

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

5 **Brian M. Schilder^{1,2}, Kitty B. Murphy^{1,2}, Robert Gordon-Smith^{1,2}, Jai
6 Chapman^{1,2}, Momoko Otani³, Nathan G. Skene^{1,2}**

7 ¹Department of Brain Sciences, Imperial College London, London, United Kingdom

8 ²UK Dementia Research Institute, London, United Kingdom

9 ³National Heart and Lung Institute, Imperial College London, London, United Kingdom

Corresponding author: Brian M. Schilder, brian_schilder@alumni.brown.edu

Corresponding author: Nathan G. Skene, n.skene@imperial.ac.uk

10 **1 Abstract**

11 Rare diseases (RDs) are an extremely heterogeneous and underserved category of
 12 medical conditions. While the majority of RDs are strongly genetic, it remains
 13 largely unknown via which physiological mechanisms genetics cause RD. Therefore,
 14 we sought to systematically characterise the cell type-specific mechanisms underlying
 15 all RD phenotypes with a known genetic cause by leveraging the Human Pheno-
 16 type Ontology and transcriptomic single-cell atlases of the entire human body from
 17 embryonic, foetal, and adult samples. In total we identified significant associations
 18 between 201 cell types and 9,563/11,015 (86.6%) unique phenotypes across 8,628
 19 RDs. We estimate that this represents an over 500-fold increase in the collective
 20 knowledge of RD phenotype-cell type mechanisms.

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

29 **2 Introduction**

30 While rare diseases (RDs) are individually uncommon, they collectively account
 31 for an enormous global disease burden with over 10,000 recognised RDs affecting
 32 at least 300-400 million people globally¹ (1 in 10-20 people)². Over 75% of RDs
 33 primarily affect children with a 30% mortality rate by 5 years of age³. Despite the
 34 prevalence and severity of RDs, patients suffering from these conditions are vastly
 35 underserved due to several contributing factors. First, diagnosis is extremely chal-
 36 lenging due to the highly variable clinical presentations of many of these diseases.
 37 The diagnostic odyssey can take patients and their families decades, with an average
 38 time to diagnosis of 5 years⁴. Of those, ~46% receive at least one incorrect diagno-
 39 sis and over 75% of all patients never receive any diagnosis⁵. Second, prognosis is
 40 also made difficult by high variability in disease course and outcomes which makes
 41 matching patients with effective and timely treatment plans even more challenging.
 42 Finally, even for patients who receive an accurate diagnosis/prognosis, treatments
 43 are currently only available for less than 5% of all RDs⁶. In addition to the sci-
 44 entific challenges of understanding RDs, there are strong financial disincentives for
 45 pharmaceutical and biotechnology companies to develop expensive therapeutics for
 46 exceedingly small RD patient populations with little or no return on investment^{7,8}.
 47 Those that have been produced are amongst the world's most expensive drugs,
 48 greatly limiting patients' ability to access it^{9,10}. The provision of timely, effective
 49 and affordable care for RD patients will require substantive transformations to our
 50 existing scientific, clinical, and regulatory frameworks.

51 A major challenge in both healthcare and scientific research is the scalable exchange
 52 of information. Even in the age of electronic healthcare records (EHR) much of
 53 the information about an individual's history is currently fractured across health-
 54 care providers, often with differing nomenclatures for the same conditions. The
 55 Human Phenotype Ontology (HPO) is a hierarchically organised set of controlled
 56 clinical terms that provides a much needed common framework by which clinicians
 57 and researchers can precisely communicate patient conditions¹⁴. The HPO spans
 58 all domains of human physiology and currently describes 18536 phenotypes across
 59 10,300 RDs. Each phenotype and disease is assigned its own unique identifier and
 60 organised as a hierarchical graph, such that higher-level terms describe broad pheno-
 61 typic categories or *branches* (e.g. *HP:0033127*: 'Abnormality of the musculoskeletal
 62 system' which contains 4522 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 time¹¹. 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 to 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 DECIPHER²⁵. 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 Human³¹ and Human Cell Landscape³² 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 566-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.

3 Results

3.1 Phenotype-cell type associations

In this study we systematically investigated the cell types underlying phenotypes across the HPO. A summary of the genome-wide results stratified by single-cell atlas can be found in Table 1. Within the results using the Descartes Human single-cell atlas, 19,894/ 847,077 (2.35%) tests across 77/ 77 (100%) cell types and 7,330/11,047 (66.4%) phenotypes revealed significant phenotype-cell type associations after multiple-testing correction ($FDR_{pc} < 0.05$). Using the Human Cell Landscape single-cell atlas, 26,543/1,357,304 (1.96%) tests across 124/124 (100%) cell types and 9,038/11,047 (81.8%) phenotypes showed significant phenotype-cell

116 type associations ($FDR_{pc} < 0.05$). The median number of significantly associated
 117 phenotypes per cell type was 252 (Descartes Human) and 200 (Human Cell Land-
 118 scape), respectively.

119 Across both single-cell references, the median number of significantly associated cell
 120 types per phenotype was 3, suggesting reasonable specificity of the testing strategy.
 121 8,628/8,631 (~100%) of diseases within the HPO gene annotations showed significant
 122 cell type associations for at least one of their respective phenotypes.

123 3.2 Validation of expected phenotype-cell type relationships

124 Within each high-level branch in the HPO shown in Fig. 1b, we tested whether each
 125 cell type was more often associated with phenotypes in that branch relative to those
 126 in all other branches (including those not shown). We then checked whether each
 127 cell type was overrepresented (at $FDR_{bc} < 0.05$) within its respective on-target
 128 HPO branch, where the number of phenotypes within that branch (N_p) Abnormal-
 129 ity of the cardiovascular system: 5/95 types of ‘cardiocyte’ were overrepresented
 130 ($N_p=673$). Abnormality of the endocrine system: 3/84 types of ‘endocrine cell’ were
 131 overrepresented ($N_p=291$). Abnormality of the eye: 5/97 types of ‘photoreceptor
 132 cell/retinal cell’ were overrepresented ($N_p=721$). Abnormality of the immune sys-
 133 tem: 14/82 types of ‘leukocyte’ were overrepresented ($N_p=253$). Abnormality of
 134 the musculoskeletal system: 4/99 types of ‘cell of skeletal muscle/chondrocyte’ were
 135 overrepresented ($N_p=2153$). Abnormality of the nervous system: 17/98 types of
 136 ‘neural cell’ were overrepresented ($N_p=1645$). Abnormality of the respiratory system:
 137 3/89 types of ‘respiratory epithelial cell/epithelial cell of lung’ were overrepresented
 138 ($N_p=291$)..

139 As an additional form of validation (Fig. 1d), we tested for a relationship between
 140 phenotype-cell type association significance ($-\log_e(p_{pc})$ where \log_e denotes natu-
 141 ral log and and p_{pc} denotes uncorrected phenotype-cell type association p-values)
 142 and the proportion of on-target cell types. The list of on-target cell types were de-
 143 termined by matching each high-level HPO branch to a corresponding CL branch.
 144 These cross-ontology mappings can be found in Table 4. For this analysis we used
 145 raw p-values (p_{pc}) rather than multiple-testing corrected p-values (FDR_{pc}) to pro-
 146 vide a more dynamic range of values (as the latter can drive values to 1). All 7/7
 147 high-level HPO branches showed a consistent upwards trend towards greater pro-
 148 portions of on-target cell types with increasing degrees of significance. Furthermore,
 149 all branches also showed a proportion of on-target cell types above that expected by
 150 chance (baseline = on-target cell types / total cell types) at $-\log_e(p_{pc}) > 1$.

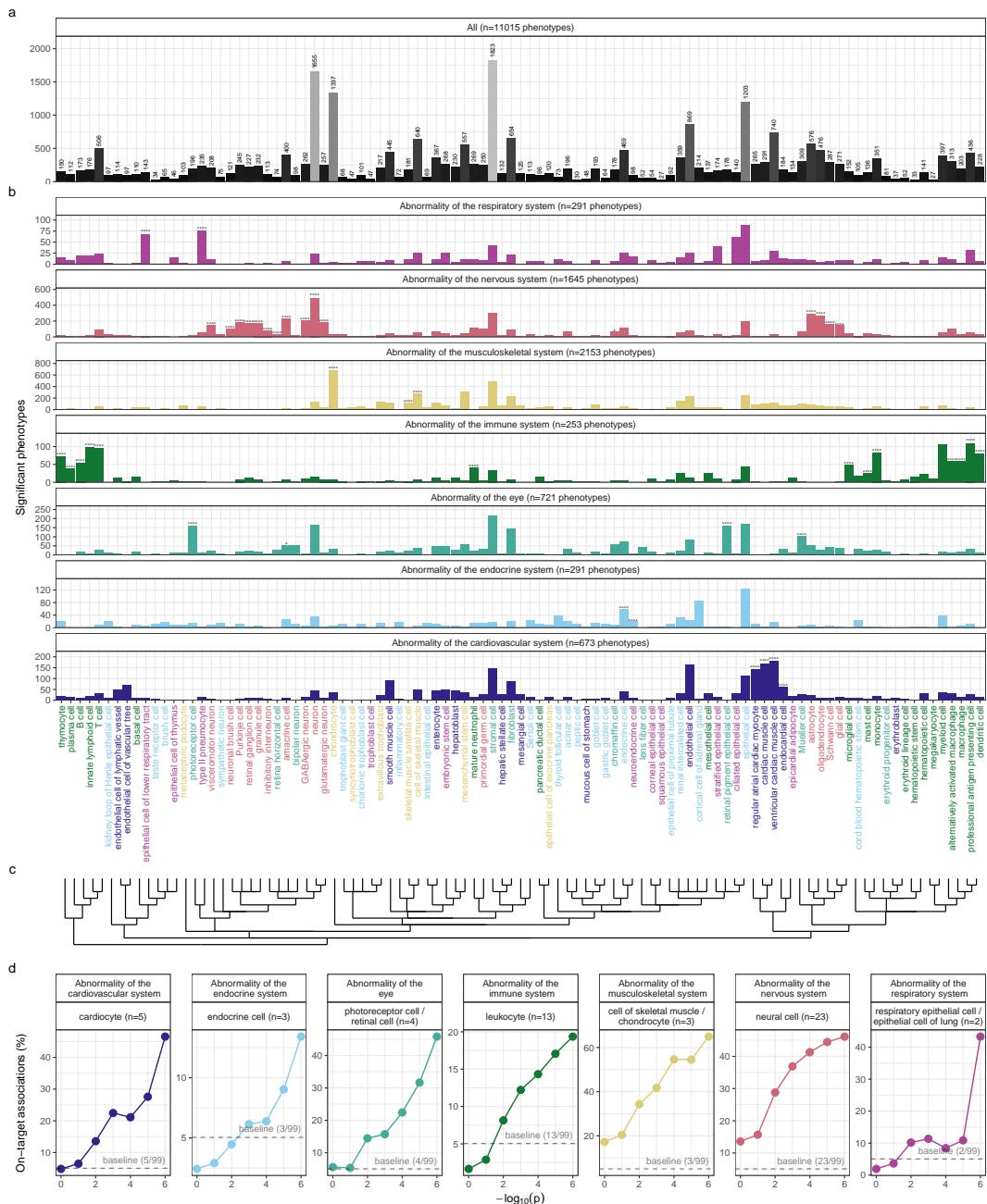


Figure 1: Summary of significant associations between phenotypes and cell types, aggregated by HPO branch. Here we show **a**, the total number of significant phenotype enrichments per cell type ($FDR_{pc} < 0.05$) across all branches of the HPO. **b**, Number of phenotype association related to several 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; $FDR_{b,c} < 1e - 04$ (****), $FDR_{b,c} < 0.001$ (**), $FDR_{bc} < 0.01$ (**), $FDR_{b,c} < 0.05$ (*). **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 CL³³ 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).

151 **3.3 Validation of inter- and intra-dataset consistency**

152 Next, we sought to validate the consistency of our results across the two single-cell
 153 reference datasets (Descartes Human vs. Human Cell Landscape) across the subset
 154 of overlapping cell types Fig. 13. In total there were 142116 phenotype-cell type
 155 associations to compare across the two datasets (across 10932 phenotypes and 13
 156 cell types annotated to the exact same CL term. We found that the correlation be-
 157 tween p-values of the two datasets was high ($\rho = 0.492, p = 2.31e - 93$). Within
 158 the subset of results that were significant in both single-cell datasets ($FDR_{pc} <$
 159 0.05), we found that correlation of the association effect size were even stronger
 160 ($\rho = 0.722, p = 2.31e - 93$). We also checked for the intra-dataset consistency
 161 between the p-values of the foetal and adult samples in the Human Cell Land-
 162 scape, showing a very similar degree of correlation as the inter-dataset comparison
 163 ($\rho = 0.436, p = 2.74e - 149$). Together, these results suggest that our approach to
 164 identifying phenotype-cell type associations is highly replicable and generalisable to
 165 new datasets.

166 **3.4 More specific phenotypes are associated with fewer genes and cell
 167 types**

168 First, we found that phenotype ontology showed a significant negative correlation
 169 with the number of genes annotated to that phenotype in the HPO data (Fig. 2a;
 170 $p = 2.23e - 308, q = 2.23e - 308, \rho = -0.267$). This is expected as broader phe-
 171 notypes tend to have large gene set annotations. Next, we reasoned that lower HPO
 172 ontology levels representing more specific phenotypes were likely to be associated
 173 with fewer, more specific subsets of cell types. This was indeed the case, as we ob-
 174 served a strongly significant negative correlation between the two variables (Fig. 2b;
 175 $p = 2.23e - 308, q = 2.23e - 308, \rho = -0.296$). We also found that the effect
 176 size of significant phenotype-cell type associations ($FDR_{pc} < 0.05$) increased with
 177 greater phenotype specificity, though the relationship was rather weak (Fig. 2c; NA).
 178 Finally, we found that the mean expression specificity of phenotype-associated genes
 179 (within the cell types significantly associated with those respective phenotypes at
 180 $FDR_{pc} < 0.05$) was positively correlated phenotype ontology depth (Fig. 2d; NA).

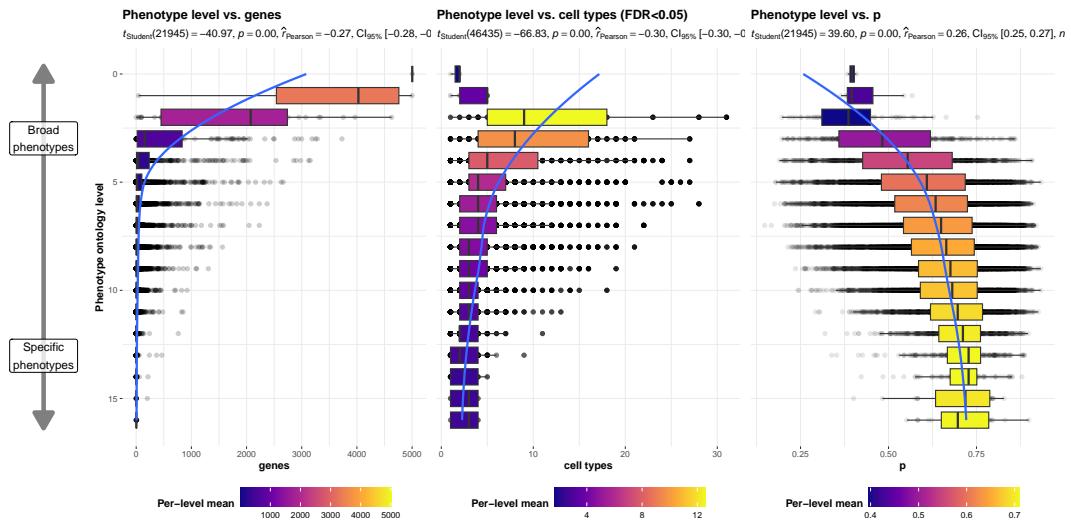


Figure 2: More specific phenotypes are associated with fewer, more specific genes and cell types. Box plots showing relationship between HPO phenotype level and **a**, the number of genes annotated to each phenotype, **b**, the number of significantly enriched cell types, **c**, the effect size of phenotype-cell type association tests at $FDR_{pc} < 0.05$, and **d**, the mean expression specificity of phenotype-associated genes in the cell types significantly associated with those respective phenotypes ($FDR_{pc} < 0.05$). 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.

181 3.5 Hepatoblasts have a unique role in recurrent Neisserial infections

182 We selected the HPO term ‘Recurrent bacterial infections’ and all of its descendants
 183 (19 phenotypes) as an example of how investigations at the level of granular pheno-
 184 types can reveal different cell type-specific mechanisms (Fig. 3). As expected, these
 185 phenotypes are primarily associated with immune cell types (e.g. macrophages, den-
 186 dritic cells, T cells, monocytes, neutrophils). Some associations confirm relationships
 187 previously suggested in the literature, such as that between ‘Recurrent staphylococ-
 188 cal infections’ and myeloid cells^{34–37}. Specifically, our results pinpoint monocytes as
 189 the most strongly associated cell subtypes ($FDR_{pc} = 1.03e - 30, B = 1.76e - 01$).

190 In contrast to all other recurrent infection types, ‘Recurrent Neisserial infections’
 191 highlighted a novel association with hepatoblasts (Descartes Human : $FDR_{pc} =$
 192 $1.13e - 06, B = 8.24e - 02$). Whilst unexpected, a convincing explanation involves
 193 the complement system, a key driver of innate immune response to Neisserial infec-
 194 tions. Hepatocytes, which derive from hepatoblasts, produce the majority of com-
 195 plement proteins³⁸, and Kupffer cells express complement receptors³⁹. In addition,
 196 individuals with deficits in complement are at high risk for Neisserial infections^{40,41},
 197 and a genome-wide association study in those with a Neisserial infection identified
 198 risk variants within complement proteins⁴². While the potential of therapeutically
 199 targeting complement in RDs (including Neisserial infections) has been proposed
 200 previously^{43,44}, performing this in a gene- and cell type-specific manner may help
 201 to improve efficacy and reduce toxicity (e.g. due to off-target effects). Importantly,
 202 there are over 56 known genes within the complement system⁴⁵, highlighting the
 203 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’ ($FDR_{pc} = 2.08e - 02, B = 5.25e - 02$) and ‘Susceptibility to chickenpox’ ($FDR_{pc} = 1.20e - 02, B = 5.49e - 02$) both of which are well-known to involve the liver^{46–48}.

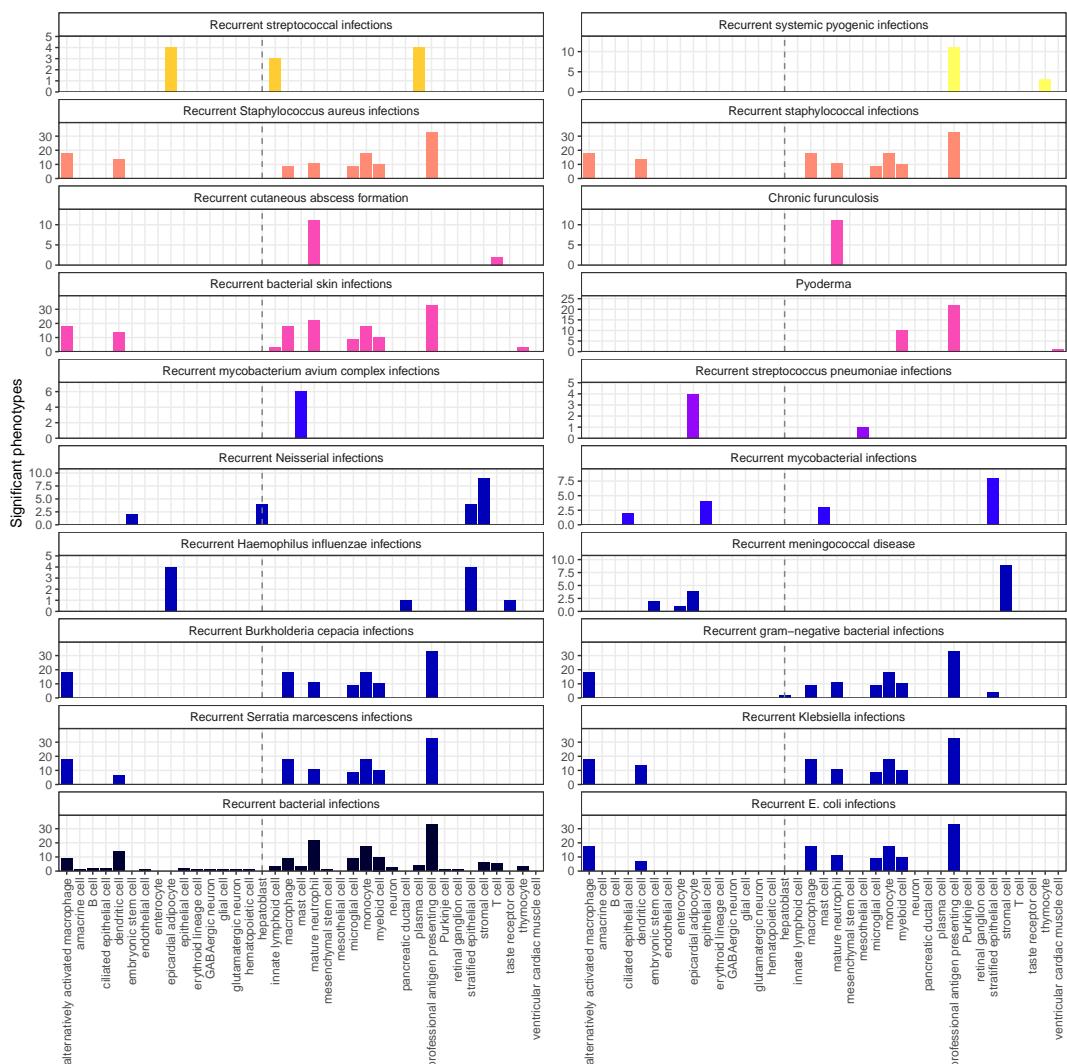


Figure 3: Hepatoblasts have a unique role in recurrent Neisserial infections. Significant phenotype-cell type tests for phenotypes within the branch ‘Recurrent bacterial infections’. Amongst all different kinds of recurrent bacterial infections, hepatoblasts (highlighted by vertical dotted lines) are 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’).

216 Next, we sought to link multi-scale mechanisms at the levels of disease, phenotype,
 217 cell type, and gene and visualise these as a network (Fig. 4). This revealed that
 218 genetic deficiencies in different complement system genes (*C5*, *C8*, and *C7*) are pri-
 219 marily mediated by different cell types (hepatoblasts, stratified epithelial cells, and
 220 stromal cells, respectively). While genes of the complement system are expressed
 221 throughout many different tissues and cell types, these results indicate that different
 222 subsets of these genes may mediate their effects through different cell types. This
 223 finding suggests that investigating (during diagnosis) and targeting (during treat-
 224 ment) different cell types may be critical for the diagnosis and treatment of these
 225 closely related, yet mechanistically distinct, diseases.

226 **3.6 Monarch Knowledge Graph recall**

227 Next, we used the Monarch Knowledge Graph (MKG) as a proxy for the field's
 228 current state of knowledge of phenotype-cell type associations. We evaluated the
 229 proportion of MKG associations that were recapitulation by our results Fig. 14. For
 230 each phenotype-cell type association in the MKG, we computed the percent of cell
 231 types recovered in our association results at a given ontological distance according to
 232 the CL ontology. An ontological distance of 0 means that our nominated cell type
 233 was as close as possible to the MKG cell type after adjusting for the cell types avail-
 234 able in our single-cell references. Instances of exact overlap of terms between the
 235 MKG and our results would qualify as an ontological distance of 0 (e.g. 'monocyte'
 236 vs. 'monocyte'). Greater ontological distances indicate further divergence between
 237 the MKG cell type and our nominated cell type. A distance of 1 indicating that the
 238 MKG cell type was one step away from our nominated cell type in the CL ontology
 239 graph (e.g. 'monocyte' vs. 'classical monocyte'). The maximum possible percent of
 240 recovered terms is capped by the percentage of MKG ground-truth phenotypes we
 241 were able to find at least one significant cell type association for at FDR_{pc} .

242 In total, our results contained at least one significant cell type associations for 90.2%
 243 of the phenotypes described in the MKG. Of these phenotypes, we captured 54.9%
 244 of the MKG phenotype-cell associations at an ontological distance of 0 (i.e. the clos-
 245 est possible Cell Ontology term match). Recall increased with greater flexibility in
 246 the matching of cell type annotations. At an ontological distance of 1 (e.g. 'mono-
 247 cyte' vs. 'classical monocyte'), we captured 78% of the MKG phenotype-cell asso-
 248 ciations. Recall reached a maximum of 90.2% at a ontological distance of 5. This
 249 recall percentage is capped by the proportion of phenotype for which we were able
 250 to find at least one significant cell type association for. It should be noted that we
 251 were unable to compute precision as the MKG (and other knowledge databases)
 252 only provide true positive associations. Identifying true negatives (e.g. a cell type is
 253 definitely never associated with a phenotype) is a fundamentally more difficult task
 254 to resolve as it would require proving the null hypothesis. Regardless, these bench-
 255 marking tests suggests that our results are able to recover the majority of known
 256 phenotype-cell type associations while proposing many new associations.

257 **3.7 Annotation of phenotypes using generative large language models**

258 Severity annotations were gathered from GPT-4 for 16880/18536 (91.06603%) HPO
 259 phenotypes. In our companion study, benchmarking tests of these results using
 260 ground-truth HPO branch annotations. For example, phenotypes within the 'Blind-
 261 ness' HPO branch (*HP:0000618*) were correctly annotated as causing blindness by
 262 GPT-4. Across all annotations, the recall rate of GPT-4 annotations was 91.43%
 263 (min=70.31%, max=100%, SD=11.76) with a mean consistency score of 91.21%
 264 (min=80.96%, max=97.48%, SD=5.739) for phenotypes whose annotation were col-
 265 lected more than once. This clearly demonstrates the ability of GPT-4 to accurately
 266 annotate phenotypes. This allowed us to begin using these annotations to compute
 267 systematically collected severity scores for all phenotypes in the HPO.

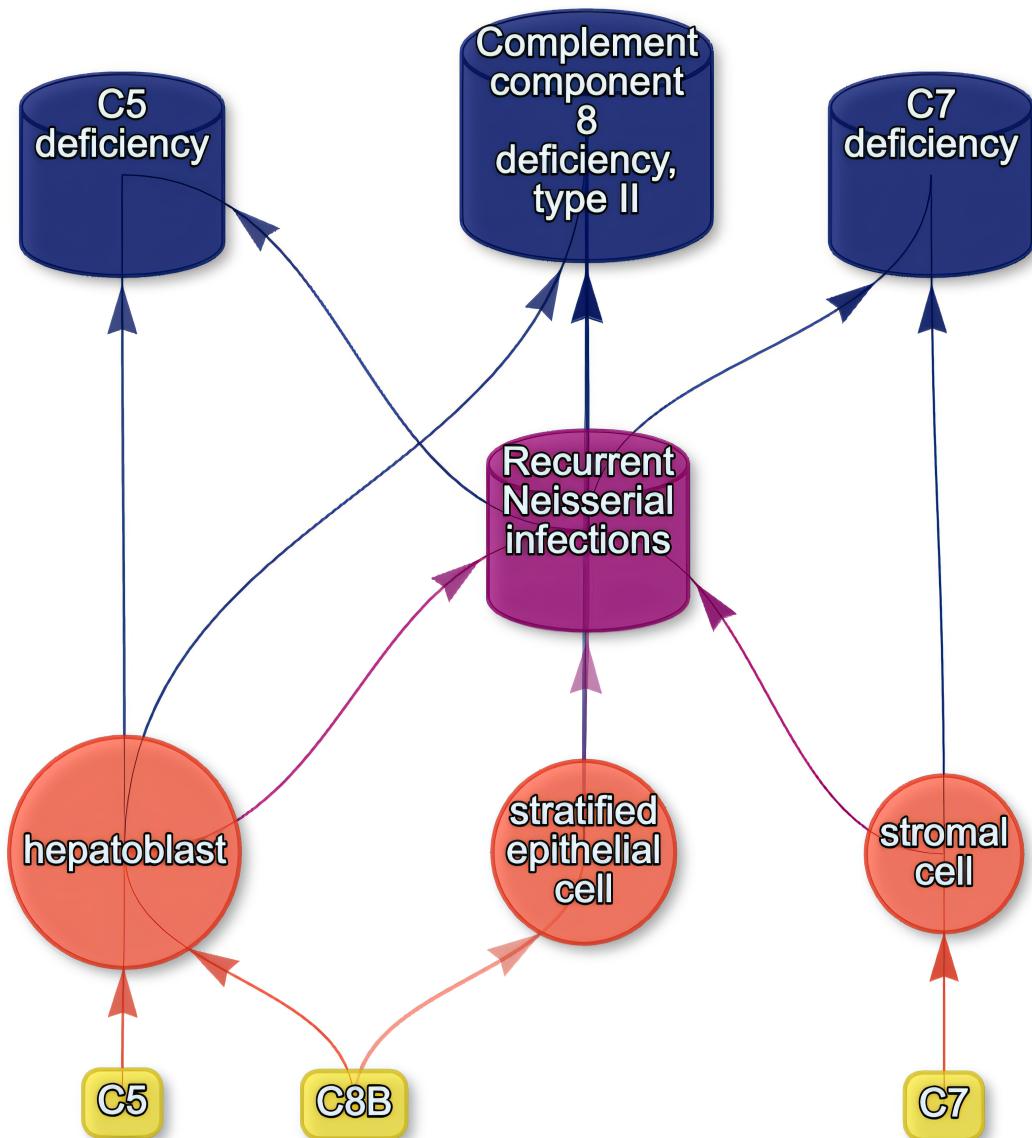


Figure 4: Multi-scale mechanisms of Recurrent Neisserial infections. Starting from the bottom of the plot, one can trace how causal genes (yellow boxes) mediate their effects through cell types (orange circles), phenotypes (purple cylinders) and ultimately diseases (blue cylinders). Cell types are connected to phenotypes via association testing ($FDR_{pc} < 0.05$), and to diseases when the symptom gene set overlap is $>25\%$. Nodes were spatially arranged using the Sugiyama algorithm⁴⁹.

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).

3.8 Congenital phenotypes are associated with foetal cell types

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 ($p = 2.3e - 200$, $\chi^2_{Pearson} = 926$, $\hat{V}_{Cramer} = 0.14$). Furthermore, increasing congenital frequency annotation (on an ordinal scale) corresponded to an increase in the proportion of foetal cell types: ‘always’=24% (n=1626 associations), ‘often’=20% (n=2965 associations), ‘rarely’=12% (n=1954 associations), ‘never’=10% (n=806 associations). This is consistent with the expected role of foetal cell types in development and the aetiology of congenital disorders.

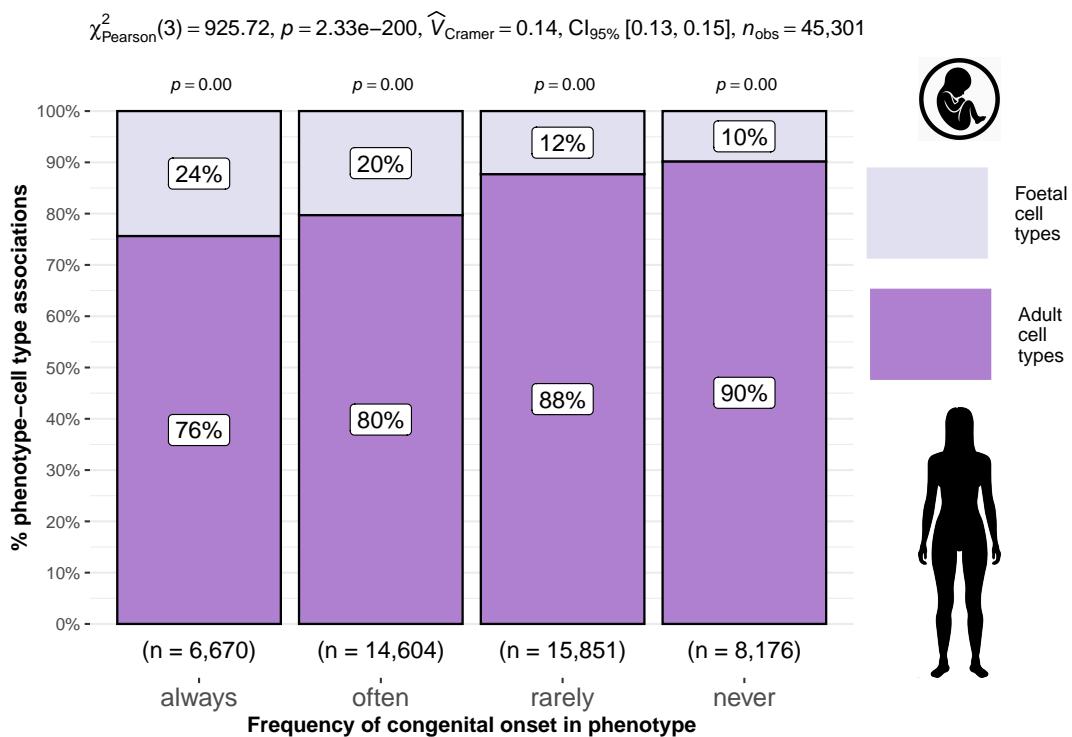


Figure 5: Congenital phenotypes are more often associated with foetal cell types. As a phenotype is more often congenital in nature, the greater proportion of foetal cell types are significantly associated with it. The summary statistics in the plot title are the results of a χ^2 tests of independence between the ordinal scale of congenital onset and the proportion of foetal cell types associated with each phenotype. The p-values above each bar are the results of an additional series of χ^2 tests to determine whether the proportion of foetal vs. non-foetal cell types significantly different differs from the proportions expected by chance. The foetal silhouette was generated with DALL-E. The adult silhouette is from phylopic.org and is freely available via CC0 1.0 Universal Public Domain Dedication.

285 **3.9 Diagnosis via cell type-specific disease prediction**

286 Using the function `MSTExplorer::predict_celltypes` we input 3 inclusion phe-
 287 notypes ('Generalized neonatal hypotonia' (*HP:0008935*), 'Scrotal hypospadias'
 288 (*HP:0012853*), 'Increased circulating progesterone' (*HP:0031216*)), 2 genes in
 289 which the patient is known to have deleterious mutations (*HSD3B2*, *HERC2*)
 290 and 1 gene in which the patient is known not to have any deleterious mutations
 291 (*SNORD115-1*). This predicted that cortical cell of adrenal gland (score sum=1.38,
 292 score mean=0.0256, score standard deviation=0.137) were the most probable cell
 293 types underlying this combination of phenotypes and genotypes (Fig. 6), which is
 294 highly consistent with existing evidence that adrenal insufficiency can cause both
 295 phenotypes via mutations in these genes^{50,51}. This was the only cell type to receive
 296 a score two standard deviations from the mean score of all cell types (mean score:
 297 0.000668).

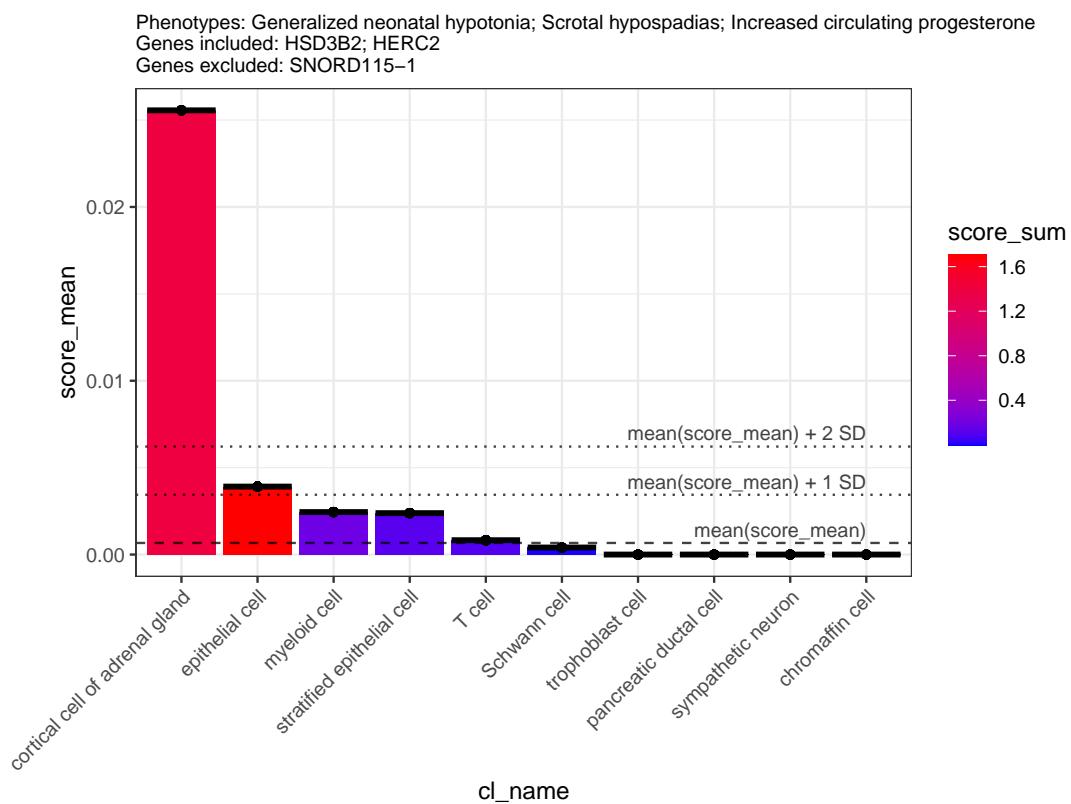


Figure 6: Diagnosis - Observed phenotypes/genotypes can be used to identify causal cell types in individuals. Our phenotype-cell type association results can be used to make predictions about which cell types are underlying a set of phenotypes observed in a given patient. Here we input three inclusion 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.

298 **3.10 Prognosis via cell type-mediated differential outcomes**

299 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
300 HPO gene annotations. Together, these hypotonia phenotypes were significantly
301 associated with 29/99 (29.29%) unique CL-aligned cell types. This reflects the highly
302 variable set of disease etiologies that can cause this broad-level phenotype. Across all
303 diseases, hypotonia phenotypes tended to be most consistently severe (lower mean
304 age of death score) when associated with the cell type inhibitory interneuron. While
305 other cell types were associated with lower mean age of death scores (e.g. stromal
306 cell, astrocyte), the severity of the outcomes were more variable.
307

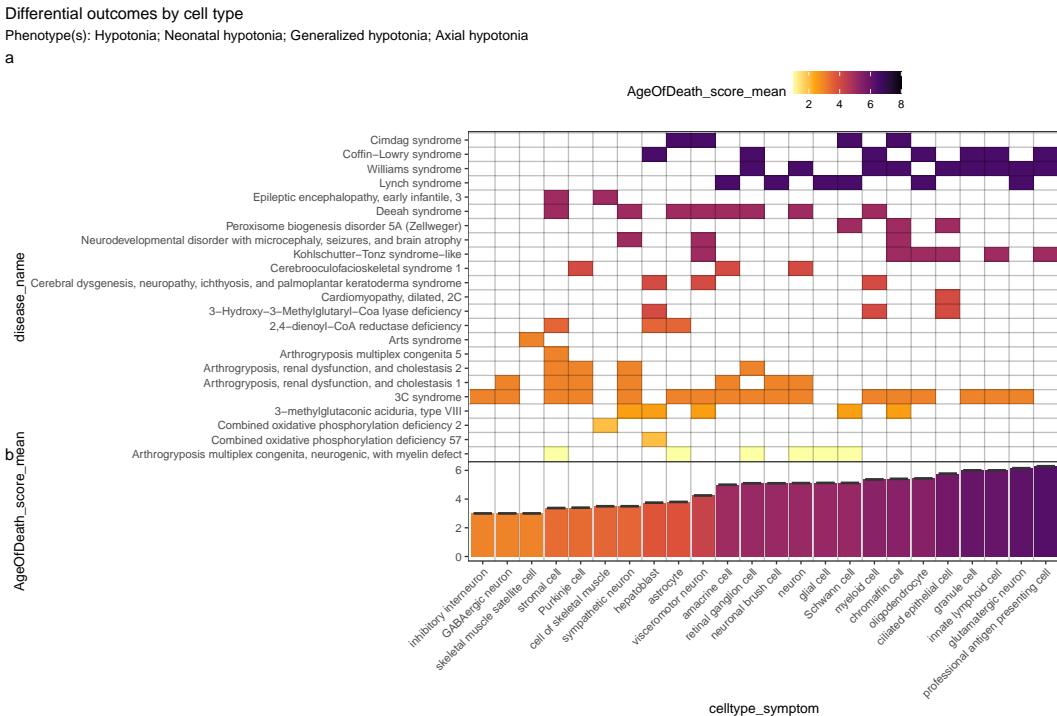


Figure 7: Prognosis - Cell types predict the probability of deadly diseases. The broad phenotype ‘Hypotonia’ and its descendants occur in many different diseases (1,832 diseases in the HPO annotations). Therefore, it can be difficult to prognose clinical outcomes of a newborn individual with hypotonia. 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 (Table 5) 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.

308 **3.11 Therapeutic target identification**

309 Next, we identified putative cell type-specific gene targets for several severe disease
310 phenotypes. This yielded putative therapeutic targets for 5244 phenotypes across
311 4805 diseases in 201 cell types and 3141 genes (Fig. 8). While this constitutes a large

312 number of genes in total, each phenotype was assigned a median of 2 gene targets
 313 (mean=3.25, min=1, max=10). Relative to the number of genes annotations per
 314 phenotype in the HPO overall (median=7, mean=62.16, min=1, max=5003) this
 315 represents a substantial decrease in the number of candidate target genes, even when
 316 excluding high-level phenotypes (HPO level>3). It is also important to note that
 317 the phenotypes in the prioritised targets list are ranked by their severity, allowing
 318 us to distinguish between phenotypes with a high medical urgency (e.g. ‘Hydranen-
 319 cephaly’) from those with lower medical urgency (e.g. ‘Hyperplastic labia majora’).
 320 This can be useful for both clinicians, biomedical scientists, and pharmaceutical
 321 manufacturers who wish to focus their research efforts on phenotypes with the great-
 322 est need for intervention.

323 Across all phenotypes, epithelial cell were most commonly implicated (837 pheno-
 324 types), followed by stromal cell (627 phenotypes), stromal cell (627 phenotypes),
 325 neuron (474 phenotypes), chondrocyte (382 phenotypes), and endothelial cell (360
 326 phenotypes). Grouped by higher-order ontology category, ‘Abnormality of the mus-
 327 culoskeletal system’ had the greatest number of enriched phenotypes (957 pheno-
 328 types, 849 genes), followed by ‘Abnormality of the nervous system’ (731 phenotypes,
 329 1135 genes), ‘Abnormality of head or neck’ (544 phenotypes, 981 genes), ‘Abnormal-
 330 ity of the genitourinary system’ (442 phenotypes, 693 genes), and ‘Abnormality of
 331 the eye’ (377 phenotypes, 549 genes).

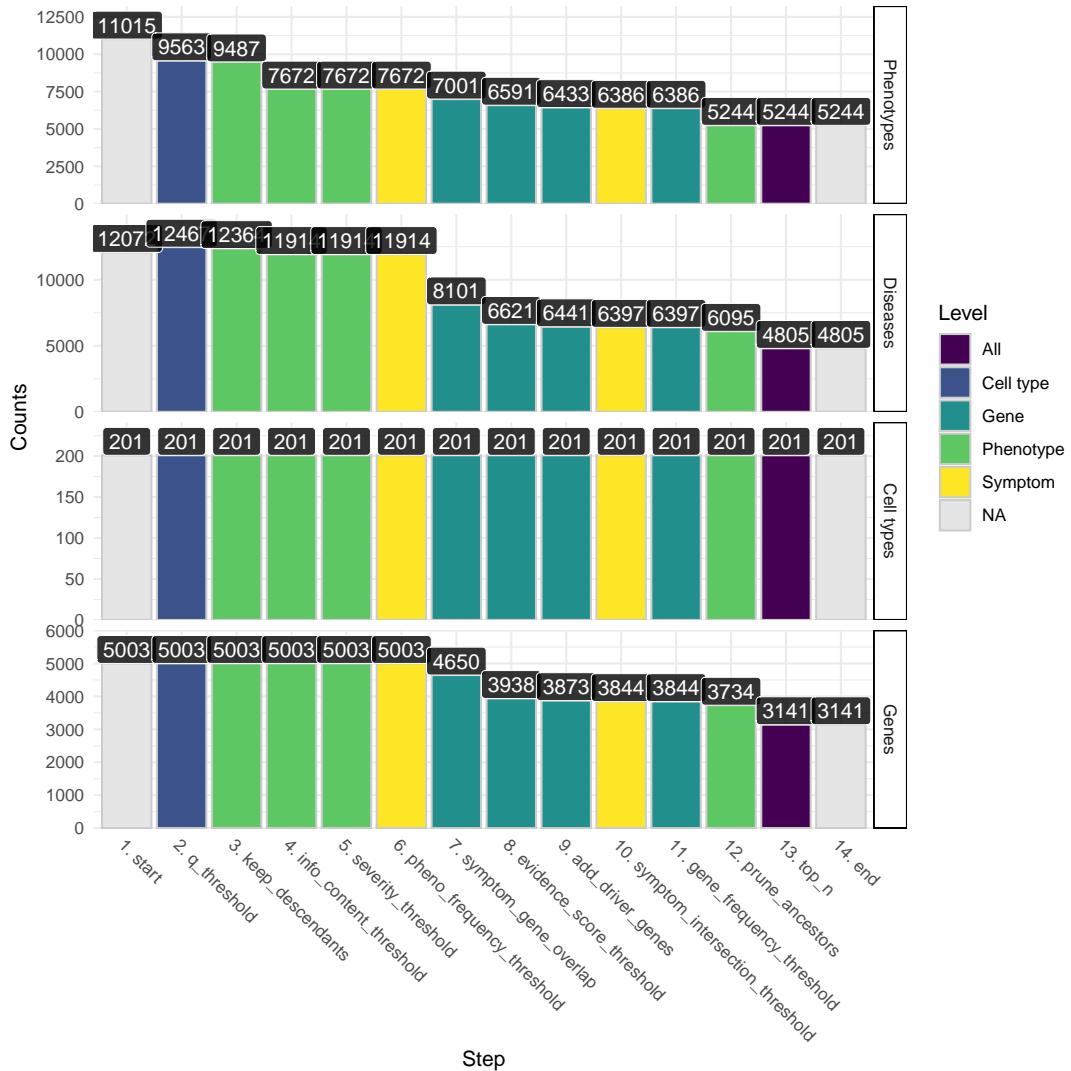


Figure 8: Therapeutics - Prioritised target filtering steps. This plot visualises the number of unique phenotype-cell type associations, cell types, genes, and phenotypes (*y-axis*) at each filtering step (*x-axis*) within the multi-scale therapeutic target prioritisation pipeline. Each step in the pipeline can be easily adjusted according to user preference and use case. See Table 2 for descriptions and criterion of each filtering step.

3.12 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 2024-04-15)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 ($p = 0.007$). 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 ($p = 0.372$), 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. 15), not just gene therapies, we recapitulated 39% 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 ($p = 1$), and highly significantly depleted for failed therapeutics ($p = 3e - 195$). This suggests that our multi-scale evidence-based prioritisation pipeline is capable of selectively identifying genes that are likely to be effective therapeutic targets.

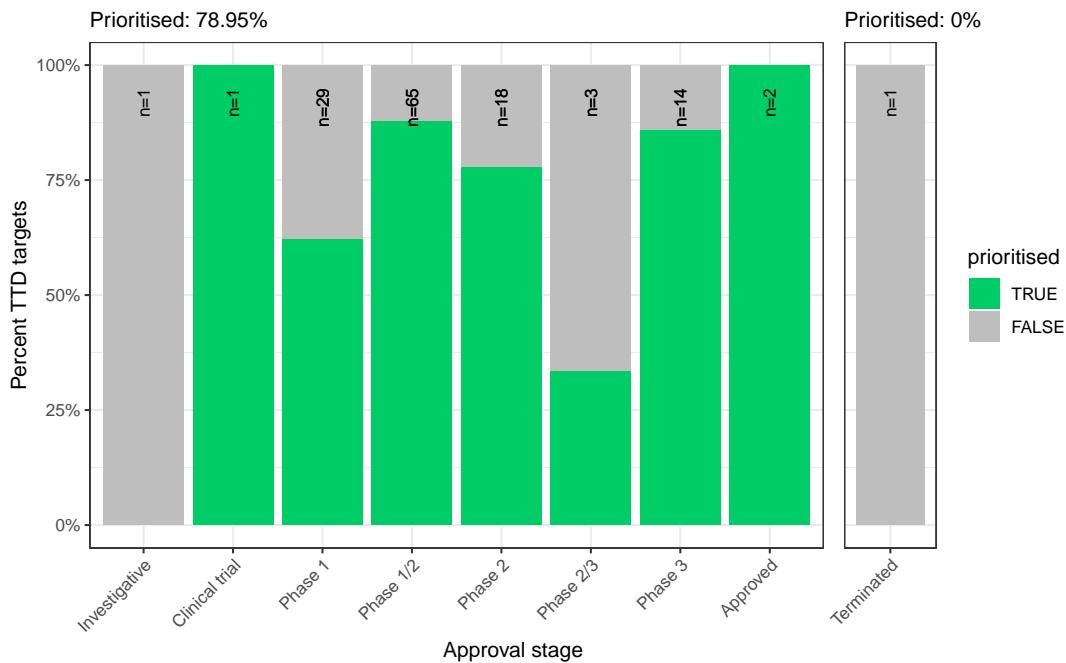


Figure 9: Therapeutics - Validation of prioritised therapeutic targets. The proportion of existing gene therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline. Therapeutics are stratified by the stage of clinical development they were at during the time of writing.

3.13 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’. Only phenotypes with a GPT severity score greater than 15 were considered to avoid overplotting and to focus on the more clinically relevant phenotypes.

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

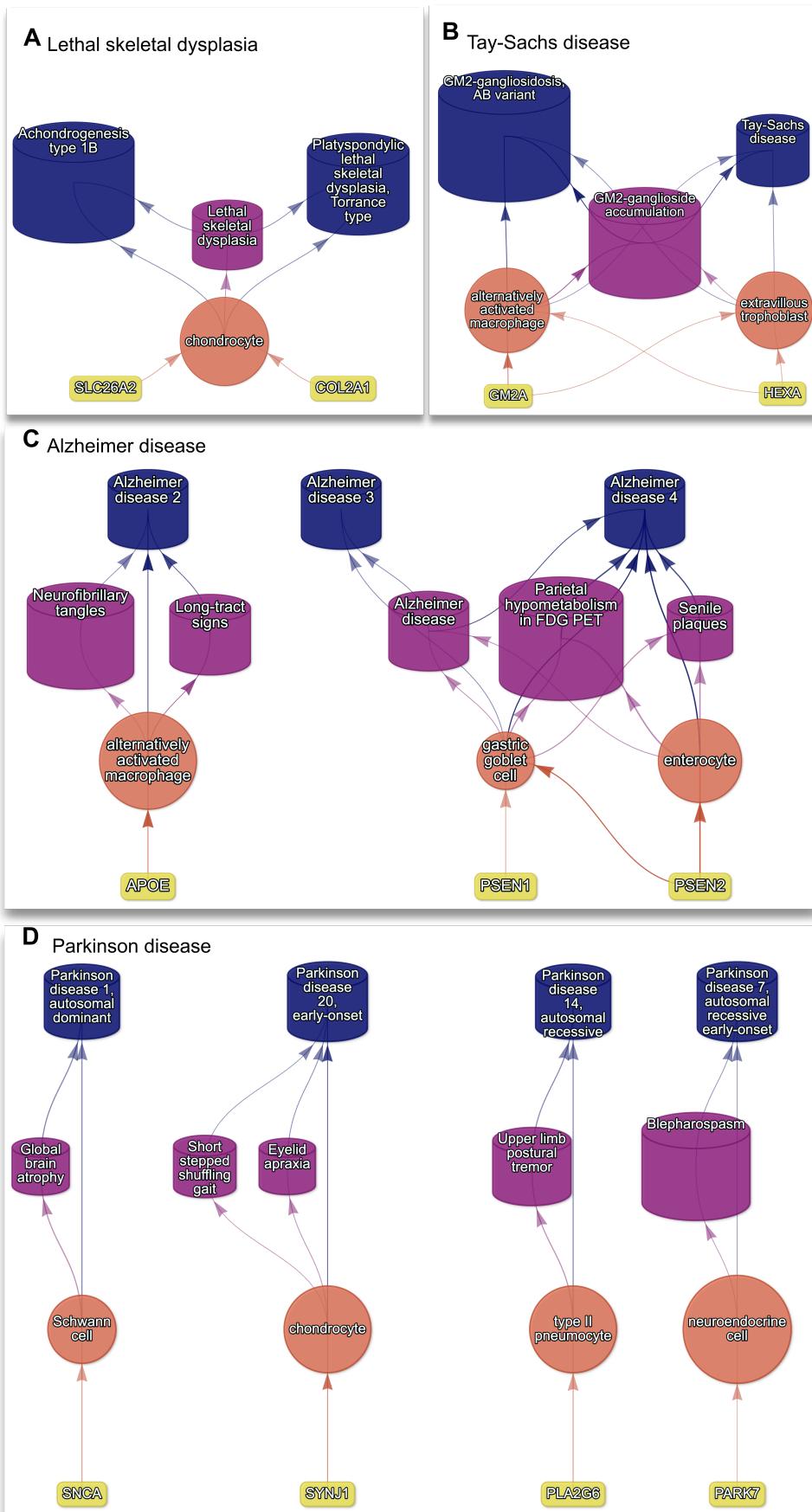


Figure 10: Example cell type-specific gene therapy targets for several severe phenotypes and their associated diseases. Each disease (blue cylinders) is connected to its phenotype (purple cylinders) based on well-established clinical observations recorded within the HPO¹¹. Phenotypes are connected to cell types (red circles) via association testing between weighted gene sets ($FDR_{pc} < 0.05$). Each cell type is connected to the prioritised gene targets (yellow boxes) based on the driver gene analysis. The thickness of the edges connecting the nodes represent the (mean) fold-change from the bootstrapped enrichment tests. Nodes were spatially arranged using the Sugiyama algorithm⁴⁹.

366 surgical intervention is insufficient, targeting these genes within chondrocytes may
 367 prove a viable long-term solution for children suffering from lethal skeletal dysplasia.

368 Tay-Sachs disease (TSD) is a devastating hereditary condition in which children are
 369 born appearing healthy, which gradually degrades leading to death after 3-5 years.
 370 The underlying cause is the toxic accumulation of gangliosides in the nervous sys-
 371 tem due to a loss of the enzyme produced by *HEXA*. While this could in theory be
 372 corrected with gene editing technologies, there remain some outstanding challenges.
 373 One of which is identifying which cell types should be targeted to ensure the most
 374 effective treatments. Here we identified alternatively activated macrophages as the
 375 cell type most strongly associated with ‘GM2-ganglioside accumulation’. The role
 376 of aberrant macrophage activity in the regulation of ganglioside levels is supported
 377 by observation that gangliosides accumulate within macrophages in TSD⁵³, as well
 378 as experimental evidence in rodent models^{54,55,56}. Our results not only corroborate
 379 these findings, but propose macrophages as the primary causal cell type in TSD,
 380 making it the most promising cell type to target in therapies.

381 Another challenge in TSD is early detection and diagnosis, before irreversible dam-
 382 age has occurred. Our pipeline implicated extravillous trophoblasts of the placenta
 383 in ‘GM2-ganglioside accumulation’. While not necessarily a target for gene ther-
 384 apy, checking these cells *in utero* for an absence of *HEXA* may serve as a viable
 385 biomarker as these cells normally express the gene at high levels. Early detection of
 386 TSD may lengthen the window of opportunity for therapeutic intervention⁵⁷, espe-
 387 cially when genetic sequencing is not available or variants of unknown significance
 388 are found within *HEXA*⁵⁸.

389 Alzheimer disease (AD) is the most common neurodegenerative condition. It is
 390 characterised by a set of variably penetrant phenotypes including memory loss, cog-
 391 nitive decline, and cerebral proteinopathy. Interestingly, we found that different
 392 forms of early onset AD (which are defined by the presence of a specific disease
 393 gene) are each associated with different cell types via different phenotypes. For ex-
 394 ample, AD 3 and AD 4 are primarily associated with cells of the digestive system
 395 ('enterocyte', 'gastric goblet cell') and are implied to be responsible for the pheno-
 396 types 'Senile plaques', 'Alzheimer disease', 'Parietal hypometabolism in FDG PET'.
 397 Meanwhile, AD 2 is primarily associated with immune cells ('alternatively activated
 398 macrophage') and is implied to be responsible for the phenotypes 'Neurofibrillary
 399 tangles', 'Long-tract signs'. This suggests that different forms of AD may be driven
 400 by different cell types and phenotypes, which may help to explain its variability in
 401 onset and clinical presentation.

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

416 It should be noted that the HPO only includes gene annotations for the monogenic
 417 forms of AD and PD. However it has previously been shown that there is at least
 418 partial overlap in their phenotypic and genetic aetiology with respect to their com-

mon forms. Thus understanding the monogenic forms of these diseases may shed light onto their more common counterparts.

421 3.14 Experimental model translatability

422 We computed interspecies translatability scores using a combination of both ontological (SIM_o) and genotypic (SIM_g) similarity relative to each homologous human
 423 phenotype and its associated genes ?@fig-animal-models. In total, we mapped
 424 278 non-human phenotypes (in *Caenorhabditis elegans*, *Danio rerio*, *Mus musculus*,
 425 *Rattus norvegicus*) to 849 homologous human phenotypes. Amongst the 5244 phe-
 426 notype within our prioritised therapy targets, 353 had viable animal models in at
 427 least on non-human species. Per species, the number of homologous phenotypes was:
 428 *Danio rerio* (n=213), *Mus musculus* (n=150), *Caenorhabditis elegans* (n=35), *Rat-*
 429 *tus norvegicus* (n=3). Amongst our prioritised targets with a GPT-4 severity score
 430 of >10, the phenotypes with the greatest animal model similarity were ‘Anterior
 431 vertebral fusion’ ($SIM_{o,g} = 0.967$), ‘Disc-like vertebral bodies’ ($SIM_{o,g} = 0.964$),
 432 ‘Metaphyseal enchondromatosis’ ($SIM_{o,g} = 0.946$), ‘Peripheral retinal avasculariza-
 433 tion’ ($SIM_{o,g} = 0.943$), ‘Retinal vascular malformation’ ($SIM_{o,g} = 0.943$).
 434

435 4 Discussion

436 Across the 201 cell types and 11,047 RD-associated phenotypes investigated, more
 437 than 46,437 significant phenotype-cell type relationships were discovered. The ex-
 438 amples we have highlighted above recapitulate well-known relationships, provide
 439 additional cellular context to many of these known relationships, and discover novel
 440 relationships at multiple biological scales.

441 Investigating RDs at the level of phenotypes offers several key advantages. First,
 442 the vast majority of RDs only have one associated gene (7671/8631 diseases = 89%).
 443 Aggregating gene sets across diseases into phenotype-centric “buckets” permits
 444 sufficiently well-powered analyses, with an average of ~76 genes per phenotype (me-
 445 dian=7) see Fig. 12. Second, we hypothesise that these phenotype-level gene sets
 446 converge on a limited number of molecular and cellular pathways. Perturbations
 447 to these pathways manifest as one or more phenotypes which, when considered to-
 448 gether, tend to be clinically diagnosed as a certain disease. Third, RDs are often
 449 highly heterogeneous in their clinical presentation across individuals, leading to the
 450 creation of an ever increasing number of disease subtypes (some of which only have
 451 a single documented case). In contrast, a phenotype-centric approach enables us
 452 to more accurately describe a particular individual’s version of a disease without
 453 relying on the generation of additional disease subcategories. By characterising an
 454 individual’s precise phenotypes over time, we may better understand the underly-
 455 ing biological mechanisms that have caused their condition. However, in order to
 456 achieve a truly precision-based approach to clinical care, we must first characterise
 457 the molecular and cellular mechanisms that cause the emergence of each phenotype.
 458 Here, we provide a highly reproducible framework that enables this at the scale of
 459 the entire genome. This presents an opportunity to design basket trials of patients
 460 with different diseases but overlapping phenotypes and cellular mechanisms¹⁷. It
 461 may be especially helpful for complex patients with diagnostically ambiguous sets of
 462 phenotypes who would otherwise be excluded from traditional clinical trials⁵⁹.

463 It was paramount to the success of this study to ensure our results were anchored in
 464 ground-truth benchmarks, generated falsifiable hypotheses, and rigorously guarded
 465 against false-positive associations. Extensive validation using multiple approaches
 466 demonstrated that our methodology consistently recapitulates expected phenotype-
 467 cell type associations (Fig. 1-Fig. 5). This was made possible by the existence of
 468 comprehensive, structured ontologies for all phenotypes (HPO) and cell types (CL),
 469 which provide an abundance of clear and falsifiable hypotheses for which to test
 470 our predictions against. Several key examples include 1) strong enrichment of as-
 471 sociations 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. 13). 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. 10).

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 diagnosis⁶⁰. 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. 10).

Unfortunately, there are currently only treatments available for less than 5% of RDs⁶. Novel technologies including CRISPR, prime editing, antisense oligonucleotides, viral vectors, and/or lipid nanoparticles, have been undergone significant advances in the last several years⁶¹⁻⁶⁵ and proven remarkable clinical success in an increasing number of clinical applications⁶⁶⁻⁶⁹. 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 technologies⁷⁰ 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)⁷¹. While these technologies

525 have the potential to revolutionise RD medicine, their successful application is de-
526 pendent on first understanding the mechanisms causing each disease.

527 To address this critical gap in knowledge, we used our results to create a repro-
528ducible and customisable pipeline to nominate cell type-resolved therapeutic targets
529 (Fig. 8-Fig. 10). Targeting cell type-specific mechanisms underlying granular RD
530 phenotypes can improve therapeutic effectiveness by treating the causal root of an
531 individual's conditions^{62,72}. A cell type-specific approach also helps to reduce the
532 number of harmful side effects caused by unintentionally delivering the therapeutic
533 to off-target tissues/cell types (which may induce aberrant gene activity), espe-
534 cially when combined with technologies that can target cell surface antigens (e.g.
535 viral vectors)⁷³. This has the additional benefit of reducing the minimal effective
536 dose of a therapeutic, which can be both immunogenic and extremely financially
537 costly^{9,10,61,64}. Here, we demonstrate the utility of a high-throughput evidence-based
538 approach to RD therapeutics discovery by highlighting several of the most promis-
539 ing therapeutic candidates. Our pipeline takes into account a myriad of factors,
540 including the strength of the phenotype-cell type associations, symptom-cell type
541 associations, cell type-specificity of causal genes, the severity and frequency of the
542 phenotypes, suitability for gene therapy delivery systems (e.g. recombinant adeno-
543 associated viral vectors (rAAV)), as well as a quantitative analysis of phenotypic
544 and genetic animal model translatability (**?@fig-animal-models**). We validated
545 these candidates by comparing the proportional overlap with gene therapies that are
546 presently in the market or undergoing clinical trials, in which we recovered 79% of
547 all active gene therapies and 0% of failed gene therapies (Fig. 9, Fig. 15). Despite
548 nominating a large number of putative targets, hypergeometric tests confirmed that
549 our targets were strongly enriched for targets of existing therapies that are either
550 approved or currently undergoing clinical trials.

551 It should be noted that our study has several key limitations. First, while our cell
552 type datasets are amongst the most comprehensive human scRNA-seq references cur-
553 rently available, they are nevertheless missing certain tissues, cell types (e.g. sperma-
554 tocytes, oocytes), and life stages (post-natal childhood, senility). It is also possible
555 that we have not captured certain cell state signatures that only occur in disease
556 (e.g. disease-associated microglia^{74,75}). Though we reasoned that using only control
557 cell type signatures would mitigate bias towards any particular disease, and avoid
558 degradation of gene signatures due to loss of function mutations. Second, the collec-
559 tive knowledge of gene-phenotype and gene-disease associations is far from complete
560 and we fully anticipate that these annotations will continue to expand and change
561 well into the future. It is for this reason we designed this study to be easily repro-
562 duced within a single containerised script so that we (or others) may rerun it with
563 updated datasets at any point. Finally, causality is notoriously difficult to prove
564 definitively from associative testing alone, and our study is not exempt from this
565 rule. Despite this, there are several reasons to believe that our approach is able to
566 better approximate causal relationships than traditional approaches. First, we did
567 not intentionally preselect any subset of phenotypes or cell types to investigate here.
568 Along with a scaling prestep during linear modelling, this means that all the results
569 are internally consistent and can be directly compared to one another (in stark con-
570 trast to literature meta-analyses). Furthermore, for the phenotype gene signatures
571 we used expert-curated GenCC annotations^{76,77} to weight the current strength of
572 evidence supporting a causal relationship between each gene and phenotype. This is
573 especially important for phenotypes with large genes lists (thousands of annotations)
574 for which some of the relationships may be tenuous. Within the cell type references,
575 we deliberately chose to use specificity scores (rather than raw gene expression) as
576 this normalisation procedure has previously been demonstrated to better distinguish
577 between signatures of highly similar cell types/subtypes⁷⁸.

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 Cell-typing 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.

590 5 Conclusions

591 Ultimately, our primary objective was to develop a methodology capable of generating high-throughput phenome-wide predictions while preserving the accuracy and 592 clinical utility typically associated with more narrowly focused studies. With the 593 rapid advancement of gene therapy technologies, and a regulatory landscape that is 594 evolving to better meet the needs of a large and diverse patient population, there is 595 finally momentum to begin to realise the promise of personalised medicine. This has 596 especially important implications for the global RD community which has remained 597 relatively neglected. Here, we lay out the groundwork necessary for this watershed 598 moment by providing a scalable, cost-effective, and fully reproducible means of 599 resolving the multi-scale, cell-type specific mechanisms of virtually all rare diseases. 600

601 6 Methods

602 6.1 Human Phenotype Ontology

603 The latest version of the HPO (release releases) was downloaded from the EMBL- 604 EBI Ontology Lookup Service⁷⁹ and imported into R using the `HPOExplorer` pack- 605 age. This R object was used to extract ontological relationships between phenotypes 606 as well as to assign absolute and relative ontological levels to each phenotype. The 607 latest version of the HPO phenotype-to-gene mappings and phenotype annotations 608 were downloaded from the official HPO GitHub repository and imported into R 609 using `HPOExplorer`. This contains lists of genes associated with phenotypes via 610 particular diseases, formatted as three columns in a table (gene, phenotype, disease).

611 However, not all genes have equally strong evidence of causality with a disease or 612 phenotype, especially when considering that the variety of resources used to gener- 613 ate these annotations (OMIM, Orphanet, DECIPHER) use variable methodologies 614 (e.g. expert-curated review of the medical literature vs. automated text mining 615 of the literature). Therefore we imported data from the Gene Curation Coalition 616 (GenCC)^{76,77}, which (as of 2024-04-13) 21796 evidence scores across 7229 diseases 617 and 5142 genes. Evidence scores are defined by GenCC using a standardised ordi- 618 nal rubric which we then encoded as a semi-quantitative score ranging from 0 (no 619 evidence of disease-gene relationship) to 6 (strongest evidence of disease-gene rela- 620 tionship) (see Table 3). As each Disease-Gene pair can have multiple entries (from 621 different studies) with different levels of evidence, we then summed evidence scores 622 per Disease-Gene pair to generate aggregated Disease-by-Gene evidence scores. This 623 procedure can be described as follows.

624 Let us denote:

- 625 • D as diseases.
- 626 • P as phenotypes in the HPO.
- 627 • G as genes

- 628 • S as the evidence scores describing the strength of the relationship between
 629 each Disease-Gene pair.
 630 • M_{ij} as the aggregated Disease-by-Gene evidence score matrix.

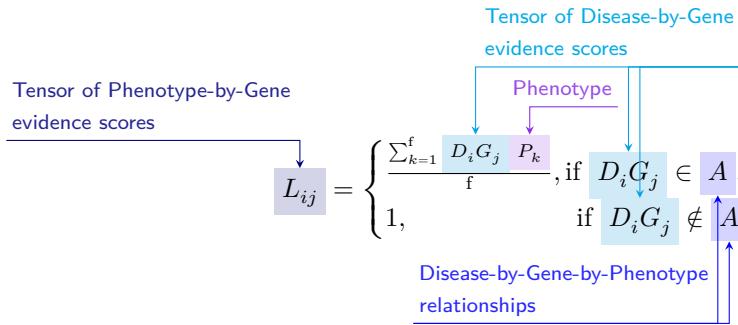
$$M_{ij} = \sum_{k=1}^f D_i G_j S_k$$

631 Next, we extracted Disease-Gene-Phenotype relationships from the annotations file
 632 distributed by the HPO (*phenotype_to_genes.txt*). This provides a list of genes asso-
 633 ciated with phenotypes via particular diseases, but does not include any strength of
 634 evidence scores.

635 Here we define: - A_{ijk} as the Disease-Gene-Phenotype relationships. - D_i as the i th
 636 disease. - G_j as the j th gene. - P_k as the k th phenotype.

$$A_{ijk} = D_i G_j P_k$$

637 In order to assign evidence scores to each Phenotype-Gene relationship, we com-
 638 bined the aforementioned datasets from GenCC (M_{ij}) and HPO (A_{ijk}) by merging
 639 on the gene and disease ID columns. For each phenotype, we then computed the
 640 mean of Disease-Gene scores across all diseases for which that phenotype is a symp-
 641 tom. This resulted in a final 2D tensor of Phenotype-by-Gene evidence scores (L_{ij}):



642 Construction of the tensor of Phenotype-by-Gene evidence scores.
 643
 644

645 Histograms of evidence score distributions at each step in processing can be found in
 646 Fig. 11.
 647

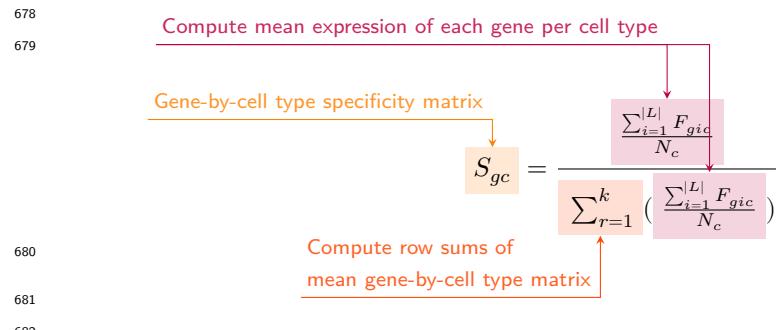
6.2 Single-cell transcriptomic atlases

648 In this study, the gene by cell type specificity matrix was constructed using the
 649 Descartes Human transcriptome atlas of foetal gene expression, which contains a
 650 mixture of single-nucleus and single-cell RNA-seq data (collected with sci-RNA-
 651 seq3)³¹. This dataset contains 377,456 cells representing 77 distinct cell types across
 652 15 tissues. All 121 human foetal samples ranged from 72 to 129 days in estimated
 653 postconceptual age. To independently replicate our findings, we also used the Hu-
 654 man Cell Landscape which contains single-cell transcriptomic data (collected with
 655 microwell-seq) from embryonic, foetal, and adult human samples across 49 tissues³².

656 Specificity matrices were generated separately for each transcriptomic atlas using
 657 the R package EWCE (v1.11.3)⁷⁸. Within each atlas, cell types were defined using
 658 the authors' original freeform annotations in order to preserve the granularity of
 659 cell subtypes as well as incorporate expert-identified rare cell types. Cell types were
 660

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. 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:



6.3 Phenotype-cell type associations

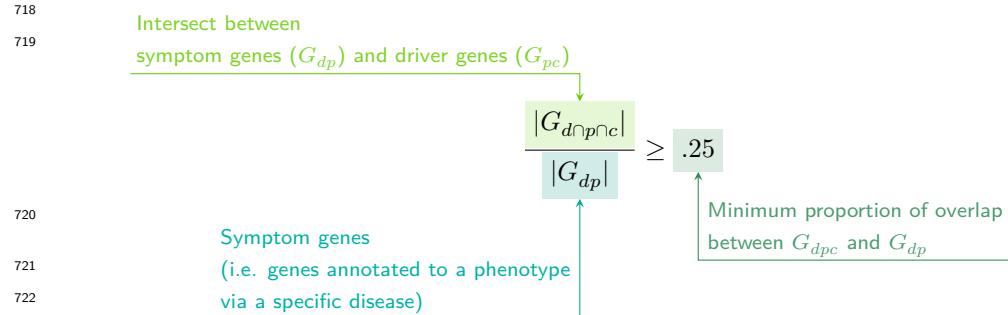
To test for relationships between each pairwise combination of phenotype (n=11,047) and cell type (n=201) we ran a series of univariate generalised linear models implemented via the `stats:::glm` function in R. First, we filtered the gene-by-phenotype evidence score matrix (L_{ij}) and the gene-by-cell type expression specificity matrix (S_{gc}) 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 (2,204,381 association tests total), we applied Benjamini-Hochberg false discovery rate⁸⁰ (denoted as FDR_{pc}) 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_{pc} was stringently controlled for across all tests performed in this study.

6.4 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_{dp} (see Fig. 12). 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

713 associated cell type. We then computed the intersect between symptom genes (G_{dp})
 714 and driver genes (G_{pc}), resulting in the gene subset $G_{d\cap p\cap c}$. Only $G_{d\cap p\cap c}$ gene sets
 715 with 25% or greater overlap with the symptom gene subset (G_{dp}) were kept. This
 716 procedure was repeated for all phenotype-cell type-disease triads, which can be
 717 summarised as follows:



723 6.5 Validation of expected phenotype-cell type relationships

724 We first sought to confirm that our tests (across both single-cell references) were
 725 able to recover expected phenotype-cell type relationships across seven high-level
 726 branches within the HPO (Fig. 1), including abnormalities of the cardiovascular
 727 system, endocrine system, eye, immune system, musculoskeletal system, nervous
 728 system, and respiratory system. Within each branch the number of significant tests
 729 in a given cell type were plotted (Fig. 1b). Mappings between freeform annota-
 730 tions (the level at which we performed our phenotype-cell type association tests)
 731 provided by the original atlas authors and their closest CL term equivalents were
 732 provided by CellxGene²⁹. CL terms along the *x-axis* of Fig. 1b were assigned colours
 733 corresponding to which HPO branch showed the greatest number of enrichments
 734 (after normalising within each branch to account for differences in scale). The nor-
 735 malised colouring allows readers to quickly assess which HPO branch was most often
 736 associated with each cell type, while accounting for differences in the number of
 737 phenotypes across branches. We then ran a series of Analysis of Variance (ANOVA)
 738 tests to determine whether (within a given branch) a given cell type was more of-
 739 ten enriched ($FDR_{pc} < 0.05$) within that branch relative to all of the other HPO
 740 branches of an equivalent level in the ontology (including all branches not shown in
 741 Fig. 1b). After applying Benjamini-Hochberg multiple testing correction⁸⁰ (denoted
 742 as $FDR_{b,c}$), we annotated each respective branch-by-cell type bar according to the
 743 significance (**** : $FDR_{b,c} < 1e - 04$, *** : $FDR_{b,c} < 0.001$, ** : $FDR_{b,c} < 0.01$, *
 744 : $FDR_{b,c} < 0.05$). Cell types in Fig. 1a-b were ordered along the *x-axis* according to
 745 a dendrogram derived from the CL ontology (Fig. 1c), which provides ground-truth
 746 semantic relationships between all cell types (e.g. different neuronal subtypes are
 747 grouped together).

748 As an additional measure of the accuracy of our phenotype-cell types test results
 749 we identified conceptually matched branches across the HPO and the CL (Fig. 1d
 750 and Table 4). For example, ‘Abnormality of the cardiovascular system’ in the HPO
 751 was matched with ‘cardiocytes’ in the CL which includes all cell types specific to the
 752 heart. Analogously, ‘Abnormality of the nervous system’ in the HPO was matched
 753 with ‘neural cell’ in the CL which includes all descendant subtypes of neurons and
 754 glia. This cross-ontology matching was repeated for each HPO branch and can be
 755 referred to as on-target cell types. Within each branch, the $-\log_{10}(FDR_{pc})$ values of
 756 on-target cell types were binned by rounding to the nearest integer (*x-axis*) and the
 757 percentage of tests for on-target cell types relative to all cell types were computed at
 758 each bin (*y-axis*) (Fig. 1d). The baseline level (dotted horizontal line) illustrates the
 759 percentage of on-target cell types relative to the total number of observed cell types.
 760 Any percentages above this baseline level represent greater than chance representa-
 761 tion of the on-target cell types in the significant tests.

6.6 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 biology⁸¹. 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 ability to generate associations was dependent on the existence of gene annotations within the HPO. 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 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. We also adjusted ontological distance by computing the ratio between the observed ontological distance and the smallest possible ontological distance for that cell type given the cell type that were available in our references ($dist_{adjusted} = \left(\frac{dist_{observed}+1}{dist_{minimum}+1} \right) - 1$). This provides a way of accurately measuring how dissimilar our identified cell types were for each phenotype-cell type association (Fig. 14).

6.7 Annotation of phenotypes using generative large language models

Only a small fraction of 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 severity⁸². 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 16880/18536 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.

$$NSS_p = \frac{\sum_{j=1}^m (F_{pj} \times W_j)}{\sum_{j=1}^m (\max\{F_j\} \times W_j)} \times 100$$

Sum of weighted annotation values across all metrics

Numerically encoded annotation value of metric j for phenotype p

Weight for metric j

Theoretical maximum severity score

810
811**6.8 Congenital phenotypes are associated with foetal cell types**

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 authors³²) 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.

6.9 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).

6.10 Prognosis via cell type-mediated differential outcomes

The phenotype hypotonia is associated with diseases that range in severity from benign to debilitating to fatal⁸³. 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 (Table 5). 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.

6.11 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. Each step is described in detail in Table 2.

860 **6.12 Therapeutic target validation**

861 To assess whether our prioritised therapeutic targets were likely to be viable, we
 862 computed the overlap between our gene targets and those of existing gene therapies
 863 at various stages of clinical development (Fig. 9). Gene targets were obtained for
 864 each therapy from the Therapeutic Target Database (TTD; release 2024-04-15) and
 865 mapped onto standardised HUGO Gene Nomenclature Committee (HGNC) gene
 866 symbols using the `orthogene` R package. We stratified our overlap metrics accord-
 867 ing to whether the therapies had failed (unsuccessful clinical trials or withdrawn),
 868 or were non-failed (successful or ongoing clinical trials). We then conducted hyper-
 869 geometric tests to determine whether the observed overlap between our prioritised
 870 targets and the non-failed therapy targets was significantly greater than expected
 871 by chance (i.e. enrichment). We also conducted a second hypergeometric test to
 872 determine whether the observed overlap between our prioritised targets and the
 873 failed therapy targets was significantly less than expected by chance (i.e. depletion).
 874 Finally, we repeated the analysis against all therapeutic targets, not just those of
 875 gene therapies, to determine whether our prioritised targets had relevance to other
 876 therapeutic modalities.

877 **6.13 Experimental model translatability**

878 To improve the likelihood of successful translation between preclinical animal models
 879 and human patients, we created an interspecies translatability prediction tool for
 880 each phenotype nominated by our gene therapy prioritised pipeline ([?@fig-animal-
 881 models](#)). First, we extracted ontological similarity scores of homologous phenotypes
 882 across species from the MKG⁸¹. Briefly, the ontological similarity scores (SIM_o) are
 883 computed for each homologous pair of phenotypes across two ontologies by calculat-
 884 ing the overlap in homologous phenotypes that are ancestors or descendants of the
 885 target phenotype. Next, we generated genotypic similarity scores (SIM_g) for each
 886 homologous phenotype pair by computing the proportion of 1:1 orthologous genes
 887 using gene annotation from their respective ontologies. Interspecies orthologs were
 888 also obtained from the MKG. Finally, both scores are multiplied together to yield a
 889 unified ontological-genotypic similarity score ($SIM_{o,g}$).

890 **6.14 Novel R packages**

891 To facilitate all analyses described in this study and to make them more easily repro-
 892ducible by others, we created several open-source R packages. `KGExplorer` imports
 893 and analyses large-scale biomedical knowledge graphs and ontologies. `HPOExplorer`
 894 aids in managing and querying the directed acyclic ontology graph within the HPO.
 895 `MSTExplorer` facilitates the efficient analysis of many thousands of phenotype-cell
 896 type association tests, and provides a suite of multi-scale therapeutic target prioriti-
 897 sation and visualisation functions. These R packages also include various functions
 898 for distributing the post-processed results from this study in an organised, tabular
 899 format. Of note, `MSTExplorer::load_example_results` loads all summary statis-
 900 tics from our phenotype-cell type tests performed here.

901 **6.15 Rare Disease Celltyping Portal**

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

910 **7 Tables**

Table 1: Summary statistics of enrichment results stratified by single-cell atlas. Summary statistics at multiple levels (tests, cell types, phenotypes, diseases, cell types per phenotype, phenotypes per cell type) stratified by the single-cell atlas that was used as a cell type signature reference (Descartes Human or Human Cell Atlas).

	DescartesHuman	HumanCellLandscape	all
tests significant	19,894	26,543	46,437
tests	847,077	1,357,304	2,204,381
tests significant (%)	2.35	1.96	2.11
cell types significant	77	124	201
cell types	77	124	201
cell types significant (%)	100	100	100
phenotypes significant	7,330	9,038	9,563
phenotypes tested	11,001	10,946	11,015
phenotypes	11,047	11,047	11,047
phenotypes significant (%)	66.4	81.8	86.6
diseases significant	8,628	8,627	8,628
diseases	8,631	8,631	8,631
diseases significant (%)	100	100	100
cell types per phenotype (mean)	1.81	2.42	4.22
cell types per phenotype (median)	1	2	3
cell types per phenotype (min)	0	0	0
cell types per phenotype (max)	31	28	59
phenotypes per cell type (mean)	258	214	231
phenotypes per cell type (median)	252	200	209
phenotypes per cell type (min)	71	57	57
phenotypes per cell type (max)	694	735	735

Table 2: Description of each filtering step performed in the multi-scale therapeutic target prioritisation pipeline. ‘Level’ indicates the biological scale at which the step is applied to.

level	step	description
NA	1. start	NA
Cell type	2. q threshold	Keep only cell type-phenotype association results at $q \leq 0.05$.
Phenotype	3. keep descendants	Remove phenotypes belonging to a certain branch of the HPO, as defined by an ancestor term.
Phenotype	4. info content threshold	Keep only phenotypes with a minimum information criterion score (computed from the HPO).
Phenotype	5. severity threshold	Keep only phenotypes with mean Severity equal to or below the threshold.
Symptom	6. 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).
Gene	7. symptom gene overlap	Ensure that genes nominated at the phenotype-level also appear in the genes overlapping at the cell type-specific symptom-level.
Gene	8. evidence score threshold	Remove genes that are below an aggregate phenotype-gene evidence score threshold.
Gene	9. add driver genes	Keep only genes that are driving the association with a given phenotype (inferred by the intersection of phenotype-associated genes and gene with high specificity quantiles in the target cell type).
Symptom	10. symptom intersection threshold	Minimum proportion of genes overlapping between a symptom gene list (phenotype-associated genes in the context of a particular disease) and the phenotype-cell type association driver genes.
Gene	11. 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).
Phenotype	12. prune ancestors	Remove redundant ancestral phenotypes when at least one of their descendants already exist.
All	13. top n	Sort candidate targets by a preferred order of metrics and only return the top N targets per cell type-phenotype combination.
NA	14. end	NA

911 8 Data and Code Availability

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

- 914 • [Human Phenotype Ontology](#)
- 915 • [GenCC](#)
- 916 • [Descartes Human scRNA-seq atlas](#)
- 917 • [Human Cell Landscape scRNA-seq atlas](#)
- 918 • [Rare Disease Celltyping Portal](#)
- 919 • [KGExplorer](#)
- 920 • [HPOExplorer](#)
- 921 • [MSTExplorer](#)
- 922 • [Code to replicate analyses](#)
- 923 • [Cell type-specific gene target prioritisation](#)
- 924 • [Complement system gene list](#)

925 9 Acknowledgements

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

930 9.1 Funding

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

935 References

- 936 1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A* **179**, 885–892
 (2019).
- 937 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).
- 938 3. Rare diseases BioResource.
- 939 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).
- 940 5. Molster, C. *et al.* Survey of healthcare experiences of australian adults living with
 rare diseases. *Orphanet J. Rare Dis.* **11**, 30 (2016).
- 941 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).
- 942 7. Institute of Medicine (US) Committee on Accelerating Rare Diseases Research
 and Orphan Product Development, Field, M. J. & Boat, T. F. *Coverage and Re-
 imbursement: Incentives and Disincentives for Product Development*. (National
 Academies Press (US), 2010).
- 943 8. Yates, N. & Hinkel, J. The economics of moonshots: Value in rare disease drug
 development. *Clin. Transl. Sci.* **15**, 809–812 (2022).
- 944 9. Nuijten, M. Pricing zolgensma - the world’s most expensive drug. *J Mark Access
 Health Policy* **10**, 2022353 (2022).
- 945 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).
- 946 11. Gargano, M. A. *et al.* The human phenotype ontology in 2024: Phenotypes around
 the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).

- 947 12. Köhler, S. *et al.* Expansion of the human phenotype ontology (HPO) knowledge
base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
- 948 13. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**,
D1207–D1217 (2021).
- 949 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).
- 950 15. Osmond, M. *et al.* Outcome of over 1500 matches through the matchmaker ex-
change for rare disease gene discovery: The 2-year experience of Care4Rare canada.
Genet. Med. **24**, 100–108 (2022).
- 951 16. Philippakis, A. A. *et al.* The matchmaker exchange: A platform for rare disease
gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).
- 952 17. Zanello, G. *et al.* Targeting shared molecular etiologies to accelerate drug develop-
ment for rare diseases. *EMBO Mol. Med.* **15**, e17159 (2023).
- 953 18. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare dis-
eases: Analysis of the orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173
(2020).
- 954 19. Rare diseases, common challenges. *Nat. Genet.* **54**, 215 (2022).
- 955 20. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: Leverag-
ing knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–
D1043 (2019).
- 956 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).
- 957 22. McKusick, V. A. Mendelian inheritance in man and its online version, OMIM. *Am.*
J. Hum. Genet. **80**, 588–604 (2007).
- 958 23. Maiella, S., Rath, A., Angin, C., Mousson, F. & Kremp, O. [Orphanet and its con-
sortium: Where to find expert-validated information on rare diseases]. *Rev. Neurol.*
169 Suppl 1, S3–8 (2013).
- 959 24. Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E. en & Cornel, M. C. [Or-
phanet: A european database for rare diseases]. *Ned. Tijdschr. Geneeskde.* **152**,
518–519 (2008).
- 960 25. Firth, H. V. *et al.* DECIPHER: Database of chromosomal imbalance and pheno-
type in humans using ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
- 961 26. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applica-
tions of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
- 962 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).
- 963 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).
- 964 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).
- 965 30. Svensson, V., Veiga Beltrame, E. da & Pachter, L. A curated database reveals
trends in single-cell transcriptomics. *Database* **2020**, (2020).
- 966 31. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
- 967 32. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature*
581, 303–309 (2020).
- 968 33. Diehl, A. D. *et al.* The cell ontology 2016: Enhanced content, modularization, and
ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
- 969 34. Heim, C. E. *et al.* Myeloid-derived suppressor cells contribute to staphylococcus
aureus orthopedic biofilm infection. *J. Immunol.* **192**, 3778–3792 (2014).
- 970 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).

- 971 36. Stoll, H. *et al.* Staphylococcal enterotoxins Dose-Dependently modulate the generation of Myeloid-Derived suppressor cells. *Front. Cell. Infect. Microbiol.* **8**, 321 (2018).
- 972 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).
- 973 38. Zhou, Z., Xu, M.-J. & Gao, B. Hepatocytes: A key cell type for innate immunity. *Cell. Mol. Immunol.* **13**, 301–315 (2016).
- 974 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).
- 975 40. Ladhami, S. N. *et al.* Invasive meningococcal disease in patients with complement deficiencies: A case series (2008-2017). *BMC Infect. Dis.* **19**, 522 (2019).
- 976 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).
- 977 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).
- 978 43. Lung, T. *et al.* The complement system in liver diseases: Evidence-based approach and therapeutic options. *J Transl Autoimmun* **2**, 100017 (2019).
- 979 44. Reis, E. S. *et al.* Applying complement therapeutics to rare diseases. *Clin. Immunol.* **161**, 225–240 (2015).
- 980 45. Seal, R. L. *et al.* Genenames.org: The HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).
- 981 46. Al-Hamoudi, W. K. Severe autoimmune hepatitis triggered by varicella zoster infection. *World J. Gastroenterol.* **15**, 1004–1006 (2009).
- 982 47. Brewer, E. C. & Hunter, L. Acute liver failure due to disseminated varicella zoster infection. *Case Reports Hepatol* **2018**, 1269340 (2018).
- 983 48. Eshchar, J., Reif, L., Waron, M. & Alkan, W. J. Hepatic lesion in chickenpox. A case report. *Gastroenterology* **64**, 462–466 (1973).
- 984 49. Sugiyama, K., Tagawa, S. & Toda, M. Methods for visual understanding of hierarchical system structures. *IEEE Trans. Syst. Man Cybern.* **11**, 109–125 (1981).
- 985 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).
- 986 51. Utsch, B., Albers, N. & Ludwig, M. Genetic and molecular aspects of hypospadias. *Eur. J. Pediatr. Surg.* **14**, 297–302 (2004).
- 987 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).
- 988 53. Fenderson, B. A. Chapter 6 - developmental and genetic diseases. in *Pathology secrets (third edition)* (ed. Damjanov, I.) 98–119 (Mosby, 2009). doi:[10.1016/B978-0-323-05594-9.00006-4](https://doi.org/10.1016/B978-0-323-05594-9.00006-4).
- 989 54. Vilcaes, A. A., Garbarino-Pico, E., Torres Demichelis, V. & Daniotti, J. L. [Ganglioside synthesis by plasma membrane-associated sialyltransferase in macrophages](#). *International Journal of Molecular Sciences* **21**, 1063 (2020).
- 990 55. Yohe, H. C., Coleman, D. L. & Ryan, J. L. [Ganglioside alterations in stimulated murine macrophages](#). *Biochimica et Biophysica Acta (BBA) - Biomembranes* **818**, 81–86 (1985).
- 991 56. Demir, S. A., Timur, Z. K., Ateş, N., Martínez, L. A. & Seyran tepe, V. [GM2 ganglioside accumulation causes neuroinflammation and behavioral alterations in a mouse model of early onset tay-sachs disease](#). *Journal of Neuroinflammation* **17**, 277 (2020).
- 992 57. Solovyeva, V. V. *et al.* [New approaches to tay-sachs disease therapy](#). *Frontiers in Physiology* **9**, (2018).

- 993 58. Hoffman, J. D. *et al.* Next-generation DNA sequencing of HEXA: A step in the
right direction for carrier screening. *Molecular Genetics & Genomic Medicine* **1**,
260–268 (2013).
- 994 59. Diaz-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).
- 995 60. Lord, J. & Baralle, D. Splicing in the diagnosis of rare disease: Advances and chal-
lenges. *Front. Genet.* **12**, 689892 (2021).
- 996 61. 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).
- 997 62. Bulaklak, K. & Gersbach, C. A. The once and future gene therapy. *Nat. Commun.*
11, 5820 (2020).
- 998 63. Godbout, K. & Tremblay, J. P. Prime editing for human gene therapy: Where are
we now? *Cells* **12**, (2023).
- 999 64. Kohn, D. B., Chen, Y. Y. & Spencer, M. J. Successes and challenges in clinical
gene therapy. *Gene Ther.* **30**, 738–746 (2023).
- 1000 65. Zhao, Z., Shang, P., Mohanraju, P. & Geijsen, N. Prime editing: Advances and
therapeutic applications. *Trends Biotechnol.* **41**, 1000–1012 (2023).
- 1001 66. Darroo, J. J. Luxturna: FDA documents reveal the value of a costly gene therapy.
Drug Discov. Today **24**, 949–954 (2019).
- 1002 67. Mendell, J. R. *et al.* Single-Dose Gene-Replacement therapy for spinal muscular
atrophy. *N. Engl. J. Med.* **377**, 1713–1722 (2017).
- 1003 68. 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).
- 1004 69. Russell, S. *et al.* Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2)
in patients with RPE65-mediated inherited retinal dystrophy: A randomised, con-
trolled, open-label, phase 3 trial. *Lancet* **390**, 849–860 (2017).
- 1005 70. Lu, C.-F. FDA takes first step toward international regulation of gene therapies to
treat rare diseases. (2024).
- 1006 71. Brown, D. G., Wobst, H. J., Kapoor, A., Kenna, L. A. & Southall, N. Clinical de-
velopment times for innovative drugs. *Nat. Rev. Drug Discov.* **21**, 793–794 (2022).
- 1007 72. 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).
- 1008 73. Zhou, Q. & Buchholz, C. J. Cell type specific gene delivery by lentiviral vectors:
New options in immunotherapy. *Oncimmunology* **2**, e22566 (2013).
- 1009 74. Keren-shaul, H. *et al.* A unique microglia type associated with restricting develop-
ment of alzheimer 's disease. *Cell* **169**, 1276–1290.e17 (2017).
- 1010 75. Deczkowska, A. *et al.* Disease-associated microglia: A universal immune sensor of
neurodegeneration. *Cell* **173**, 1073–1081 (2018).
- 1011 76. DiStefano, M. T. *et al.* The gene curation coalition: A global effort to harmonize
gene-disease evidence resources. *Genet. Med.* **24**, 1732–1742 (2022).
- 1012 77. DiStefano, M. *et al.* P451: The gene curation coalition works to resolve discrep-
ancies in gene-disease validity assertions. *Genetics in Medicine Open* **1**, 100498
(2023).
- 1013 78. 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).
- 1014 79. Côté, R. *et al.* The ontology lookup service: Bigger and better. *Nucleic Acids Res.*
38, W155–60 (2010).
- 1015 80. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and
powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).

- 1016 81. Putman, T. E. *et al.* The monarch initiative in 2024: An analytic platform in-
tegrating phenotypes, genes and diseases across species. *Nucleic Acids Res.* **52**,
D938–D949 (2024).
- 1017 82. Lazarin, G. A. *et al.* Systematic classification of disease severity for evaluation of
expanded carrier screening panels. *PLoS One* **9**, e114391 (2014).
- 1018 83. Ahmed, M. I., Iqbal, M. & Hussain, N. A structured approach to the assessment of
a floppy neonate. *J. Pediatr. Neurosci.* **11**, 2–6 (2016).
- 1019
- 1020

1021 **10 Supplementary Materials**
 1022 **10.1 Supplementary Figures**

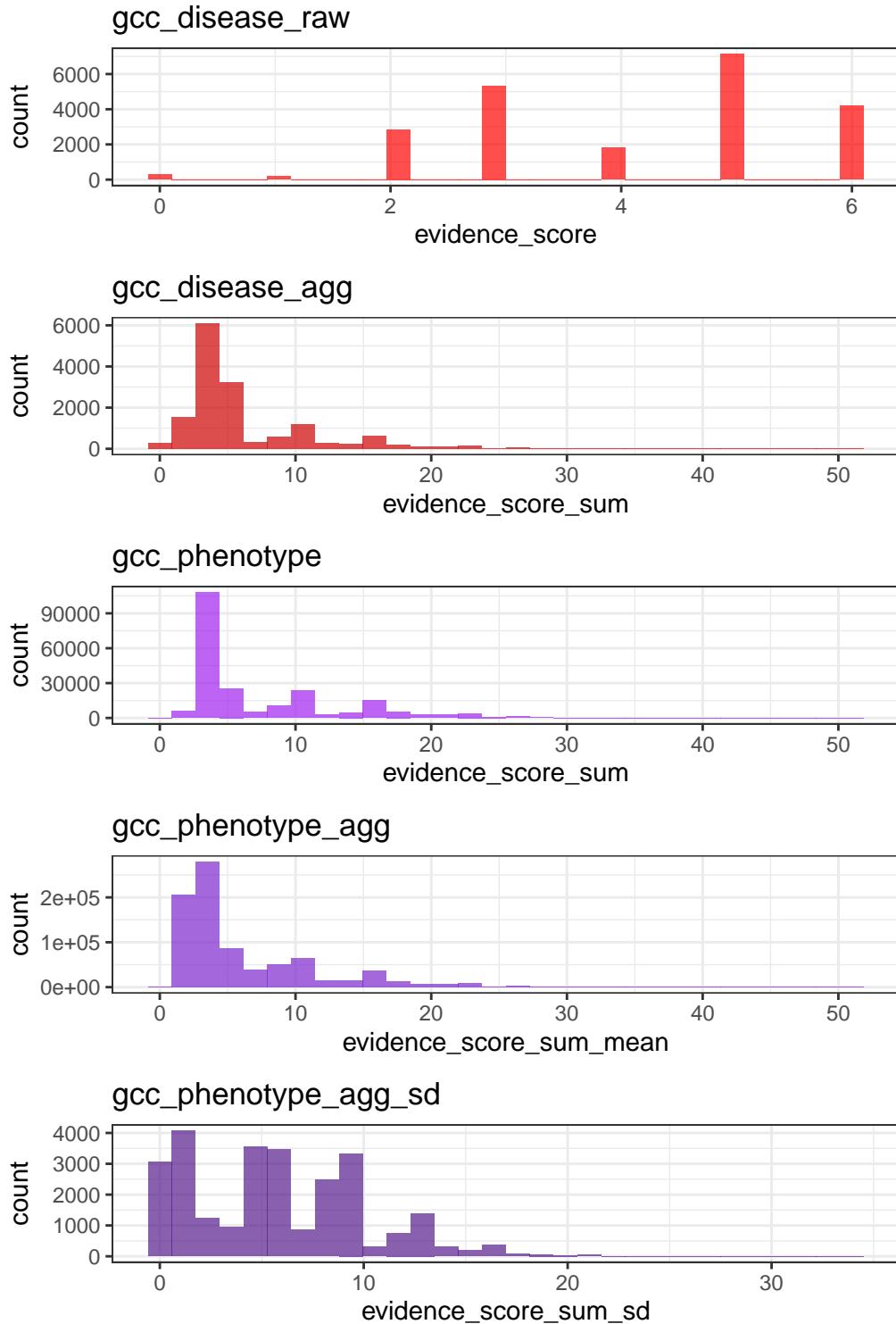


Figure 11: Distribution of evidence scores at each processing step.

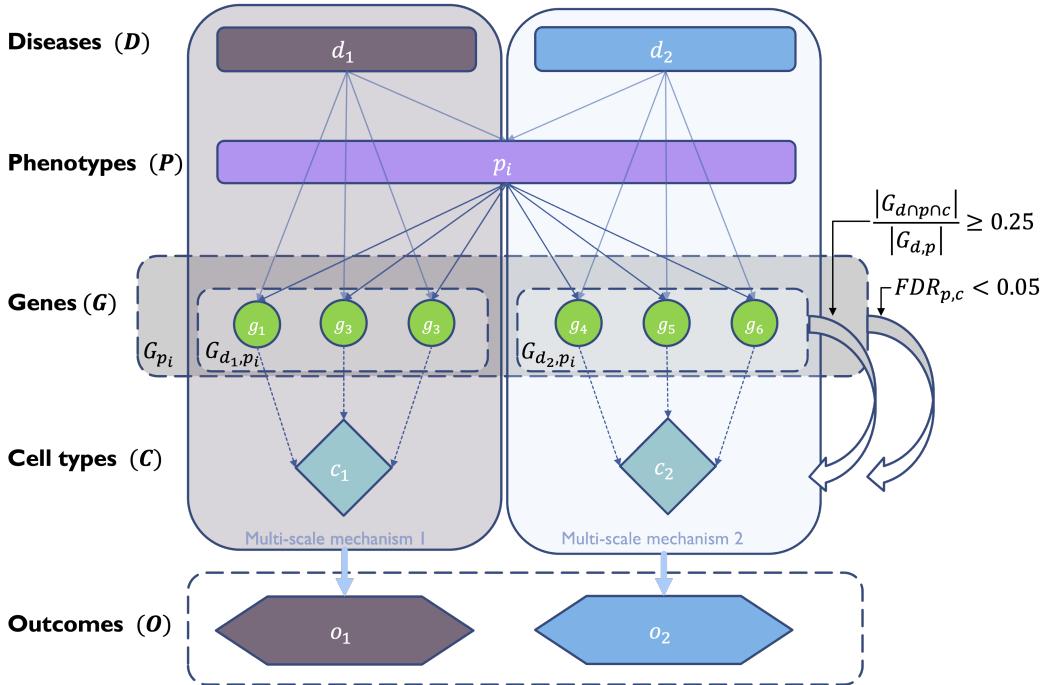


Figure 12: Diagrammatic overview of multi-scale disease investigation strategy. Here we provide an abstract example of differential disease aetiology across multiple scales: diseases (D), phenotypes (P), cell types (C), genes (G), and clinical outcomes (O). In the HPO, genes are assigned to phenotypes via particular diseases (G_{dp}). 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 ($FDR_{pc} < 0.05$). 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%.

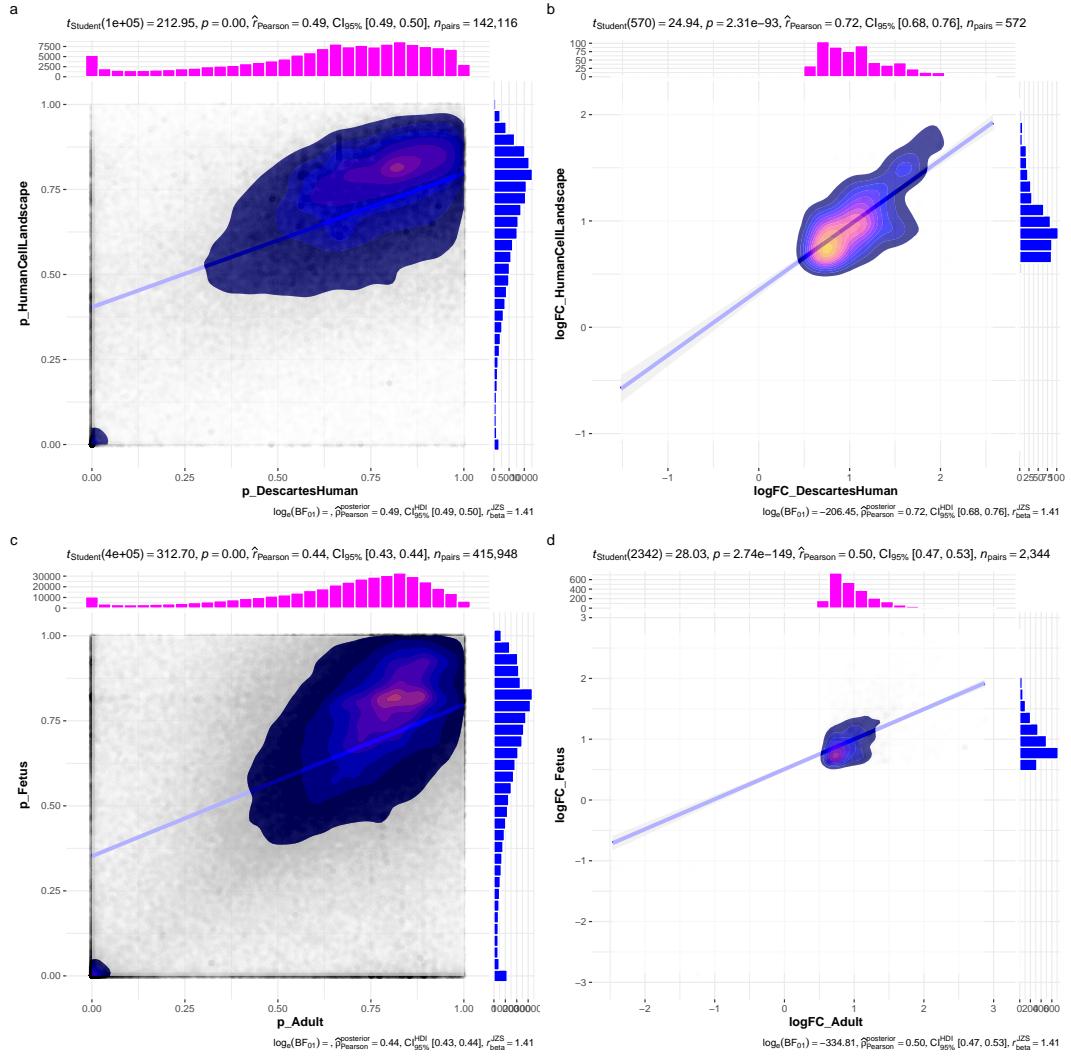


Figure 13: Inter- and intra-dataset validation across the different CellTypeDataset (CTD) and developmental stages. Correlations are computed using Pearson correlation coefficient. Point density is plotted using a 2D kernel density estimate. **a** Correlation between the uncorrected p-values from all 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 ($FDR_{pc} < 0.05$) 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 ($FDR_{pc} < 0.05$) using the Human Cell Landscape fetal samples vs. Human Cell Landscape adult samples.

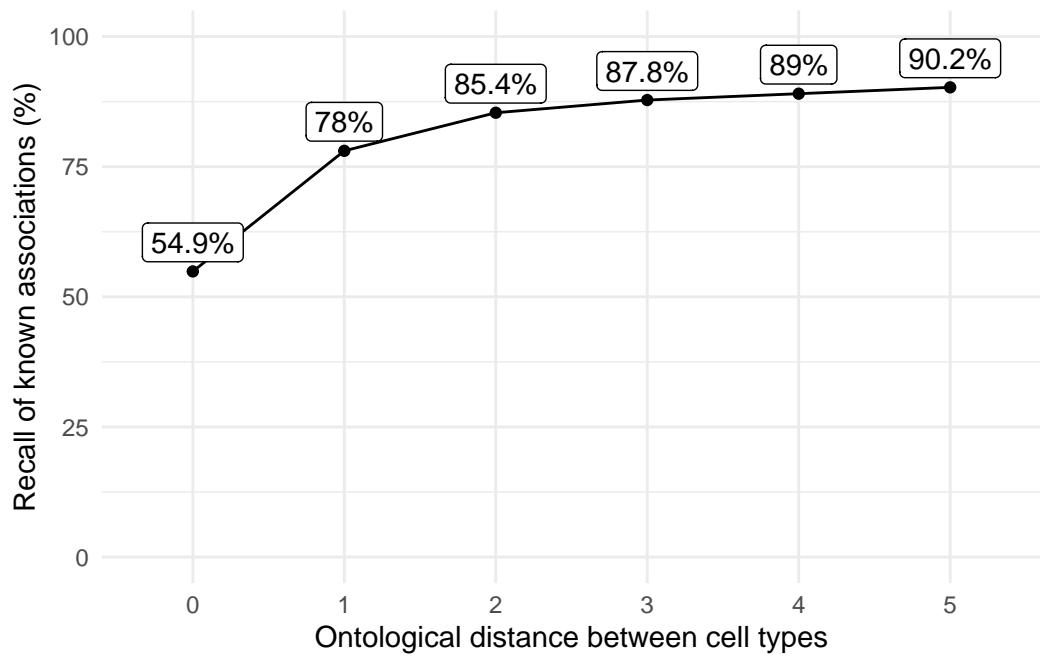


Figure 14: Recall of ground-truth Monarch Knowledge Graph phenotype-cell type relationships at each ontological distance between cell types according to the Cell Ontology.

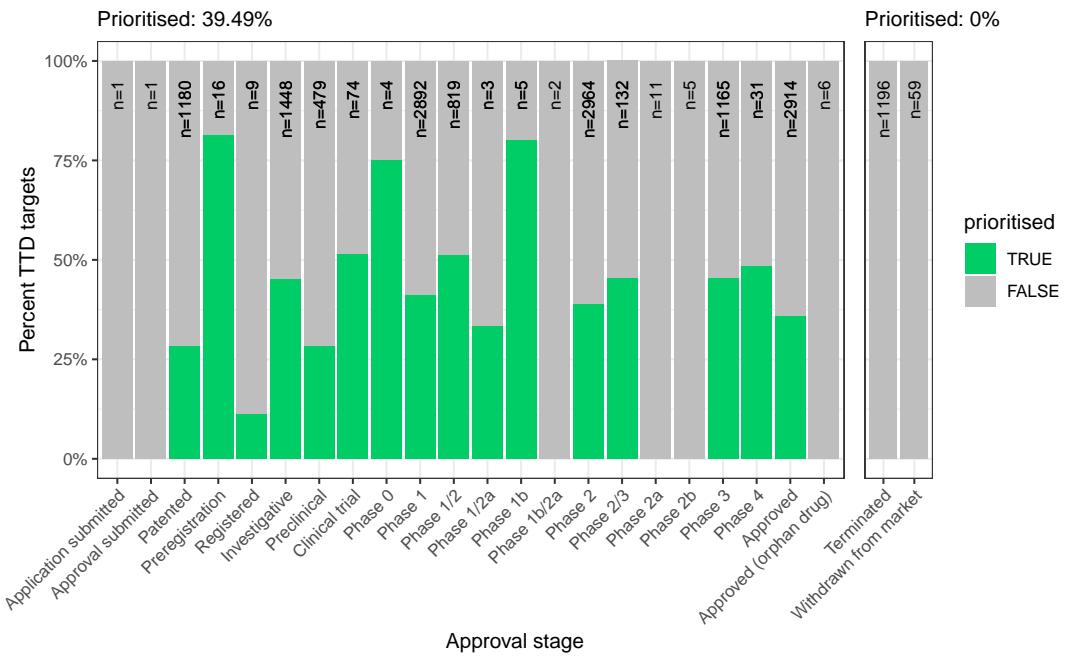
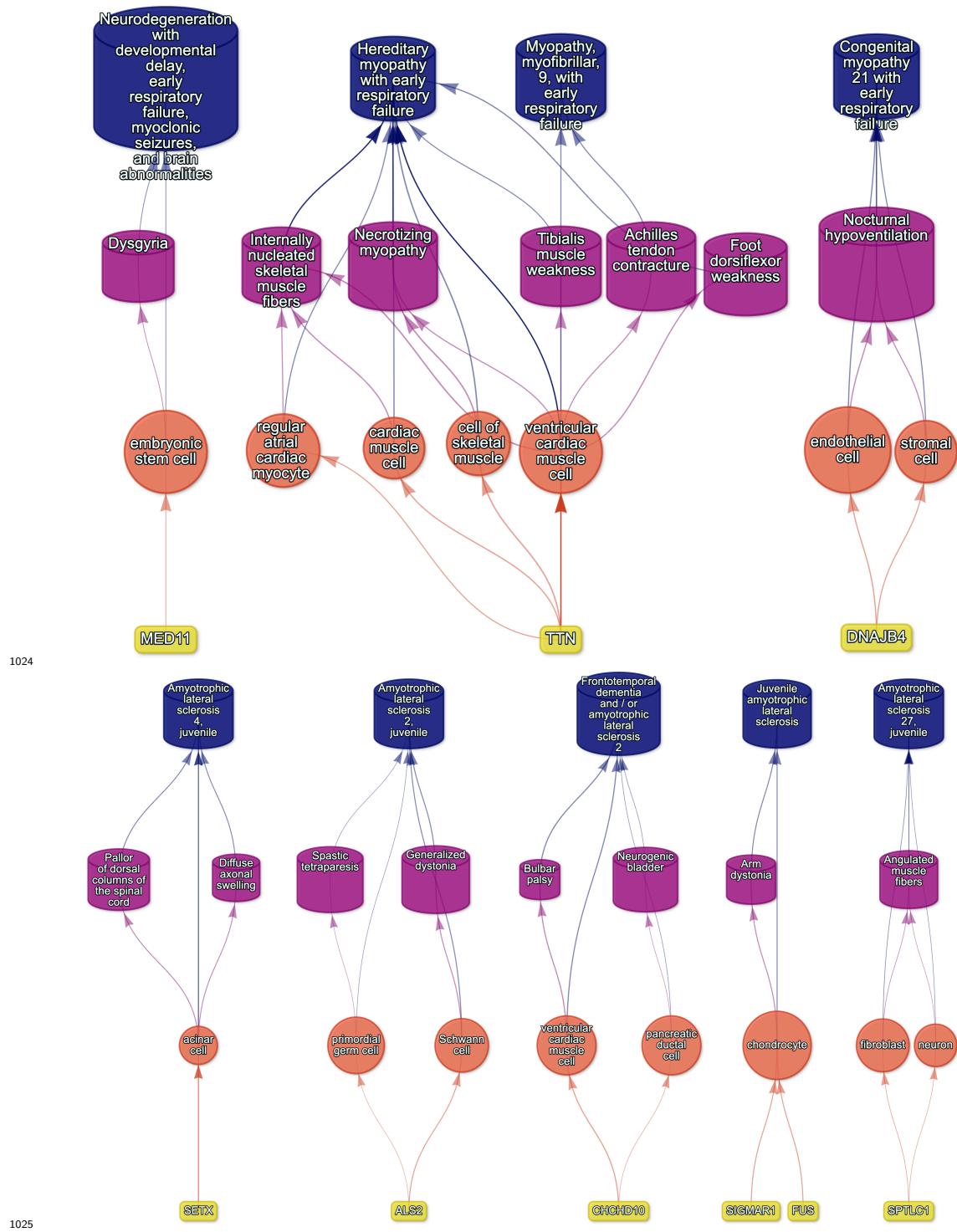
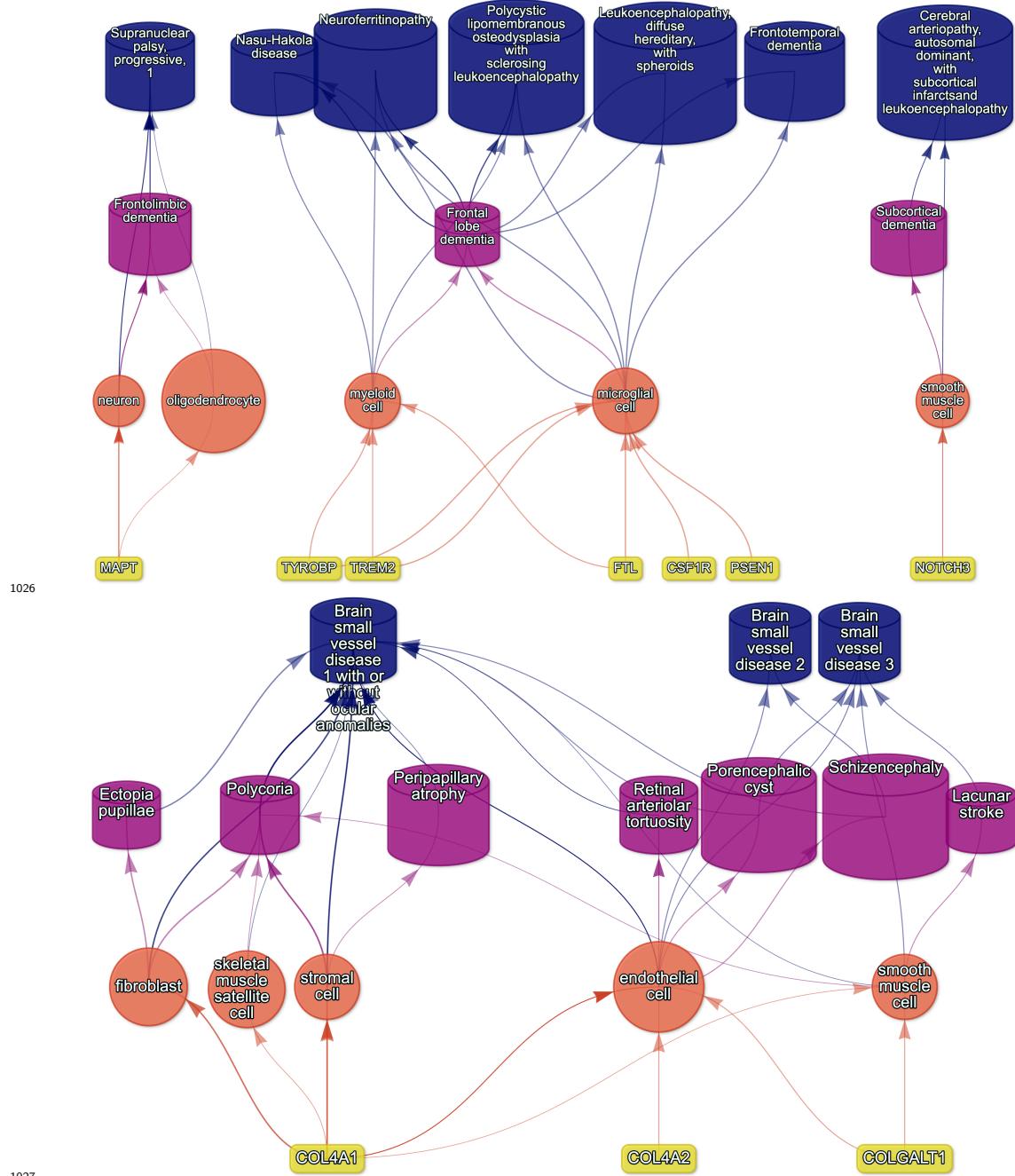


Figure 15: Therapeutics - Validation of prioritised therapeutic targets. Proportion of existing all therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline.

1023 (skipping until **simona** is fixed)





1028 Example cell type-specific gene therapy targets for several severe phenotypes and
 1029 their associated diseases. Each disease (blue cylinders) is connected to its phenotype
 1030 (purple cylinders) based on well-established clinical observations recorded within the
 1031 HPO¹¹. Phenotypes are connected to cell types (red circles) via association testing
 1032 between weighted gene sets ($FDR_{pc} < 0.05$). Each cell type is connected to the
 1033 prioritised gene targets (yellow boxes) based on the driver gene analysis. The thickness
 1034 of the edges connecting the nodes represent the (mean) fold-change from the
 1035 bootstrapped enrichment tests. Nodes were spatially arranged using the Sugiyama
 1036 algorithm⁴⁹.

10.2 Supplementary Tables

Table 3: Encodings for GenCC evidence scores. Assigned numeric values for the GenCC evidence levels.

classification_curie	classification_title	encoding
GENCC:100001	Definitive	6
GENCC:100002	Strong	5
GENCC:100003	Moderate	4
GENCC:100009	Supportive	3
GENCC:100004	Limited	2
GENCC:100005	Disputed Evidence	1
GENCC:100008	No Known Disease Relationship	0
GENCC:100006	Refuted Evidence	0

Table 4: On-target cell types for each HPO ancestral branch.

hpo_branch	cl_branch	cl_name	cl_id
Abnormality of the cardiovascular system	cardiocyte	cardiac muscle cell	CL:0000746
Abnormality of the cardiovascular system	cardiocyte	regular atrial cardiac myocyte	CL:0002129
Abnormality of the cardiovascular system	cardiocyte	endocardial cell	CL:0002350
Abnormality of the cardiovascular system	cardiocyte	epicardial adipocyte	CL:1000309
Abnormality of the cardiovascular system	cardiocyte	ventricular cardiac muscle cell	CL:2000046
Abnormality of the endocrine system	endocrine cell	endocrine cell	CL:0000163
Abnormality of the endocrine system	endocrine cell	neuroendocrine cell	CL:0000165
Abnormality of the endocrine system	endocrine cell	chromaffin cell	CL:0000166
Abnormality of the eye	photoreceptor cell / retinal cell	photoreceptor cell	CL:0000210
Abnormality of the eye	photoreceptor cell / retinal cell	amacrine cell	CL:0000561
Abnormality of the eye	photoreceptor cell / retinal cell	Mueller cell	CL:0000636
Abnormality of the eye	photoreceptor cell / retinal cell	retinal pigment epithelial cell	CL:0002586
Abnormality of the immune system	leukocyte	T cell	CL:0000084
Abnormality of the immune system	leukocyte	mature neutrophil	CL:0000096
Abnormality of the immune system	leukocyte	mast cell	CL:0000097
Abnormality of the immune system	leukocyte	microglial cell	CL:0000129
Abnormality of the immune system	leukocyte	professional antigen presenting cell	CL:0000145
Abnormality of the immune system	leukocyte	macrophage	CL:0000235
Abnormality of the immune system	leukocyte	B cell	CL:0000236
Abnormality of the immune system	leukocyte	dendritic cell	CL:0000451
Abnormality of the immune system	leukocyte	monocyte	CL:0000576
Abnormality of the immune system	leukocyte	plasma cell	CL:0000786
Abnormality of the immune system	leukocyte	alternatively activated macrophage	CL:0000890
Abnormality of the immune system	leukocyte	thymocyte	CL:0000893
Abnormality of the immune system	leukocyte	innate lymphoid cell	CL:0001065

Table 4: On-target cell types for each HPO ancestral branch.

hpo_branch	cl_branch	cl_name	cl_id
Abnormality of the musculoskeletal system	cell of skeletal muscle / chondrocyte	chondrocyte	CL:0000138
Abnormality of the musculoskeletal system	cell of skeletal muscle / chondrocyte	cell of skeletal muscle	CL:0000188
Abnormality of the musculoskeletal system	cell of skeletal muscle / chondrocyte	skeletal muscle satellite cell	CL:0000594
Abnormality of the nervous system	neural cell	bipolar neuron	CL:0000103
Abnormality of the nervous system	neural cell	granule cell	CL:0000120
Abnormality of the nervous system	neural cell	Purkinje cell	CL:0000121
Abnormality of the nervous system	neural cell	glial cell	CL:0000125
Abnormality of the nervous system	neural cell	astrocyte	CL:0000127
Abnormality of the nervous system	neural cell	oligodendrocyte	CL:0000128
Abnormality of the nervous system	neural cell	microglial cell	CL:0000129
Abnormality of the nervous system	neural cell	neuroendocrine cell	CL:0000165
Abnormality of the nervous system	neural cell	chromaffin cell	CL:0000166
Abnormality of the nervous system	neural cell	photoreceptor cell	CL:0000210
Abnormality of the nervous system	neural cell	inhibitory interneuron	CL:0000498
Abnormality of the nervous system	neural cell	neuron	CL:0000540
Abnormality of the nervous system	neural cell	neuronal brush cell	CL:0000555
Abnormality of the nervous system	neural cell	amacrine cell	CL:0000561
Abnormality of the nervous system	neural cell	GABAergic neuron	CL:0000617
Abnormality of the nervous system	neural cell	Mueller cell	CL:0000636
Abnormality of the nervous system	neural cell	glutamatergic neuron	CL:0000679
Abnormality of the nervous system	neural cell	retinal ganglion cell	CL:0000740
Abnormality of the nervous system	neural cell	retina horizontal cell	CL:0000745
Abnormality of the nervous system	neural cell	Schwann cell	CL:0002573
Abnormality of the nervous system	neural cell	retinal pigment epithelial cell	CL:0002586
Abnormality of the nervous system	neural cell	visceromotor neuron	CL:0005025

Table 4: On-target cell types for each HPO ancestral branch.

hpo_branch	cl_branch	cl_name	cl_id
Abnormality of the nervous system	neural cell	sympathetic neuron	CL:0011103
Abnormality of the respiratory system	respiratory epithelial cell / epithelial cell of lung	type II pneumocyte	CL:0002063
Abnormality of the respiratory system	respiratory epithelial cell / epithelial cell of lung	epithelial cell of lower respiratory tract	CL:0002632

Table 5: Encodings for Age of Death scores. Assigned numeric values for the Age of Death scores within the HPO annotations.

hpo_id	hpo_name	encoding
HP:0003826	Stillbirth	1
HP:0005268	Miscarriage	1
HP:0034241	Prenatal death	1
HP:0003811	Neonatal death	2
HP:0001522	Death in infancy	3
HP:0003819	Death in childhood	4
HP:0011421	Death in adolescence	5
HP:0100613	Death in early adulthood	6
HP:0033763	Death in adulthood	7
HP:0033764	Death in middle age	7
HP:0033765	Death in late adulthood	8