotypes 'Neurofibrillary tangles', 'Long-tract signs', 'Finger agnosia', 'Semantic dementia'. This suggests that different forms of AD may be driven by different cell types and phenotypes, which may help explain its variability in onset and clinical presentation.

Finally, Parkinson disease (PD) is characterised by motor symptoms such as tremor, rigidity, and bradykinesia. However there are a number of additional phenotypes associated with the disease that span multiple physiological systems. PD 19a and PD 8 seemed to align most closely with the canonical understanding of PD as a disease of the central nervous system in that they implicated oligodendrocytes and neurons. Though the reference datasets being used in this study were not annotated at sufficient resolution to distinguish between different subtypes of neurons, in particular dopaminergic neurons. PD 19a/8 also suggested that risk variants in *LRRK2* mediate their effects on PD through both myeloid cells and oligodendrocytes by causing gliosis of the substantia nigra. The remaining clusters of PD mechanisms revolved around chondrocytes (PD 20), amacrine cells of the eye (hereditary late-onset PD), and the respiratory/immune system (PD 14). While the diversity in cell type-specific mechanisms is somewhat surprising, it may help to explain the wide variety of cross-system phenotypes frequently observed in PD.

It should be noted that the HPO only includes gene annotations for the monogenic forms of AD and PD. However it has previously been shown that there is at least partial overlap in their phenotypic and genetic aetiology with respect to their common forms. Thus understanding the monogenic forms of these diseases may shed light onto their more common counterparts.

Experimental model translatability

We computed interspecies translatability scores using a combination of both ontological (SIM_o) and genotypic (SIM_g) similarity relative to each homologous human phenotype and its associated genes Fig. 15. In total, we mapped 278 non-human phenotypes (in *Caenorhabditis elegans*, *Danio rerio*, *Mus musculus*, *Rattus norvegicus*) to 849 homologous human phenotypes. Amongst the 5287 phenotype within our prioritised therapy targets, 356 had viable animal models in at least on non-human species. Per species, the number of homologous phenotypes was: *Danio rerio* (n=214), *Mus musculus* (n=152), *Caenorhabditis elegans* (n=35), *Rattus norvegicus* (n=3). Amongst our prioritised targets with a GPT-4 severity score of >10, the phenotypes with the greatest animal model similarity were 'Anterior vertebral fusion' ($SIM_{o,g} = 0.967$), 'Disc-like vertebral bodies' ($SIM_{o,g} = 0.964$), 'Metaphyseal enchondromatosis' ($SIM_{o,g} = 0.946$), 'Peripheral retinal avascularization' ($SIM_{o,g} = 0.943$), 'Retinal vascular malformation' ($SIM_{o,g} = 0.943$).

Discussion

Across the 201 cell types and 11,047 RD-associated phenotypes investigated, more than 46,514 significant phenotype-cell type relationships were discovered. The examples we have highlighted above recapitulate well-known relationships, provide additional cellular context to many of these known relationships, and discover novel relationships at multiple biological scales.

Investigating RDs at the level of phenotypes offers several key advantages. First, the vast majority of RDs only have one associated gene (7671/8631 diseases = 89%). Aggregating gene sets across diseases into phenotype-centric "buckets" permits sufficiently well-powered analyses, with an average of ~76 genes per phenotype (median=7) see Fig. 11. Second, we hypothesise that these phenotype-level gene sets converge on a limited number of molecular and cellular pathways. Perturbations to these pathways manifest as one or more phenotypes which, when considered together, tend to be clinically diagnosed as a certain disease. Third, RDs are often highly heterogeneous in their clinical presentation across individuals, leading to the creation of an ever increasing number of disease subtypes (some of which only have a single documented case). In contrast, a phenotype-centric approach enables us to more accurately describe a particular individual's version of a disease without relying on the generation of additional disease subcategories. By characterising an individual's precise phenotypes over time, we may better understand the underlying biological mechanisms that have caused their condition. However, in order to achieve a truly precision-based approach to clinical care, we must first characterise the molecular and cellular mechanisms that cause the emergence of each phenotype. Here, we provide a highly reproducible framework that enables this at the scale of the entire phenome. This presents an opportunity to design basket trials of patients with different diseases but overlapping phenotypes and cellular mechanisms[17]. It may be especially helpful for complex patients with diagnostically ambiguous sets of phenotypes who would otherwise be excluded from traditional clinical trials[53].

It was paramount to the success of this study to ensure our results were anchored in ground-truth benchmarks, generated falsifiable hypotheses, and rigorously guarded against false-positive associations. Extensive validation using multiple approaches demonstrated that our methodology consistently recapitulates expected phenotype-cell type associations (Fig. 1-Fig. 5). This was made possible by the existence of comprehensive, structured ontologies for all phenotypes (HPO) and cell types (CL), which provide an abundance of clear and falsifiable hypotheses for which to test our predictions against. Several key examples include 1) strong enrichment of associations between cell types and phenotypes within the same anatomical systems (Fig. 1b-d), 2) a strong relationship between phenotype-specificity and the strength and number of cell type associations (Fig. 2), 3) identification of the precise cell subtypes involved in susceptibility to various subtypes of recurrent bacterial infections (Fig. 3), 4) a strong positive correlation between the frequency of congenital onset of a phenotype and the proportion of developmental cell types associated with it (Fig. 5)), and 5) consistent phenotype-cell type associations across multiple independent single-cell datasets (Fig. 12). Having validated our phenotype-cell type associations, we then went on to demonstrate how these results may be used in each stage of clinical care: diagnosis (Fig. 6), prognosis (Fig. 7), treatment, and therapeutics development (Fig. 9).

Diagnosis is an essential but challenging step in RD patient care. Additional phenotypes that emerge over time may assist a clinician to reach a more confident disease diagnosis. However many of these phenotypes can have a serious impact on patient quality of life or survival and avoiding them would be far better for patient outcomes. Often times phenotypes alone cannot clearly pinpoint the disease and thus a diagnosis is never reached. Having a more complete understanding of the mechanisms underlying observed phenotypes allows clinicians to far more effectively make predictions about what additional, less obvious phenotypes they should search for to confirm or reject their hypothesis of disease diagnosis (e.g. with imaging or biomarker tests).

Consider the following hypothetical scenario. A clinician observes that a newborn patient has several phenotypes ('Generalized neonatal hypotonia', 'Scrotal hypospadias', 'Increased circulating progesterone'), none of which conclusively point to a single disease diagnosis. Under the strong suspicion that the phenotypes are genetic in origin, the clinician orders whole-genome sequencing (WGS) on the patient as well as the patient's family. The clinician finds that the patient has a number of putative causal mutations, narrowing down the number of potential diseases from hundreds to just 10. Further narrowing down the possibilities at this stage can be extremely challenging even for expert clinical geneticists. However, additional knowledge of which tissues and cell types are primarily affected allow the clinician to make a series of testable hypotheses that they may begin to investigate. For example, two of the putative diseases are known to cause aberrant splicing events in a gene that is only expressed in adrenocortical cells (Fig. 6), providing justification to order a needle biopsy of the adrenal gland. RNA sequencing is performed on the tissue biopsy and it is discovered that the patient does indeed have high expression of the dysfunctional transcript, confirming the disease diagnosis[54]. This opens new avenues for the patient to receive timely and effective treatments for their specific condition, which is important as their version of

the disease tends to lead to death in early childhood if left untreated (Fig. 7). Fortunately, their diagnosis now qualifies them to participate in a clinical trial of a novel gene therapy with promising preliminary results. Furthermore, it is predicted that this patient would respond especially well to this treatment given that the mechanisms of action of the gene therapy primarily acts on adrenocortical cells (Fig. 9).

Unfortunately, there are currently only treatments available for less than 5% of RDs[6]. Novel technologies including CRISPR, prime editing, antisense oligonucleotides, viral vectors, and/or lipid nanoparticles, have been undergone significant advances in the last several years[55–59] and proven remarkable clinical success in an increasing number of clinical applications[60–63]. The U.S. Food and Drug Administration (FDA) recently announced an landmark program aimed towards improving the international regulatory framework to take advantage of the evolving gene/cell therapy technologies[64] with the aim of bringing dozens more therapies to patients in a substantially shorter timeframe than traditional pharmaceutical product development (typically 5-20 years with a median of 8.3 years)[65]. While these technologies have the potential to revolutionise RD medicine, their successful application is dependent on first understanding the mechanisms causing each disease.

To address this critical gap in knowledge, we used our results to create a reproducible and customisable pipeline to nominate cell type-resolved therapeutic targets (Fig. 13-Fig. 9). Targeting cell type-specific mechanisms underlying granular RD phenotypes can improve therapeutic effectiveness by treating the causal root of an individual's conditions[56,66]. A cell type-specific approach also helps to reduce the number of harmful side effects caused by unintentionally delivering the therapeutic to off-target tissues/cell types (which may induce aberrant gene activity), especially when combined with technologies that can target cell surface antigens (e.g viral vectors)[67]. This has the additional benefit of reducing the minimal effective dose of a therapeutic, which can be both immunogenic and extremely financially costly[9,10,55,58]. Here, we demonstrate the utility of a high-throughput evidence-based approach to RD therapeutics discovery by highlighting several of the most promising therapeutic candidates. Our pipeline takes into account a myriad of factors, including the strength of the phenotype-cell type associations, symptom-cell type associations, cell type-specificity of causal genes, the severity and frequency of the phenotypes, suitability for gene therapy delivery systems (e.g. recombinant adeno-associated viral vectors (rAAV)), as well as a quantitative analysis of phenotypic and genetic animal model translatability (Fig. 15). We validated these candidates by comparing the proportional overlap with gene therapies that are presently in the market or undergoing clinical trials, in which we recovered 79% of all active gene therapies and 0% of failed gene therapies (Fig. 8, Fig. 14). Despite nominating a large number of putative targets, hypergeometric tests confirmed that our targets were strongly enriched for targets of existing therapies that are either approved or currently undergoing clinical trials.

It should be noted that our study has several key limitations. First, while our cell type datasets are amongst the most comprehensive human scRNA-seq references currently available, they are nevertheless missing certain tissues, cell types (e.g. spermatocytes, oocytes), and life stages (post-natal childhood, senility). It is also possible that we

have not captured certain cell state signatures that only occur in disease (e.g. disease-associated microglia [**CITATION**]). Though we reasoned that using only control cell type signatures would mitigate bias towards any particular disease, and avoid degradation of gene signatures due to loss of function mutations. Second, the collective knowledge of gene-phenotype and gene-disease associations is far from complete and we fully anticipate that these annotations will continue to expand and change well into the future. It is for this reason we designed this study to be easily reproduced within a single containerised script so that we (or others) may rerun it with updated datasets at any point. Finally, causality is notoriously difficult to prove definitively from associative testing alone, and our study is not exempt from this rule. Despite this, there are several reasons to believe that our approach is able to better approximate causal relationships than traditional approaches. First, we did not intentionally preselect any subset of phenotypes or cell types to investigate here. Along with a scaling prestep during linear modelling, this means that all the results are internally consistent and can be directly compared to one another (in stark contrast to literature meta-analyses). Furthermore, for the phenotype gene signatures we used expert-curated GenCC annotations[68,69] to weight the current strength of evidence supporting a causal relationship between each gene and phenotype. This is especially important for phenotypes with large genes lists (thousands of annotations) for which some of the relationships may be tenuous. Within the cell type references, we deliberately chose to use specificity scores (rather than raw gene expression) as this normalisation procedure has previously been demonstrated to better distinguish between signatures of highly similar cell types/subtypes[70].

Moving forward, we are now actively seeking industry and academic partnerships to begin experimentally validating our multi-scale target predictions and exploring their potential for therapeutic translation. Nevertheless, there are more promising therapeutic targets here than our research group could ever hope to pursue by ourselves. In the interest of accelerating research and ensuring RD patients are able to benefit from this work as quickly as possible, we have decided to publicly release all of the results described in this study. These can be accessed in multiple ways, including through a suite of R packages as well as a web app, the Rare Disease Celltyping Portal. The latter allows our results to be easily queried, filtered, visualised, and downloaded without any knowledge of programming. Through these resources we aim to make our findings useful to a wide variety of RD stakeholders including subdomain experts, clinicians, advocacy groups, and patients.

Conclusions

Ultimately, our primary objective was to develop a methodology capable of generating high-throughput phenome-wide predictions while preserving the accuracy and clinical utility typically associated with more narrowly focused studies. With the rapid advancement of gene therapy technologies, and a regulatory landscape that is evolving to better meet the needs of a large and diverse patient population, there is finally momentum to begin to realise the promise of personalised medicine. This has especially important implications for the global RD community which has remained relatively

neglected. Here, we lay out the groundwork necessary for this watershed moment by providing a scalable, cost-effective, and fully reproducible means of resolving the multi-scale, cell-type specific mechanisms of virtually all rare diseases.

Methods

Human Phenotype Ontology

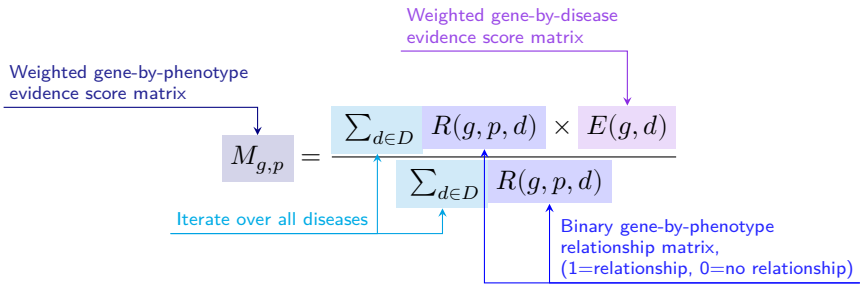
The latest version of the HPO (release 2024-02-08) was downloaded from the EMBL-EBI Ontology Lookup Service[71] and imported into R using the `HPOExplorer` package. This R object was used to extract ontological relationships between phenotypes as well as to assign absolute and relative ontological levels to each phenotype. The latest version of the HPO phenotype-to-gene mappings and phenotype annotations were downloaded from the official HPO GitHub repository and imported into R using `HPOExplorer`. This contains lists of genes associated with phenotypes via particular diseases, formatted as three columns in a table (gene, phenotype, disease).

However, not all genes have equally strong evidence of causality with a disease or phenotype, especially when considering that the variety of resources used to generate these annotations (OMIM, Orphanet, DECIPHER) use variable methodologies (e.g. expert-curated review of the medical literature vs. automated text mining of the literature). Therefore we imported data from the Gene Curation Coalition (GenCC)[68,69], which (as of 2024-03-01) 21798 evidence scores across 7229 diseases and 5142 genes. Evidence scores are defined by GenCC using a standardised ordinal rubric which we then encoded as a semi-quantitative score ranging from 0 (no evidence of disease-gene relationship) to 6 (strongest evidence of disease-gene relationship) (see **?@tbl-gencc**). We then summed evidence scores per disease, merged this table with the HPO disease-phenotype-gene annotation table, and then cast the data into a gene-by-phenotype matrix filled with the aggregated mean evidence score. This can be expressed as the following equations.

Let us denote:

- D as the set of d diseases.
- p as a phenotype.
- g as a gene.

The final evidence-weighted gene-by-phenotype matrix ($M_{g,p}$) can be expressed as:



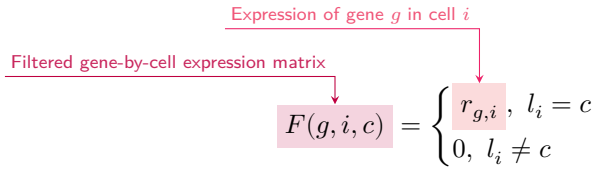
Histograms of evidence score distributions at each step in processing can be found in Fig. 10.

Single-cell transcriptomic atlases

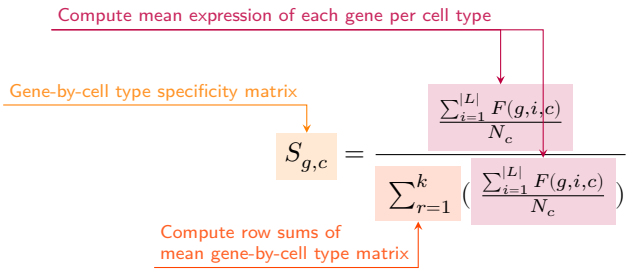
In this study, the gene by cell type specificity matrix was constructed using the Descartes Human transcriptome atlas of foetal gene expression, which contains a mixture of single-nucleus and single-cell RNA-seq data (collected with sci-RNA-seq3)[31]. This dataset contains 377,456 cells representing 77 distinct cell types across 15 tissues. All 121 human foetal samples ranged from 72 to 129 days in estimated postconceptual age. To independently replicate our findings, we also used the Human Cell Landscape which contains single-cell transcriptomic data (collected with microwell-seq) from embryonic, foetal, and adult human samples across 49 tissues[32].

Specificity matrices were generated separately for each transcriptomic atlas using the R package `EWCE` (v1.11.3)[70]. Within each atlas, cell types were defined using the authors' original freeform annotations in order to preserve the granularity of cell subtypes as well as incorporate expert-identified rare cell types. Cell types were only aligned and aggregated to the level of corresponding Cell Ontology (CL)[33] annotations afterwards when generating summary figures and performing cross-atlas analyses. Using the original gene-by-cell count matrices from each single-cell atlas, we computed gene-by-cell type expression specificity matrices as follows.

Let us denote: g as a gene, c as a cell type, and i as a single cell. Genes with very no expression across any cell types were considered to be uninformative and were therefore removed from the input gene-by-cell matrix $F(g, i, c)$.



Next, we calculated the mean expression per cell type and normalised the resulting matrix to transform it into a gene-by-cell type expression specificity matrix ($S_{g,c}$). In other words, each gene in each cell type had a 0-1 score where 1 indicated the gene was mostly specifically expressed in that particular cell type relative to all other cell types. This procedure was repeated separately for each of the single-cell atlases and can be summarised as:



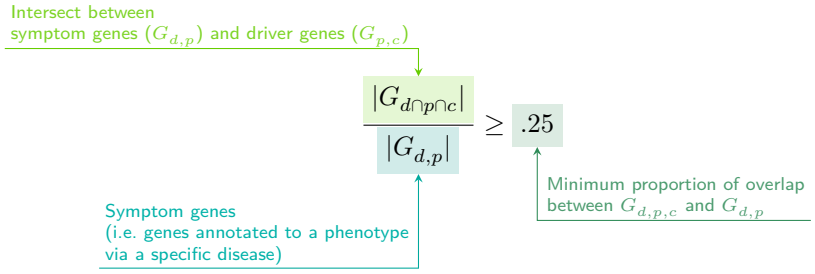
Phenotype-cell type associations

To test for relationships between each pairwise combination of phenotype (n=11,047) and cell type (n=201) we ran a series of univariate generalised linear models implemented via the `stats::glm` function in R. First, we filtered the gene-by-phenotype evidence score matrix ($M_{g,p}$) and the gene-by-cell type expression specificity matrix ($S_{g,c}$) to only include genes present in both matrices (n=4,949 genes in the Descartes Human analyses; n=4,653 genes in the Human Cell Landscape analyses). Then, within each matrix any rows or columns with a sum of 0 were removed as these were uninformative data points that did not vary. To improve interpretability of the results β coefficient estimates across models (i.e. effect size), we performed a scaling prestep on all dependent and independent variables. Initial tests showed that this had virtually no impact on the total number of significant results or any of the benchmarking metrics based on p-value thresholds Fig. 1. This scaling prestep improved our ability to rank cell types by the strength of their association with a given phenotype as determined by separate linear models.

We repeated the aforementioned procedure separately for each of the single-cell references. Once all results were generated using both cell type references (2,206,994 association tests total), we applied Benjamini-Hochberg false discovery rate[72] (denoted as $FDR_{p,c}$) to account for multiple testing. Of note, we applied this correction across all results at once (as opposed to each single-cell reference separately) to ensure the $FDR_{p,c}$ was stringently controlled for across all tests performed in this study.

Symptom-cell type associations

Here we define a symptom as a phenotype as it presents within the context of the specific disease. The features of a given symptom can be described as the subset of genes annotated to phenotype p via a particular disease d, denoted as $G_{d,p}$ (see Fig. 11). To attribute our phenotype-level cell type enrichment signatures to specific diseases, we first identified the gene subset that was most strongly driving the phenotype-cell type association by computing the intersect of genes that were both in the phenotype annotation and within the top 25% specificity percentile for the associated cell type. We then computed the intersect between symptom genes ($G_{d,p}$) and driver genes ($G_{p,c}$), resulting in the gene subset $G_{d \cap p \cap c}$. Only $G_{d \cap p \cap c}$ gene sets with 25% or greater overlap with the symptom gene subset ($G_{d,p}$) were kept. This procedure was repeated for all phenotype-cell type-disease triads, which can be summarised as follows:



Validation of expected phenotype-cell type relationships

We first sought to confirm that our tests (across both single-cell references) were able to recover expected phenotype-cell type relationships across seven high-level branches within the HPO (Fig. 1), including abnormalities of the cardiovascular system, endocrine system, eye, immune system, musculoskeletal system, nervous system, and respiratory system. Within each branch the number of significant tests in a given cell type were plotted (Fig. 1b). Mappings between freeform annotations (the level at which we performed our phenotype- cell type association tests) provided by the original atlas authors and their closest CL term equivalents were provided by CellxGene[29]. CL terms along the *x-axis* of Fig. 1b were assigned colours corresponding to which HPO branch showed the greatest number of enrichments (after normalising within each branch to account for differences in scale). The normalised colouring allows readers to quickly assess which HPO branch was most often associated with each cell type, while accounting for differences in the number of phenotypes across branches. We then ran a series of Analysis of Variance (ANOVA) tests to determine whether (within a given branch) a given cell type was more often enriched ($FDR_{p,c} < 0.05$) within that branch relative to all of the other HPO branches of an equivalent level in the ontology (including all branches not shown in Fig. 1b). After applying Benjamini-Hochberg multiple testing correction[72] (denoted as $FDR_{b,c}$), we annotated each respective branch-by-cell type bar according to the significance (**** : $FDR_{b,c} < 1e-04$, *** : $FDR_{b,c} < 0.001$, ** : $FDR_{b,c} < 0.01$, * : $FDR_{b,c} < 0.05$). Cell types in Fig. 1a-b were ordered along the *x-axis* according to a dendrogram derived from the CL ontology (Fig. 1c), which provides ground-truth semantic relationships between all cell types (e.g. different neuronal subtypes are grouped together).

As an additional measure of the accuracy of our phenotype-cell types test results we identified conceptually matched branches across the HPO and the CL (Fig. 1d and **?@tbl-celltypes**). For example, 'Abnormality of the cardiovascular system' in the HPO was matched with 'cardiocytes' in the CL which includes all cell types specific to the heart. Analogously, 'Abnormality of the nervous system' in the HPO was matched with 'neural cell' in the CL which includes all descendant subtypes of neurons and glia. This cross-ontology matching was repeated for each HPO branch and can be referred to as on-target cell types. Within each branch, the $-log_{10}(FDR_{p,c})$ values of on-target cell types were binned by rounding to the nearest integer (*x-axis*) and the percentage of tests for on-target cell types relative to all cell types were computed at

each bin (*y-axis*) (Fig. 1d). The baseline level (dotted horizontal line) illustrates the percentage of on-target cell types relative to the total number of observed cell types. Any percentages above this baseline level represent greater than chance representation of the on-target cell types in the significant tests.

Monarch Knowledge Graph recall

Finally, we gathered known phenotype-cell type relationships from the Monarch Knowledge Graph (MKG), a comprehensive database of links between many aspects of disease biology[73]. This currently includes 103 links between HPO phenotypes (n=103) and CL cell types (n=79). Of these, we only considered the 82 phenotypes that we were able to test given that our approach was reliant on gene annotations. We considered instances where we found a significant relationship between exactly matching pairs of HPO-CL terms as a hit.

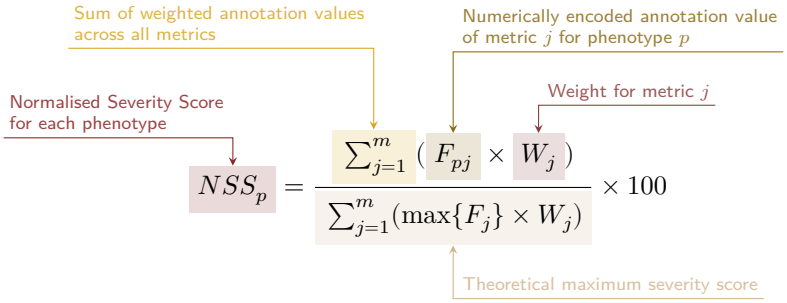
However, as the cell types in MKG were not necessarily annotated at the same level as our single-cell references, we also considered instances where the MKG cell type was an ancestor term of our cell type (e.g. 'myeloid cell' vs. 'monocyte'), or *vice versa*, as hits. Using these criteria, we determined our results recapitulated `**!!RECOMPUTE!!**`% of known phenotype-cell type relationships in the MKG. We next computed how far along the CL ontological tree we would need to travel in order to reach a common ancestor between the MKG cell type and our cell type, for each phenotype-cell type link in the MKG. This provides a metric of not just whether we recapitulated the exact cell types, but how dissimilar our identified cell types were for a given phenotype-cell type association (**?@fig-monarch-recall**).

Annotation of phenotypes using generative large language models

Only a small fraction of the the phenotypes in HPO (<1%) have metadata annotations containing information on their time course, consequences, and severity. This is due to the time-consuming nature of manually annotating thousands of phenotypes. To generate such annotations at scale, we used Generative Pre-trained Transformer 4 (GPT-4), a large language model (LLM) as implemented within OpenAI's chatGPT Application Programming Interface (API). After extensive prompt engineering and ground-truth benchmarking, we were able to acquire annotations on how often each phenotype directly causes intellectual disability, death, impaired mobility, physical malformations, blindness, sensory impairments, immunodeficiency, cancer, reduced fertility, or is associated with a congenital onset. These criteria were previously defined in surveys of medical experts as a means of systematically assessing phenotype severity[74]. Responses for each metric were provided in a consistent one-word format which could be one of: 'never', 'rarely', 'often', 'always'. This procedure was repeated in batches (to avoid exceeding token limits) until annotations were gathered for 16982/18082 HPO phenotypes.

We then encoded these responses into a semi-quantitative scoring system ('never'=0, 'rarely'=1, 'often'=2, 'always'=3), which were then weighted by multiplying a semi-subjective scoring of the relevance of each metric to the concept of severity on a scale from 1-5, with 5 being the most severe ('intellectual_disability'=5,

'death'=5, 'impaired_mobility'=4, 'physical_malformations'=3, 'blindness'=4, 'sensory_impairments'=3, 'immunodeficiency'=3, 'cancer'=3, 'reduced_fertility'=1, 'congenital_onset'=4). Finally, the product of the score was normalised to a quantitative severity score ranging from 0-100, where 100 is the theoretical maximum severity score. This phenotype severity scoring procedure can be expressed as follows.



Enrichment of foetal cell types in congenital phenotypes

The GPT-4 annotations also enabled us to assess whether foetal cell types were more often significantly associated with congenital phenotypes in our Human Cell Landscape results as this single-cell reference contained both adult and foetal versions of cell types (Fig. 5). To do this, we performed a chi-squared (χ^2) test on the proportion of significantly associated cell types containing any of the substrings 'fetal', 'fetus', 'primordial', 'hESC' or 'embryonic' (within cell types annotations from the original Human Cell Landscape authors[32]) vs. those associated without, stratified by how often the corresponding phenotype had a congenital onset according to the GPT phenotype annotations (including 'never', 'rarely', 'often', 'always'). In addition, a series of χ^2 tests were performed within each congenital onset frequency strata, to determine whether the observed proportion of foetal cell types vs. non-foetal cell types significantly deviated from the proportions expected by chance.

Diagnosis via cell type-specific disease prediction

We designed an algorithm that uses our results to predict the most likely cell types underlying a set of phenotypic and genotypic traits observed in a patient (Fig. 6). This is implemented within `MSTExplorer::predict_celltypes` and takes HPO phenotypes as inputs. It can optionally take included risk genes, excluded risk genes, included diseases and/or excluded diseases as additional inputs. It then computes the It then outputs a weighted ranking of cell types, where higher ranking indicates a higher likelihood of being the underlying mechanism of the patient's particular form of disease(s).

Prognosis via cell type-mediated differential outcomes

The phenotype hypotonia is associated with diseases that range in severity from benign to debilitating to fatal[75]. In the absence of additional information, making an accurate

diagnosis is extremely challenging even for experienced physicians. The magnitude of this challenge is highlighted by the fact that each disease is associated with anywhere between 1-595 unique phenotypes (median=61, mean=77.74) within the HPO. Conversely, each phenotype is associated with 1-5404 diseases (median=6, mean=60.74). We addressed this challenge by applying our phenotype-cell type association results in combination with expert-curated HPO annotations of clinical outcomes associated with each phenotype-disease pairing (Fig. 7). We first extracted results for the phenotype 'Hypotonia' (*HP:0001252*) and its 13 descendant subterms from our phenotype-cell type association analyses. Next, we encoded the "Age of Death" categories associated with each disease in an ordinal scale ranging from 1, corresponding to prenatal death, to 8, corresponding to death in late adulthood (**?@tbl-death**). To determine whether cell type identity significantly predicted the age of death, we conducted an ANOVA where cell type was the predictor and "Age of Death score" was the outcome.

Therapeutic target identification

We developed a systematic and automated strategy for identifying putative cell type-specific gene targets for each phenotype based on a series of filters at phenotype, cell type, and gene levels. The entire target prioritisation procedure can be replicated with a single function: `MSTExplorer::prioritise_targets`. This function automates all of the reference data gathering (e.g. phenotype metadata, cell type metadata, cell type signature reference, gene lengths, severity tiers) and takes a variety of arguments at each step for greater customisability.

Therapeutic target validation

To assess whether our prioritised therapeutic targets were likely to be viable, we computed the overlap between our gene targets and those of existing gene therapies at various stages of clinical development (Fig. 8). Gene targets were obtained for each therapy from the Therapeutic Target Database (TTD; release 2024-03-22) and mapped onto standardised HUGO Gene Nomenclature Committee (HGNC) gene symbols using the `orthogene` R package. We stratified our overlap metrics according to whether the therapies had failed (unsuccessful clinical trials or withdrawn), or were non-failed (successful or ongoing clinical trials). We then conducted hypergeometric tests to determine whether the observed overlap between our prioritised targets and the non-failed therapy targets was significantly greater than expected by chance (i.e. enrichment). We also conducted a second hypergeometric test to determine whether the observed overlap between our prioritised targets and the failed therapy targets was significantly less than expected by chance (i.e. depletion). Finally, we repeated the analysis against all therapeutic targets, not just those of gene therapies, to determine whether our prioritised targets had relevance to other therapeutic modalities.

Experimental model translatability

To improve the likelihood of successful translation between preclinical animal models and human patients, we created an interspecies translatability prediction tool for each phenotype nominated by our gene therapy prioritised pipeline (Fig. 15). First, we extracted ontological similarity scores of homologous phenotypes across species from the MKG[73]. Briefly, the ontological similarity scores (SIM_o) are computed for each homologous pair of phenotypes across two ontologies by calculating the overlap in homologous phenotypes that are ancestors or descendants of the target phenotype. Next, we generated genotypic similarity scores (SIM_g) for each homologous phenotype pair by computing the proportion of 1:1 orthologous genes using gene annotation from their respective ontologies. Interspecies orthologs were also obtained from the MKG. Finally, both scores are multiplied together to yield a unified ontological-genotypic similarity score ($SIM_{o,g}$).

Novel R packages

To facilitate all analyses described in this study and to make them more easily reproducible by others, we created several open-source R packages. `KGExplorer` imports and analyses large-scale biomedical knowledge graphs and ontologies. `HPOExplorer` aids in managing and querying the directed acyclic ontology graph within the HPO. `MSTExplorer` facilitates the efficient analysis of many thousands of phenotype-cell type association tests, and provides a suite of multi-scale therapeutic target prioritisation and visualisation functions. These R packages also include various functions for distributing the post-processed results from this study in an organised, tabular format. Of note, `MSTExplorer::load_example_results` loads all summary statistics from our phenotype-cell type tests performed here.

Rare Disease Celltyping Portal

To further increase the ease of access for stakeholders in the RD community without the need for programmatic experience, we developed a series of web apps to interactively explore, visualise, and download the results from our study. Collectively, these web apps are called the Rare Disease Celltyping Portal. The landing page for the website was made using HTML, CSS, and javascript and the web apps were created using the Shiny Web application framework for R and deployed on the shinyapps.io server. The website can be accessed here. All code used to generate the website can be found here.

Data and Code Availability

All data and code is made freely available through preexisting databases and/or GitHub repositories / software associated with this publication.

- Human Phenotype Ontology
- GenCC
- Descartes Human scRNA-seq atlas

- Human Cell Landscape scRNA-seq atlas
- Rare Disease Celltyping Portal
- `KGExplorer`
- `HPOExplorer`
- `MSTExplorer`
- Code to replicate analyses
- Cell type-specific gene target prioritisation

- Complement system gene list

Acknowledgements

We would like to thank the following individuals for their insightful feedback and assistance with data resources: Sarah J. Marzi, Gerton Lunter, Peter Robinson, Melissa Haendel, Ben Coleman, Nico Matentzoglu, Shawn T. O'Neil, Alan E. Murphy, Sarada Gurung.

Funding

This work was supported by a UK Dementia Research Institute (UK DRI) Future Leaders Fellowship [MR/T04327X/1] and the UK DRI which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

References

1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A* **179**, 885–892 (2019).
2. Zhu, Q. *et al.* An integrative knowledge graph for rare diseases, derived from the genetic and rare diseases information center (GARD). *J. Biomed. Semantics* **11**, 13 (2020).
3. Rare diseases BioResource.
4. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: Beyond the exome. *Genome Med.* **14**, 23 (2022).
5. Molster, C. *et al.* Survey of healthcare experiences of australian adults living with rare diseases. *Orphanet J. Rare Dis.* **11**, 30 (2016).
6. Halley, M. C., Smith, H. S., Ashley, E. A., Goldenberg, A. J. & Tabor, H. K. A call for an integrated approach to improve efficiency, equity and sustainability in rare disease research in the united states. *Nat. Genet.* **54**, 219–222 (2022).
7. Institute of Medicine (US) Committee on Accelerating Rare Diseases Research and Orphan Product Development, Field, M. J. & Boat, T. F. *Coverage and Reimbursement: Incentives and Disincentives for Product Development.* (National Academies Press (US), 2010).
8. Yates, N. & Hinkel, J. The economics of moonshots: Value in rare disease drug development. *Clin. Transl. Sci.* **15**, 809–812 (2022).

9. Nuijten, M. Pricing zolgensma - the world's most expensive drug. *J Mark Access Health Policy* **10**, 2022353 (2022).
10. Thielen, F. W., Heine, R. J. S. D., Berg, S. van den, Ham, R. M. T. T. & Groot, C. A. U. Towards sustainability and affordability of expensive cell and gene therapies? Applying a cost-based pricing model to estimate prices for libmeldy and zolgensma. *Cytotherapy* **24**, 1245–1258 (2022).
11. Gargano, M. A. *et al.* The human phenotype ontology in 2024: Phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
12. Köhler, S. *et al.* Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
13. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
14. Robinson, P. N. *et al.* The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
15. Osmond, M. *et al.* Outcome of over 1500 matches through the matchmaker exchange for rare disease gene discovery: The 2-year experience of Care4Rare canada. *Genet. Med.* **24**, 100–108 (2022).
16. Philippakis, A. A. *et al.* The matchmaker exchange: A platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).
17. Zanello, G. *et al.* Targeting shared molecular etiologies to accelerate drug development for rare diseases. *EMBO Mol. Med.* **15**, e17159 (2023).
18. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: Analysis of the orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
19. Rare diseases, common challenges. *Nat. Genet.* **54**, 215 (2022).
20. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
21. Amberger, J. S. & Hamosh, A. Searching online mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic phenotypes. *Curr. Protoc. Bioinformatics* **58**, 1.2.1–1.2.12 (2017).
22. McKusick, V. A. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
23. Maiella, S., Rath, A., Angin, C., Mousson, F. & Kremp, O. [Orphanet and its consortium: Where to find expert-validated information on rare diseases]. *Rev. Neurol.* **169 Suppl 1**, S3–8 (2013).
24. Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E. en & Cornel, M. C. [Orphanet: A european database for rare diseases]. *Ned. Tijdschr. Geneeskd.* **152**, 518–519 (2008).
25. Firth, H. V. *et al.* DECIPHER: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).

26. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
27. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
28. Qi, R. & Zou, Q. Trends and potential of machine learning and deep learning in drug study at Single-Cell level. *Research* **6**, 0050 (2023).
29. CZI Single-Cell Biology Program *et al.* CZ CELL×GENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv* 2023.10.30.563174 (2023).
30. Svensson, V., Veiga Beltrame, E. da & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, (2020).
31. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
32. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
33. Diehl, A. D. *et al.* The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
34. Heim, C. E. *et al.* Myeloid-derived suppressor cells contribute to staphylococcus aureus orthopedic biofilm infection. *J. Immunol.* **192**, 3778–3792 (2014).
35. Pidwill, G. R., Gibson, J. F., Cole, J., Renshaw, S. A. & Foster, S. J. The role of macrophages in staphylococcus aureus infection. *Front. Immunol.* **11**, 620339 (2020).
36. Stoll, H. *et al.* Staphylococcal enterotoxins Dose-Dependently modulate the generation of Myeloid-Derived suppressor cells. *Front. Cell. Infect. Microbiol.* **8**, 321 (2018).
37. Tebartz, C. *et al.* A major role for myeloid-derived suppressor cells and a minor role for regulatory T cells in immunosuppression during staphylococcus aureus infection. *J. Immunol.* **194**, 1100–1111 (2015).
38. Zhou, Z., Xu, M.-J. & Gao, B. Hepatocytes: A key cell type for innate immunity. *Cell. Mol. Immunol.* **13**, 301–315 (2016).
39. Dixon, L. J., Barnes, M., Tang, H., Pritchard, M. T. & Nagy, L. E. Kupffer cells in the liver. *Compr. Physiol.* **3**, 785–797 (2013).
40. Ladhani, S. N. *et al.* Invasive meningococcal disease in patients with complement deficiencies: A case series (2008-2017). *BMC Infect. Dis.* **19**, 522 (2019).
41. Rosain, J. *et al.* Strains responsible for invasive meningococcal disease in patients with terminal complement pathway deficiencies. *J. Infect. Dis.* **215**, 1331–1338 (2017).
42. The International Meningococcal Genetics Consortium. Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nature Genetics* **42**, 772–776 (2010).
43. Lung, T. *et al.* The complement system in liver diseases: Evidence-based approach and therapeutic options. *J Transl Autoimmun* **2**, 100017 (2019).

44. Reis, E. S. *et al.* Applying complement therapeutics to rare diseases. *Clin. Immunol.* **161**, 225–240 (2015).
45. Seal, R. L. *et al.* Genenames.org: The HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).
46. Al-Hamoudi, W. K. Severe autoimmune hepatitis triggered by varicella zoster infection. *World J. Gastroenterol.* **15**, 1004–1006 (2009).
47. Brewer, E. C. & Hunter, L. Acute liver failure due to disseminated varicella zoster infection. *Case Reports Hepatol* **2018**, 1269340 (2018).
48. Eshchar, J., Reif, L., Waron, M. & Alkan, W. J. Hepatic lesion in chickenpox. A case report. *Gastroenterology* **64**, 462–466 (1973).
49. Sugiyama, K., Tagawa, S. & Toda, M. Methods for visual understanding of hierarchical system structures. *IEEE Trans. Syst. Man Cybern.* **11**, 109–125 (1981).
50. Srivastava, P., Tenney, J., Lodish, M., Slavotinek, A. & Baskin, L. Utility of genetic work-up for 46, XY patients with severe hypospadias. *J. Pediatr. Urol.* **19**, 261–272 (2023).
51. Utsch, B., Albers, N. & Ludwig, M. Genetic and molecular aspects of hypospadias. *Eur. J. Pediatr. Surg.* **14**, 297–302 (2004).
52. Liu, X. *et al.* The therapeutic target database: An internet resource for the primary targets of approved, clinical trial and experimental drugs. *Expert Opin. Ther. Targets* **15**, 903–912 (2011).
53. Díaz-Santiago, E. *et al.* Phenotype-genotype comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. *PLoS Genet.* **16**, e1009054 (2020).
54. Lord, J. & Baralle, D. Splicing in the diagnosis of rare disease: Advances and challenges. *Front. Genet.* **12**, 689892 (2021).
55. Bueren, J. A. & Auricchio, A. Advances and challenges in the development of gene therapy medicinal products for rare diseases. *Hum. Gene Ther.* **34**, 763–775 (2023).
56. Bulaklak, K. & Gersbach, C. A. The once and future gene therapy. *Nat. Commun.* **11**, 5820 (2020).
57. Godbout, K. & Tremblay, J. P. Prime editing for human gene therapy: Where are we now? *Cells* **12**, (2023).
58. Kohn, D. B., Chen, Y. Y. & Spencer, M. J. Successes and challenges in clinical gene therapy. *Gene Ther.* **30**, 738–746 (2023).
59. Zhao, Z., Shang, P., Mohanraju, P. & Geijsen, N. Prime editing: Advances and therapeutic applications. *Trends Biotechnol.* **41**, 1000–1012 (2023).
60. Darrow, J. J. Luxturna: FDA documents reveal the value of a costly gene therapy. *Drug Discov. Today* **24**, 949–954 (2019).
61. Mendell, J. R. *et al.* Single-Dose Gene-Replacement therapy for spinal muscular atrophy. *N. Engl. J. Med.* **377**, 1713–1722 (2017).
62. Mueller, C. *et al.* 5 year expression and neutrophil defect repair after gene therapy in alpha-1 antitrypsin deficiency. *Mol. Ther.* **25**, 1387–1394 (2017).

63. Russell, S. *et al.* Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: A randomised, controlled, open-label, phase 3 trial. *Lancet* **390**, 849–860 (2017).
64. Lu, C.-F. FDA takes first step toward international regulation of gene therapies to treat rare diseases. (2024).
65. Brown, D. G., Wobst, H. J., Kapoor, A., Kenna, L. A. & Southall, N. Clinical development times for innovative drugs. *Nat. Rev. Drug Discov.* **21**, 793–794 (2022).
66. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: An industry perspective. *Nat. Rev. Drug Discov.* **16**, 531–543 (2017).
67. Zhou, Q. & Buchholz, C. J. Cell type specific gene delivery by lentiviral vectors: New options in immunotherapy. *Oncoimmunology* **2**, e22566 (2013).
68. DiStefano, M. T. *et al.* The gene curation coalition: A global effort to harmonize gene-disease evidence resources. *Genet. Med.* **24**, 1732–1742 (2022).
69. DiStefano, M. *et al.* P451: The gene curation coalition works to resolve discrepancies in gene-disease validity assertions. *Genetics in Medicine Open* **1**, 100498 (2023).
70. Skene, N. G. & Grant, S. G. N. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* **10**, 16 (2016).
71. Côté, R. *et al.* The ontology lookup service: Bigger and better. *Nucleic Acids Res.* **38**, W155–60 (2010).
72. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).
73. Putman, T. E. *et al.* The monarch initiative in 2024: An analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Res.* **52**, D938–D949 (2024).
74. Lazarin, G. A. *et al.* Systematic classification of disease severity for evaluation of expanded carrier screening panels. *PLoS One* **9**, e114391 (2014).
75. Ahmed, M. I., Iqbal, M. & Hussain, N. A structured approach to the assessment of a floppy neonate. *J. Pediatr. Neurosci.* **11**, 2–6 (2016).
76. Chang, C.-W., Wakeland, A. K. & Parast, M. M. Trophoblast lineage specification, differentiation and their regulation by oxygen tension. *J. Endocrinol.* **236**, R43–R56 (2018).
77. Fogarty, N. M. E., Mayhew, T. M., Ferguson-Smith, A. C. & Burton, G. J. A quantitative analysis of transcriptionally active syncytiotrophoblast nuclei across human gestation. *J. Anat.* **219**, 601–610 (2011).
78. Hu, D. & Cross, J. C. Development and function of trophoblast giant cells in the rodent placenta. *Int. J. Dev. Biol.* **54**, 341–354 (2010).

Supplementary Materials

Supplementary Figures

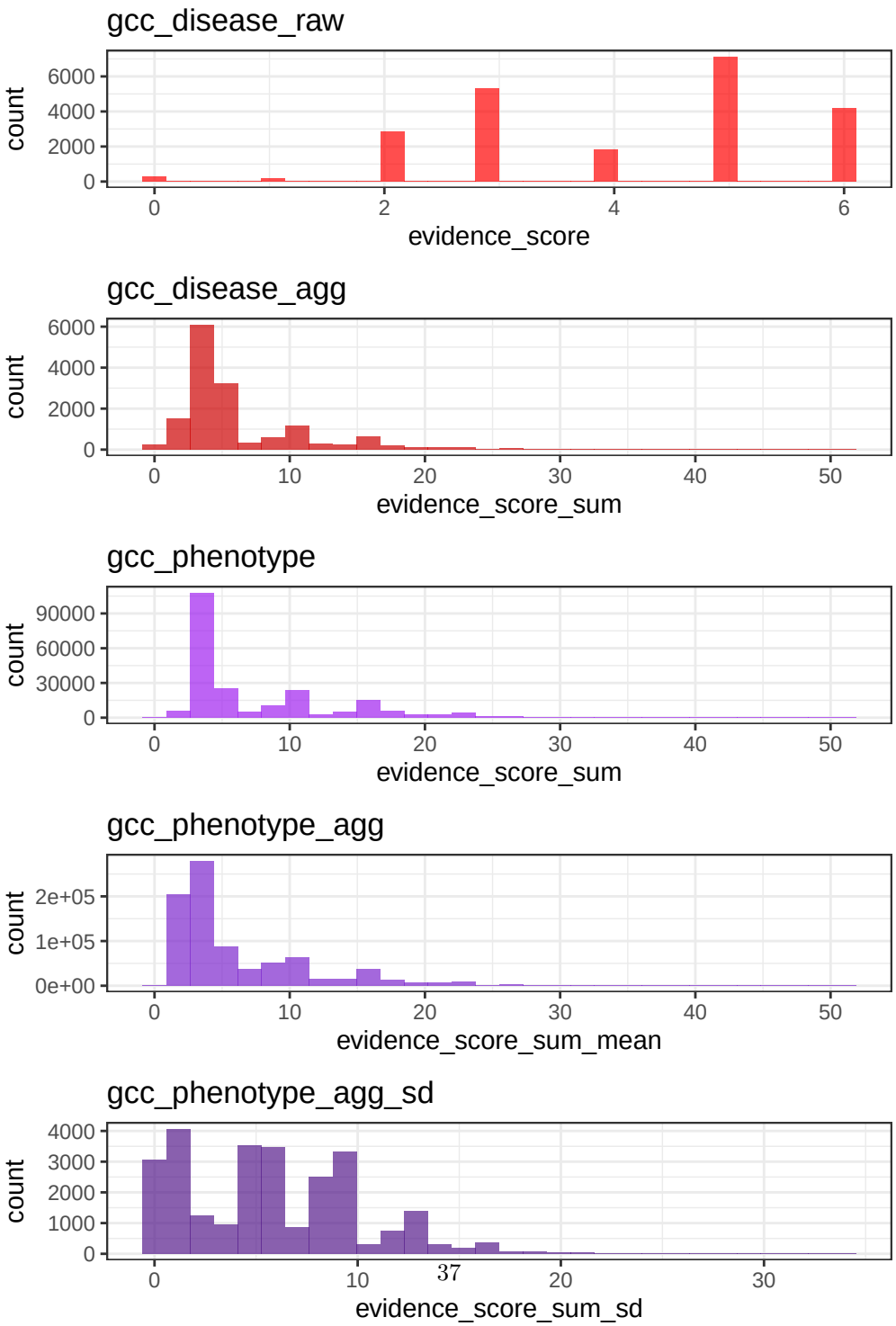


Figure 10: Distribution of evidence scores at each processing step.

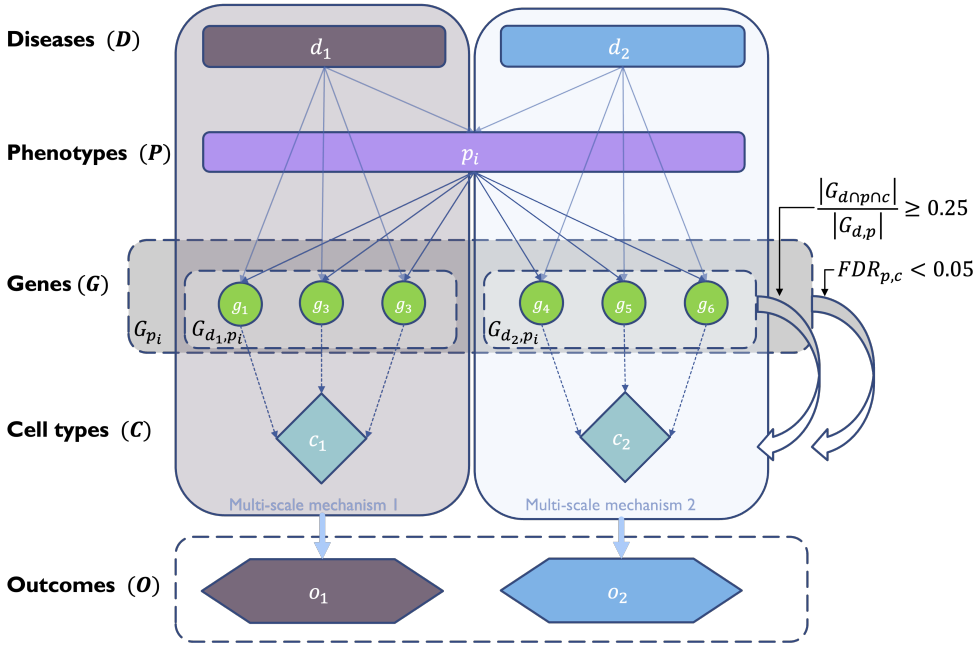
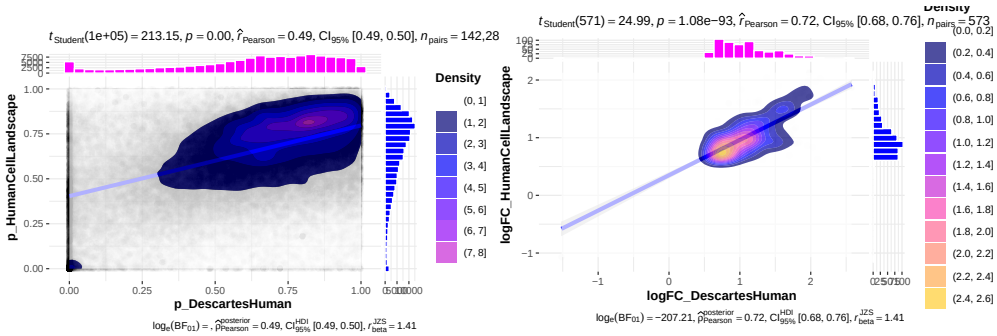
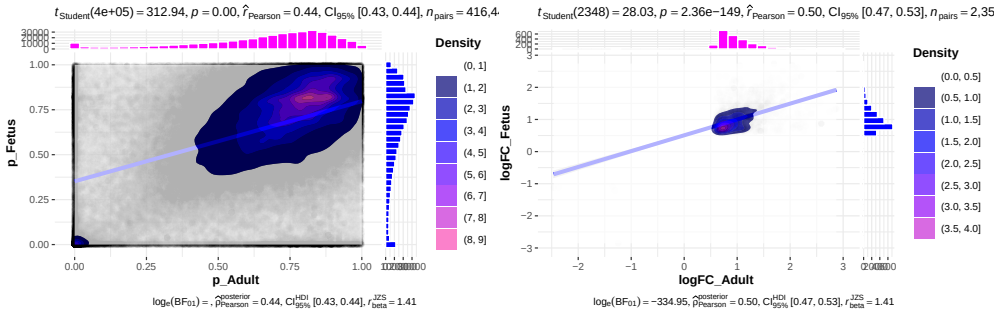


Figure 11: Diagrammatic overview of multi-scale disease investigation strategy. Here we provide an abstract example of differential disease aetiology across multiple scales: diseases (D), phenotypes (P), cell types (C), genes (G), and clinical outcomes (O). In the HPO, genes are assigned to phenotypes via particular diseases ($G_{d,p}$). Therefore, the final gene list for each phenotype is aggregated from across multiple diseases (G_p). We performed association tests for all pairwise combinations of cell types and phenotypes and filtered results after multiple testing corrections ($FDR_{p,c} < 0.05$). Each phenotype in the context of a given disease is referred to here as a symptom. Links were established between symptoms and cell types through proportional gene set overlap at a minimum threshold of 25%.



(a) Correlation between the uncorrected p-values from all phenotype-cell type association tests using the Descartes Human vs. Human Cell Landscape CTDs.

(b) Correlation between the $log_{10}(fold-change)$ from significant phenotype-cell type association tests ($FDR_{p,c} < 0.05$) using the Descartes Human vs. Human Cell Landscape CTDs.



(c) Correlation between the uncorrected p-values from all phenotype-cell type association tests using the Human Cell Landscape fetal samples vs. Human Cell Landscape adult samples.

(d) Correlation between the $log_{10}(fold-change)$ from significant phenotype-cell type association tests ($FDR_{p,c} < 0.05$) using the Human Cell Landscape fetal samples vs. Human Cell Landscape adult samples.

Figure 12: Inter- and intra-dataset validation across the different CellTypeDataset (CTD) and developmental stages. Correlations are computed using Pearson's correlation coefficient. Point density is plotted using a 2D kernel density estimate.

Figure 13: Therapeutics - Prioritised target filtering steps. This plot visualises the number of unique phenotype-cell type associations, cell types, genes, and phenotypes (*y-axis*) at each filtering step (*x-axis*) within the multi-scale therapeutic target prioritisation pipeline. Each step in the pipeline can be easily adjusted according to user preference and use case. See **Methods** for descriptions and criterion of each filtering step.**a**, The percentage of phenotypes belonging to each severity Tier after each filtering step (Tier 1 being the most severe). **b**, The number of phenotypes, cell types, associated diseases and genes remaining after each filtering step during the gene pri-

!!!RECOMPUTE!!!

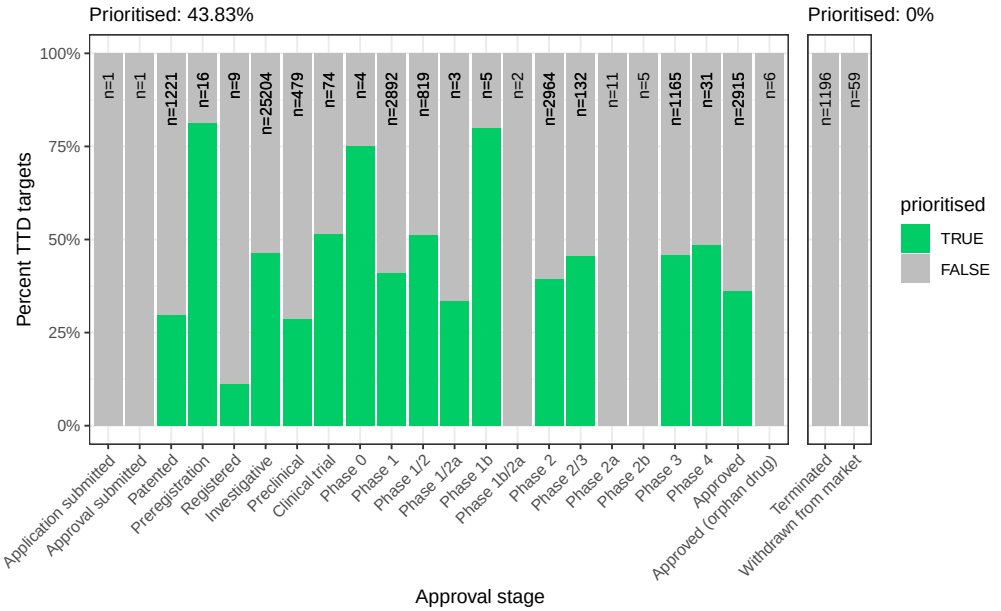


Figure 14: Therapeutics - Validation of prioritised therapeutic targets. Proportion of existing all therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline.

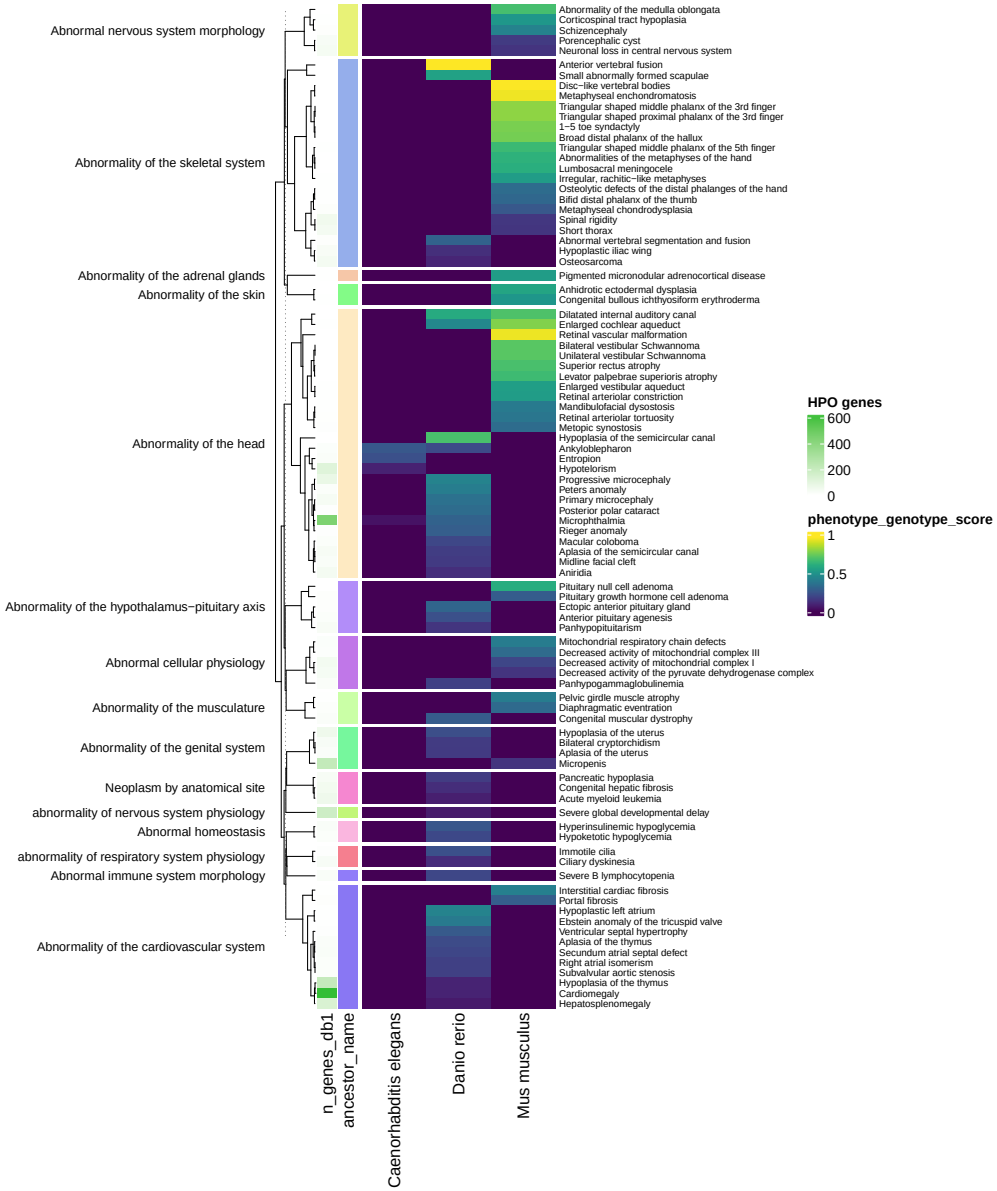


Figure 15: Identification of translatable experimental models. Interspecies translatability of human phenotypes nominated by our gene therapy prioritised pipeline. Above, our combined ontological-genotypic similarity score ($SIM_{o,g}$) is displayed as the heatmap fill colour stratified by the model organism (*x-axis*). An additional column ("n_genes_db1" on the far left) displays the total number of unique genes annotated to the phenotypic within the HPO. Phenotypes are clustered according to their ontological similarity in the HPO (*y-axis*).

Supplementary Methods

Therapeutics: Gene therapy target identification

Descriptions of each step in the prioritisation pipeline are as follows:

1. **start**: All phenotype-cell type association results.
2. **q_threshold**: Keep only results that were significant after multiple-testing correction (q<0.05).
3. **fold_threshold**: Keep only results with fold change>=1.
4. **keep_ont_levels**: Keep only phenotypes at certain absolute ontology levels within the HPO.
5. **keep_onsets**: Keep only phenotypes with postnatal age of onsets to circumvent technical and ethical challenges associated with antenatal gene therapeutics delivery.
6. **keep_tiers**: Keep only phenotypes with high severity Tiers.
    1. We used a combination of manual curation and automated text-based substring queries to assign each phenotype a severity Tier as characterised in a survey of healthcare professionals[74].
    2. Tier 1: Diseases that shortened life span in adolescence or earlier or resulted in intellectual disability.
    3. Tier 2: Diseases that shortened lifespan prematurely in adulthood, or resulted in impaired mobility or internal physical malformation.
    4. Tier 3: Diseases causing sensory impairments (hearing, vision, touch, pain, or other), immunodeficiency/cancer, mental illness, or dysmorphic features.
    5. Tier 4: Diseases that reduce fertility. Of the 49 phenotypes that were available in this severity ranking, we selected three that were classified as Tier 1 (the most severe disease category): mental deterioration, coma and respiratory failure.
7. **severity_threshold**: Keep only phenotypes with mean severity score equal to or below the threshold.
    1. Severity scores were computed by assigning each severity modifier term found in the HPO annotations a numerical value. In order of increasing severity:
    2. HP:0012825 “Mild” (Severity_score=4)
    3. HP:0012827 “Borderline” (Severity_score=3)
    4. HP:0012828 “Severe” (Severity_score=2)
    5. HP:0012829 “Profound” (Severity_score=1)
8. **pheno_frequency_threshold**: Keep only phenotypes with mean frequency equal to or above the threshold (i.e. how frequently a phenotype is associated with any diseases in which it occurs).
    1. Keep phenotypes with a mean frequency 10% or are NA by default.
9. **keep_celltypes**: Keep only terminally differentiated cell types.
    1. Of the 77 cell types tested in the Descartes cell type reference, the 40 terminally differentiated cell types were identified through a literature search. Of these, three (extravillous trophoblasts, syncytiotrophoblasts and trophoblast giant cells) were excluded as they only played a role in pregnancy[76–78], which

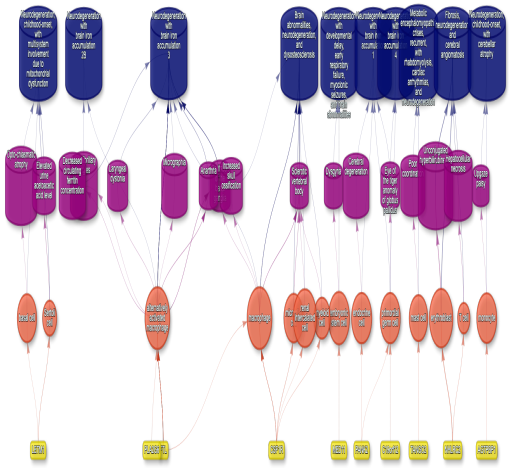
respiratory failure

(a) Respiratory failure

amyotrophic lateral sclerosis

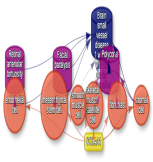
(b) Amyotrophic lateral sclerosis

neurodegeneration



(c) Neurodegeneration

small vessel disease



would raise additional technical and ethical challenges as rAAV therapy has not yet been used to target foetuses in clinical trials.

10. **keep_seqnames**: Remove genes on non-standard chromosomes.
    1. Only keep chromosomes 1-22, X, and Y.
11. **gene_size**: Keep only genes <4.3kb in length.
    1. Due to limitations in the length of the gene that can be carried by the rAAV vector, genes with a length of >4.3kb were excluded.
12. **keep_biotypes**: Keep only genes belonging to certain biotypes (e.g. "protein_coding", "processed_transcript", "snRNA", "lincRNA", "snoRNA", "IG_C_gene").
    1. Keep all biotypes by default.
13. **gene_frequency_threshold**: Keep only genes at or above a certain mean frequency threshold (i.e. how frequently a gene is associated with a given phenotype when observed within a disease).
    1. Keep genes with a mean frequency 10% or are NA by default.
14. **keep_specificity_quantiles**: Keep only genes in top specificity quantiles from the cell type dataset.
    1. To further narrow down genes, we extracted relevant metrics from the Descartes reference for each gene in each cell type. These included mean expression, specificity, and specificity quantiles (using 40 bins). Only genes with the most specific quantiles (39-40) were included for further analysis, as cell type-specific genes may be less likely to have off-target effects in other cell types.
15. **keep_mean_exp_quantiles**: Keep only genes in top mean expression quantiles from the cell type dataset
16. **end**: Final table of prioritised cell type- / phenotype-specific gene targets.

Finally, for more comprehensive target search, the we removed the filters for onsets (keep_onsets=NULL), Tier (keep_tiers=NULL), severity (severity_threshold=NULL), as well as relaxed the filters for phenotype frequency threshold (pheno_frequency_threshold=c(10,NA)), gene frequency threshold (gene_frequency_threshold = c(10,NA)), gene specificity quantiles (keep_specificity_quantiles = seq(20,40)), and gene expression quantiles (keep_mean_exp_quantiles = seq(20,40)).