

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

<sup>4</sup> Brian M. Schilder      Kitty B. Murphy      Robert Gordon-Smith      Jai Chapman  
<sup>5</sup>                          Momoko Otani                          Nathan G. Skene

<sup>6</sup> 2024-07-03

## 7 Abstract

8 Rare diseases (RDs) are an extremely heterogeneous and underserved category of medical conditions. While  
9 the majority of RDs are strongly genetic, it remains largely unknown via which physiological mechanisms  
10 genetics cause RD. Therefore, we sought to systematically characterise the cell type-specific mechanisms  
11 underlying all RD phenotypes with a known genetic cause by leveraging the Human Phenotype Ontology  
12 and transcriptomic single-cell atlases of the entire human body from embryonic, foetal, and adult samples.  
13 In total we identified significant associations between 201 cell types and 9,575/11,028 (86.7%) unique phe-  
14 notypes across 8,628 RDs. This greatly the collective knowledge of RD phenotype-cell type mechanisms.  
15 Next, developed a pipeline to identify cell type-specific targets for phenotypes ranked by metrics of severity  
16 (e.g. lethality, motor/mental impairment) and compatibility with gene therapy (e.g. filtering out physical  
17 malformations). Furthermore, we have made these results entirely reproducible and freely accessible to the  
18 global community to maximise their impact. To summarise, this work represents a significant step forward  
19 in the mission to treat patients across an extremely diverse spectrum of serious RDs.

## 20 Introduction

21 While rare diseases (RDs) are individually uncommon, they collectively account for an enormous global  
22 disease burden with over 10,000 recognised RDs affecting at least 300-400 million people globally<sup>1</sup> (1 in  
23 10-20 people)<sup>2</sup>. Over 75% of RDs primarily affect children with a 30% mortality rate by 5 years of age<sup>3</sup>.  
24 Despite the prevalence and severity of RDs, patients suffering from these conditions are vastly underserved  
25 due to several contributing factors. First, diagnosis is extremely challenging due to the highly variable  
26 clinical presentations of many of these diseases. The diagnostic odyssey can take patients and their families  
27 decades, with an average time to diagnosis of 5 years<sup>4</sup>. Of those, ~46% receive at least one incorrect  
28 diagnosis and over 75% of all patients never receive any diagnosis<sup>5</sup>. Second, prognosis is also made difficult  
29 by high variability in disease course and outcomes which makes matching patients with effective and timely  
30 treatment plans even more challenging. Finally, even for patients who receive an accurate diagnosis/prognosis,  
31 treatments are currently only available for less than 5% of all RDs<sup>6</sup>. In addition to the scientific challenges of  
32 understanding RDs, there are strong financial disincentives for pharmaceutical and biotechnology companies  
33 to develop expensive therapeutics for exceedingly small RD patient populations with little or no return  
34 on investment<sup>7,8</sup>. Those that have been produced are amongst the world's most expensive drugs, greatly  
35 limiting patients' ability to access it<sup>9,10</sup>. New high-throughput approaches for the development of rare disease  
36 therapeutics could greatly reduce costs (for manufacturers and patients) and accelerate the timeline from  
37 discovery to delivery.

38 A major challenge in both healthcare and scientific research is the lack of standardised medical terminology.  
39 Even in the age of electronic healthcare records (EHR) much of the information about an individual's history  
40 is currently fractured across healthcare providers, often with differing nomenclatures for the same conditions.

41 The Human Phenotype Ontology (HPO) is a hierarchically organised set of controlled clinical terms that  
42 provides a much needed common framework by which clinicians and researchers can precisely communi-  
43 cate patient conditions<sup>14</sup>. The HPO spans all domains of human physiology and currently describes 18082  
44 phenotypes across 10,300 RDs. Each phenotype and disease is assigned its own unique identifier and organ-  
45 ised as a hierarchical graph, such that higher-level terms describe broad phenotypic categories or *branches*  
46 (e.g. *HP:0033127*: ‘Abnormality of the musculoskeletal system’ which contains 4495 unique phenotypes) and  
47 lower-level terms describe increasingly precise phenotypes (e.g. *HP:0030675*: ‘Contracture of proximal inter-  
48 phalangeal joints of 2nd-5th fingers’). It has already been integrated into healthcare systems and clinical  
49 diagnostic tools around the world, with increasing adoption over time<sup>11</sup>. Standardised frameworks like the  
50 HPO also allow us to aggregate relevant knowledge about the molecular mechanisms underlying each RD.

51 Over 80% of RDs have a known genetic cause<sup>15,16</sup>. Since 2008, the HPO has been continuously updated  
52 using curated knowledge from the medical literature, as well as by integrating databases of expert validated  
53 gene-phenotype relationships, such as OMIM<sup>17-19</sup>, Orphanet<sup>20,21</sup>, and DECIPHER<sup>22</sup>. Many of these gene  
54 annotations are manually or semi-manually curated by expert clinicians from case reports of rare disease  
55 patients in which the causal gene is identified through whole exome or genome sequencing. Currently, the  
56 HPO contains gene annotations for 11,047 phenotypes across 8,631 diseases. Yet genes alone do not tell the  
57 full story of how RDs come to be, as their expression and functional relevance varies drastically across the  
58 multitude of tissues and cell types contained within the human body. Our knowledge of the physiological  
59 mechanisms via which genetics cause pathogenesis is lacking for most RDs, severely hindering our ability to  
60 effectively diagnose, prognosis and treat RD patients.

61 Our knowledge of cell type-specific biology has exploded over the course of the last decade and a half,  
62 with numerous applications in both scientific and clinical practices<sup>23-25</sup>. In particular, single-cell RNA-seq  
63 (scRNA-seq) has allowed us to quantify the expression of every gene (i.e. the transcriptome) in individual  
64 cells. More recently, comprehensive single-cell transcriptomic atlases across tissues have also emerged<sup>26,27</sup>.  
65 In particular, the Descartes Human<sup>28</sup> and Human Cell Landscape<sup>29</sup> projects provide comprehensive multi-  
66 system scRNA-seq atlases in embryonic, foetal, and adult human samples from across the human body.  
67 These datasets provide data-driven gene signatures for hundreds of cell subtypes. Given that many disease-  
68 associated genes are expressed in some cell types but not others, we can infer that disruptions to these genes  
69 will have varying impact across cell types. By comparing the aggregated disease gene annotations with  
70 cell type-specific expression profiles, we can therefore uncover the cell types and tissues via which diseases  
71 mediate their effects.

72 Here, we combine and extend several of the most comprehensive genomic and transcriptomic resources  
73 currently available to systematically uncover the cell types underlying granular phenotypes across 8,628  
74 diseases. This information is essential for the development of novel therapeutics, especially gene therapy  
75 modalities such as adeno-associated viral (AAV) vectors in which advancement have been made in their

76 ability selectively target specific cell types<sup>30,31</sup>. Precise knowledge of relevant cell types and tissues causing  
77 the disease can improve safety by minimising harmful side effects in off-target cell types and tissues. It  
78 can also enhance efficacy by efficiently delivering expensive therapeutic payloads to on-target cell types and  
79 tissues. For example, if a phenotype primarily effects retinal cells, then the gene therapy would be optimised  
80 for delivery to retinal cells of the eye. Using this information, we developed a high-throughput pipeline for  
81 comprehensively nominating cell type-resolved gene therapy targets across thousands of RD phenotypes. As  
82 a prioritisation tool, we sorted these targets based on the severity of their respective phenotypes, using a  
83 generative AI-based approach<sup>32</sup>. Together, our study dramatically expands the available knowledge of the  
84 cell types, organ systems and life stages underlying RD phenotypes.

## 85 Results

### 86 Phenotype-cell type associations

87 In this study we systematically investigated the cell types underlying phenotypes across the HPO. For  
88 each phenotype we created a list of associated genes weighted by the strength of the evidence supporting  
89 those associations, imported from the Gene Curation Coalition (GenCC)<sup>33</sup>. Analogously, we created gene  
90 expression profiles for each cell type in our scRNA-seq atlases and then applied normalisation to compute  
91 how specific the expression of each gene is to each cell type. To assess consistency in the phenotype-cell type  
92 associations, we used multiple scRNA-seq atlases: Descartes Human (~4 million single-nuclei and single-cells  
93 from 15 fetal tissues)<sup>28</sup> and Human Cell Landscape (~703,000 single-cells from 49 embryonic, fetal and adult  
94 tissues)<sup>29</sup>. To identify phenotype-cell type relationships, we ran a series of linear regression models to test  
95 for the relationship between each combination of phenotype and cell type. We applied multiple testing  
96 correction to control the false discovery rate (FDR) across all tests.

97 Within the results using the Descartes Human single-cell atlas, 19,929/ 848,078 (2.35%) tests across 77/  
98 77 (100%) cell types and 7,340/11,047 (66.4%) phenotypes revealed significant phenotype-cell type asso-  
99 ciations after multiple-testing correction (FDR<0.05). Using the Human Cell Landscape single-cell atlas,  
100 26,585/1,358,916 (1.96%) tests across 124/124 (100%) cell types and 9,049/11,047 (81.9%) phenotypes showed  
101 significant phenotype-cell type associations (FDR<0.05). The median number of significantly associated phe-  
102 notypes per cell type was 252 (Descartes Human) and 200 (Human Cell Landscape), respectively.

103 Across both single-cell references, the median number of significantly associated cell types per phenotype was  
104 3, suggesting reasonable specificity of the testing strategy. Within the HPO, 8,628/8,631 (~100%) of diseases  
105 gene annotations showed significant cell type associations for at least one of their respective phenotypes. A  
106 summary of the genome-wide results stratified by single-cell atlas can be found in Table 2.

<sup>107</sup> **Validation of expected phenotype-cell type relationships**

<sup>108</sup> We intuitively expect that abnormalities of an organ system will often be driven by cell types within that  
<sup>109</sup> system. The HPO has broad categories at the higher level of the ontology, enabling us to systematically test  
<sup>110</sup> this. For example, phenotypes associated with the heart should generally be caused by cell types of the heart  
<sup>111</sup> (i.e. cardiocytes), while abnormalities of the nervous system should largely be caused by neural cells. There  
<sup>112</sup> will of course be exceptions to this. For example, some immune disorders can cause intellectual disability  
<sup>113</sup> through neurodegeneration. Nevertheless, it is reasonable to expect that abnormalities of the nervous system  
<sup>114</sup> will be most often associated with neural cells. All cell types in our single-cell reference atlases were mapped  
<sup>115</sup> onto the Cell Ontology (CL); a controlled vocabulary of cell types organised into hierarchical branches  
<sup>116</sup> (e.g. neural cell include neurons and glia, which in turn include their respective subtypes).

<sup>117</sup> Here, we consider a cell type to be *on-target* relative to a given HPO branch if it belongs to one of the  
<sup>118</sup> matched CL branches (see Table 1). Within each high-level branch in the HPO shown in Fig. 1b, we tested  
<sup>119</sup> whether each cell type was more often associated with phenotypes in that branch relative to those in all  
<sup>120</sup> other branches (including those not shown). We then checked whether each cell type was overrepresented  
<sup>121</sup> (at FDR<0.05) within its respective on-target HPO branch, where the number of phenotypes within that  
<sup>122</sup> branch. Indeed, we found that all 7 HPO branches were disproportionately associated with on-target cell  
<sup>123</sup> types from their respective organ systems.

Table 1: Cross-ontology mappings between HPO and CL branches. The last two columns represent the number of cell types that were overrepresented in the on-target HPO branch and the total number of cell types in that branch. A disaggregated version of this table with all descendant cell type names is available in Table 5.

| HPO branch                                | Phenotypes |   | Cell types<br>(overrepresented) | Cell types<br>(total) |
|---|------------|---|---------------------------------|-----------------------|
|   | (total)    | CL branch   |                                 |                       |
| Abnormality of the cardiovascular system  | 673        | cardiocyte  | 5                               | 6                     |
| Abnormality of the endocrine system       | 291        | endocrine cell                                      | 3                               | 4                     |
| Abnormality of the eye                    | 721        | photoreceptor cell/retinal cell                     | 5                               | 5                     |
| Abnormality of the immune system          | 255        | leukocyte   | 14                              | 14                    |
| Abnormality of the musculoskeletal system | 2155       | cell of skeletal muscle/chondrocyte                 | 4                               | 4                     |
| Abnormality of the nervous system         | 1647       | neural cell   | 17                              | 24                    |
| Abnormality of the respiratory system     | 292        | respiratory epithelial cell/epithelial cell of lung | 3                               | 3                     |

124 In addition to binary metrics of a cell type being associated with a phenotype or not, we also used association  
 125 test p-values as a proxy for the strength of the association. We hypothesized that the more significant the  
 126 association between a phenotype and a cell type, the more likely it is that the cell type is on-target for its  
 127 respective HPO branch. To evaluate whether this, we grouped the association  $-\log_{10}(\text{p-values})$  into 6 bins.  
 128 For each HPO-CL branch pairing, we then calculated the proportion of on-target cell types within each bin.  
 129 We found that the proportion of on-target cell types increased with increasing significance of the association  
 130 ( $\rho=0.63$ ,  $p=1.1\text{e-}06$ ). For example, abnormalities of the nervous system with  $-\log_{10}(\text{p-values}) = 1$ , only  
 131 16% of the associated cell types were neural cells. Whereas for those with  $-\log_{10}(\text{p-values}) = 6$ , 46% were  
 132 neural cells despite the fact that this class of cell types only constituted 23% of the total cell types tested  
 133 (i.e. the baseline). This shows that the more significant the association, the more likely it is that the cell  
 134 type is on-target.

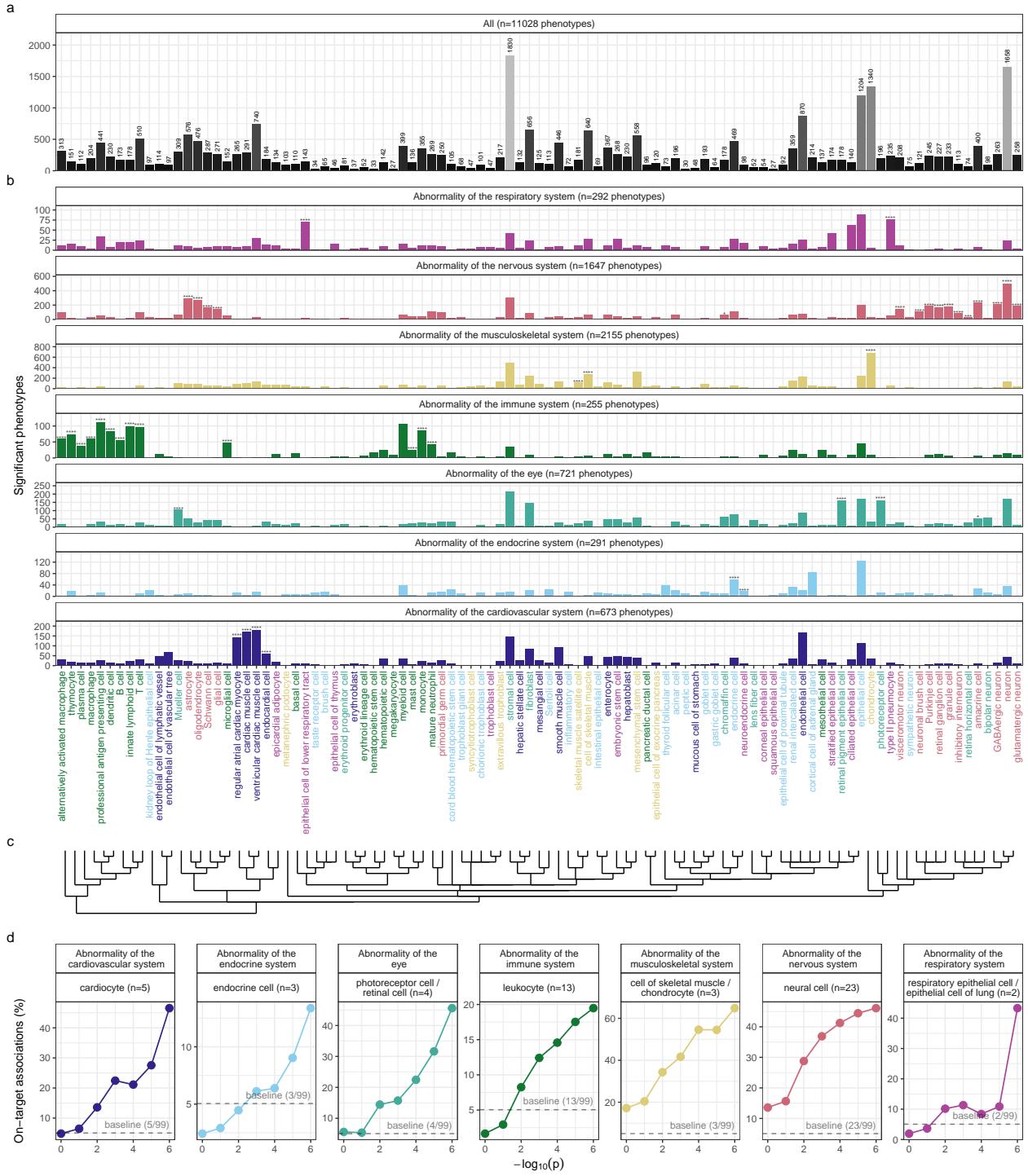


Figure 1: High-throughput analysis reveals cell types underlying thousands of rare disease phenotypes. **a**, Some cell types are much more commonly associated with phenotypes than others. Bar height indicates the total number of significant phenotype enrichments per cell type ( $FDR < 0.05$ ) across all branches of the HPO. **b**, Analyses reveal expected and novel cell type associations within high-level HPO branches. 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;  $FDR < 0.0001$  (\*\*\*\*),  $FDR < 0.001$  (\*\*),  $FDR < 0.01$  (\*\*),  $FDR < 0.05$  (\*). **c**, Ontological relatedness of cell types in the Cell Ontology (CL)<sup>34</sup>. **d**, The proportion of on-target associations (*y-axis*) increases with greater test significance (*x-axis*). Percentage of significant phenotype associations with on-target cell types (second row of facet labels), respective to the HPO branch.

135 **Validation of inter- and intra-dataset consistency**

136 Next, we sought to validate the consistency of our results across the two single-cell reference datasets  
137 (Descartes Human vs. Human Cell Landscape) across the subset of overlapping cell types Fig. 10. In total  
138 there were 142285 phenotype-cell type associations to compare across the two datasets (across 10945 phe-  
139 notypes and 13 cell types annotated to the exact same CL term. We found that the correlation between  
140 p-values of the two datasets was high ( $\rho = 0.491956950302773, p = 1.07617274060444e - 93$ ). Within the  
141 subset of results that were significant in both single-cell datasets (FDR<0.05), we found that correlation  
142 of the association effect size were even stronger ( $\rho = 0.722784999300949, p = 1.07617274060444e - 93$ ).  
143 We also checked for the intra-dataset consistency between the p-values of the foetal and adult samples in  
144 the Human Cell Landscape, showing a very similar degree of correlation as the inter-dataset comparison  
145 ( $\rho = 0.436339765865796, p = 2.36197328541783e - 149$ ). Together, these results suggest that our approach  
146 to identifying phenotype-cell type associations is highly replicable and generalisable to new datasets.

147 **More specific phenotypes are associated with fewer genes and cell types**

148 Higher levels of the ontology are broad classes of phenotype (e.g. ‘Abnormality of the nervous system’) while  
149 the lower levels can get very detailed (e.g. ‘Spinocerebellar atrophy’). The higher level phenotypes inherit  
150 all genes associated with lower level phenotypes, so naturally they have more genes than the lower level  
151 phenotypes (Fig. 2a;  $p = 2.2250738585072e - 308, \rho = -0.263403620608294$ ).

152 Next, we reasoned that the more detailed and specific a phenotype is, the more likely it is to be driven by one  
153 cell type. For example, while ‘Neurodevelopmental abnormality’ could plausibly be driven by any/all cell  
154 types in the brain, it is more likely that ‘Impaired visuospatial constructive cognition’ is driven by a single  
155 cell type. This was indeed the case, as we observed a strongly significant negative correlation between the  
156 two variables (Fig. 2b;  $p = 2.2250738585072e - 308, \rho = -0.292677384995276$ ). We also found that the  
157 phenotype-cell type association p-values increased with greater phenotype specificity, reflecting the decreasing  
158 overall number of associated cell types at each ontological level (Fig. 2c;  $p = 2.2250738585072e - 308, \rho =$   
159  $0.256729073040334$ ).

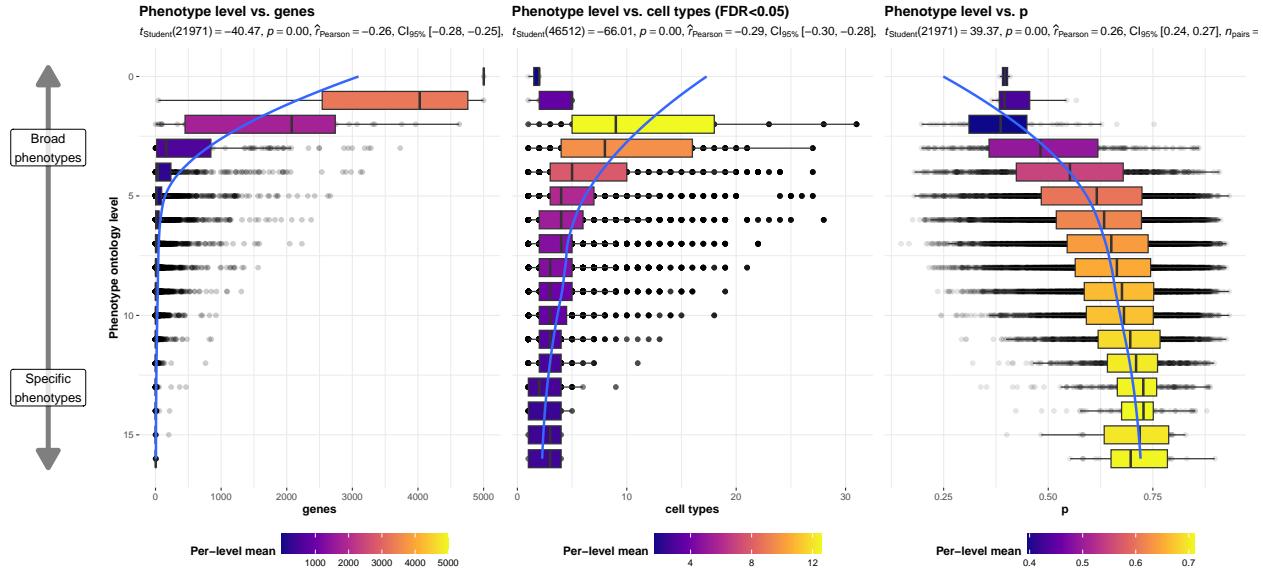


Figure 2: More specific phenotypes are associated with fewer, more specific genes and cell types. Box plots showing relationship between HPO phenotype level and **a**, the number of genes annotated to each phenotype, **b**, the number of significantly enriched cell types, **c**, the p-values of phenotype-cell type association tests. 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.

## 160 Hepatoblasts have a unique role in recurrent Neisserial infections

161 We selected the HPO term ‘Recurrent bacterial infections’ and all of its descendants (19 phenotypes) as an  
 162 example of how investigations at the level of granular phenotypes can reveal different cell type-specific  
 163 mechanisms (Fig. 3). As expected, these phenotypes are primarily associated with immune cell types  
 164 (e.g. macrophages, dendritic cells, T cells, monocytes, neutrophils). Some associations confirm relation-  
 165 ships previously suggested in the literature, such as that between ‘Recurrent staphylococcal infections’ and  
 166 myeloid cells<sup>35–38</sup>. Specifically, our results pinpoint monocytes as the most strongly associated cell subtypes  
 167 ( $FDR = 1.02624301552218e - 30, B = 0.17635450011961$ ).

168 In contrast to all other recurrent infection types, ‘Recurrent Neisserial infections’ highlighted a novel asso-  
 169 ciation with hepatoblasts (Descartes Human :  $FDR = 1.13424027668278e - 06, B = 0.0823733563618383$ ).  
 170 Whilst unexpected, a convincing explanation involves the complement system, a key driver of innate immune  
 171 response to Neisserial infections. Hepatocytes, which derive from hepatoblasts, produce the majority of com-  
 172 plement proteins<sup>39</sup>, and Kupffer cells express complement receptors<sup>40</sup>. In addition, individuals with deficits  
 173 in complement are at high risk for Neisserial infections<sup>41,42</sup>, and a genome-wide association study in those  
 174 with a Neisserial infection identified risk variants within complement proteins<sup>43</sup>. While the potential of ther-  
 175 apeutically targeting complement in RDs (including Neisserial infections) has been proposed previously<sup>44,45</sup>,  
 176 performing this in a gene- and cell type-specific manner may help to improve efficacy and reduce toxicity  
 177 (e.g. due to off-target effects). Importantly, there are over 56 known genes within the complement system<sup>46</sup>,

<sup>178</sup> highlighting the need for a systematic, evidence-based approach to identify effective gene targets.

<sup>179</sup> Also of note, despite the fact that our datasets contain both hepatoblasts and their mature counterpart,  
<sup>180</sup> hepatocytes, only the hepatoblasts showed this association. This suggests that the genetic factors that  
<sup>181</sup> predispose individuals for risk of Neisserial infections are specifically affecting hepatoblasts before they  
<sup>182</sup> become fully differentiated. It is also notable that these phenotypes were the only ones within the ‘Recur-  
<sup>183</sup> rent bacterial infections’ branch, or even the broader ‘Recurrent infections’ branch, perhaps indicating a  
<sup>184</sup> unique role for hepatoblasts in recurrent infectious disease. The only phenotypes within the even broader  
<sup>185</sup> ‘Abnormality of the immune system’ HPO branch that significantly associated with mature hepatocytes  
<sup>186</sup> were ‘Pancreatitis’ ( $FDR = 0.0207647478699714, B = 0.0525112272785126$ ) and ‘Susceptibility to chicken-  
<sup>187</sup> pox’ ( $FDR = 0.0119527486705115, B = 0.0549042507312806$ ) both of which are well-known to involve the  
<sup>188</sup> liver<sup>47-49</sup>.

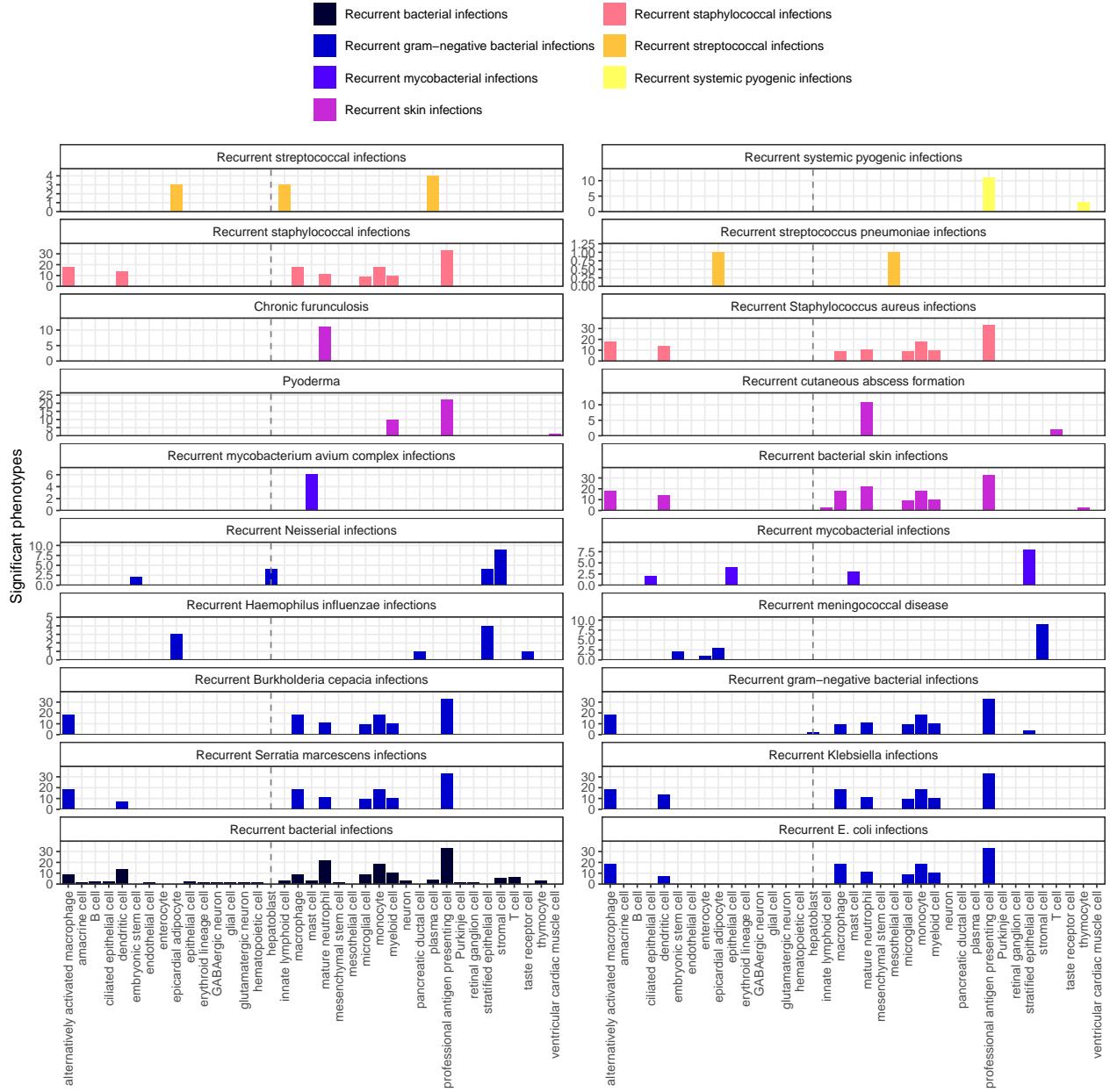


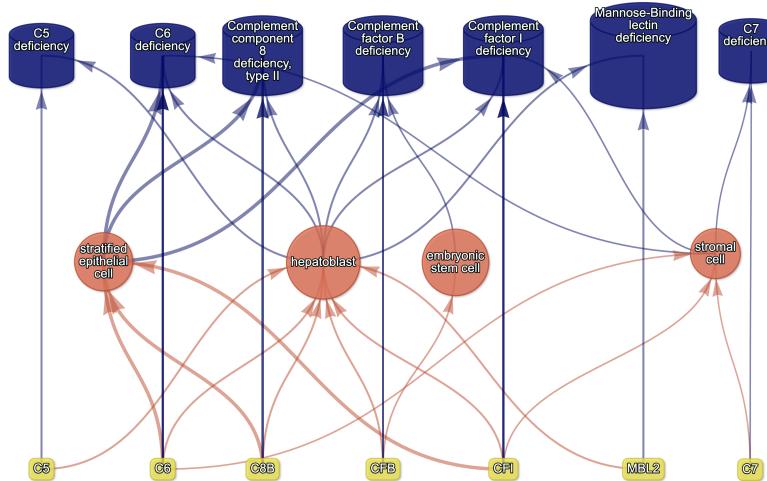
Figure 3: Hepatoblasts have a unique role in recurrent Neisserial infections. Significant phenotype-cell type tests for phenotypes within the branch ‘Recurrent bacterial infections’. Amongst all different kinds of recurrent bacterial infections, hepatoblasts (highlighted by vertical dotted lines) are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

189 Phenotypes can be associated with multiple diseases, cell types and genes. In addition to hepatoblasts,  
 190 ‘Recurrent Neisserial infections’ were also associated with stromal cells ( $FDR=4.6e-06$ ,  $\beta=0.079$ ), stratified  
 191 epithelial cells ( $FDR=1.7e-23$ ,  $\beta=0.15$ ), and embryonic stem cells ( $FDR=5.4e-05$ ,  $\beta=0.074$ ). RNI is a phe-  
 192 notype of 7 different diseases (‘C5 deficiency’, ‘C6 deficiency’, ‘C7 deficiency’, ‘Complement component 8

<sup>193</sup> deficiency, type II', 'Complement factor B deficiency', 'Complement factor I deficiency', 'Mannose-Binding  
<sup>194</sup> lectin deficiency').

<sup>195</sup> Next, we sought to link multi-scale mechanisms at the levels of disease, phenotype, cell type, and gene and  
<sup>196</sup> visualise these as a network (Fig. 4). This revealed that genetic deficiencies in different complement system  
<sup>197</sup> genes (e.g. *C5*, *C8*, and *C7*) are primarily mediated by different cell types (hepatoblasts, stratified epithelial  
<sup>198</sup> cells, and stromal cells, respectively). While genes of the complement system are expressed throughout many  
<sup>199</sup> different tissues and cell types, these results indicate that different subsets of these genes may mediate their  
<sup>200</sup> effects through different cell types. This finding suggests that investigating (during diagnosis) and targeting  
<sup>201</sup> (during treatment) different cell types may be critical for the diagnosis and treatment of these closely related,  
<sup>202</sup> yet mechanistically distinct, diseases.

a



b

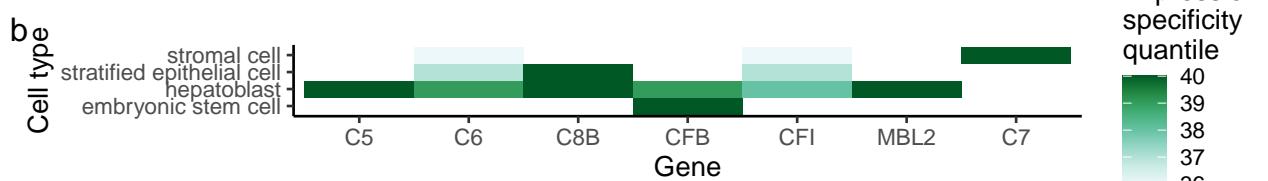


Figure 4: Multi-scale mechanisms of Recurrent Neisserial infections. Starting from the bottom of the plot, one can trace how causal genes (yellow boxes) mediate their effects through cell types (orange circles), phenotypes (purple cylinders) and ultimately diseases (blue cylinders). Cell types are connected to phenotypes via association testing ( $FDR < 0.05$ ), and to diseases when the symptom gene set overlap is  $> 25\%$ . Only the top driver genes (specificity quantiles  $> 75\%$ ) mediating each phenotype-cell type association are shown. Nodes were spatially arranged using the Sugiyama algorithm<sup>130</sup>.

203 Multi-scale mechanisms of Recurrent Neisserial infections. Starting from the bottom of the plot, one can  
204 trace how causal genes (yellow boxes) mediate their effects through cell types (orange circles), phenotypes  
205 (purple cylinders) and ultimately diseases (blue cylinders). Cell types are connected to phenotypes via  
206 association testing ( $FDR < 0.05$ ), and to diseases when the symptom gene set overlap is  $> 25\%$ . Only the top  
207 driver genes (specificity quantiles  $> 75\%$ ) mediating each phenotype-cell type association are shown. Nodes  
208 were spatially arranged using the Sugiyama algorithm<sup>50</sup>.

209 **Monarch Knowledge Graph recall**

210 Next, we used the Monarch Knowledge Graph (MKG) as a proxy for the field's current state of knowledge of  
211 phenotype-cell type associations. We evaluated the proportion of MKG associations that were recapitulation  
212 by our results Fig. 11. For each phenotype-cell type association in the MKG, we computed the percent of  
213 cell types recovered in our association results at a given ontological distance according to the CL ontology.  
214 An ontological distance of 0 means that our nominated cell type was as close as possible to the MKG  
215 cell type after adjusting for the cell types available in our single-cell references. Instances of exact overlap  
216 of terms between the MKG and our results would qualify as an ontological distance of 0 (e.g. 'monocyte'  
217 vs. 'monocyte'). Greater ontological distances indicate further divergence between the MKG cell type and  
218 our nominated cell type. A distance of 1 indicating that the MKG cell type was one step away from our  
219 nominated cell type in the CL ontology graph (e.g. 'monocyte' vs. 'classical monocyte'). The maximum  
220 possible percent of recovered terms is capped by the percentage of MKG ground-truth phenotypes we were  
221 able to find at least one significant cell type association for at  $FDR_{pc}$ .

222 In total, our results contained at least one significant cell type associations for 90% of the phenotypes de-  
223 scribed in the MKG. Of these phenotypes, we captured 55% of the MKG phenotype-cell associations at an  
224 ontological distance of 0 (i.e. the closest possible Cell Ontology term match). Recall increased with greater  
225 flexibility in the matching of cell type annotations. At an ontological distance of 1 (e.g. 'monocyte' vs. 'clas-  
226 sical monocyte'), we captured 77% of the MKG phenotype-cell associations. Recall reached a maximum of  
227 90% at a ontological distance of 5. This recall percentage is capped by the proportion of phenotype for  
228 which we were able to find at least one significant cell type association for. It should be noted that we  
229 were unable to compute precision as the MKG (and other knowledge databases) only provide true positive  
230 associations. Identifying true negatives (e.g. a cell type is definitely never associated with a phenotype) is  
231 a fundamentally more difficult task to resolve as it would require proving the null hypothesis. Regardless,  
232 these benchmarking tests suggests that our results are able to recover the majority of known phenotype-cell  
233 type associations while proposing many new associations.

234 **Annotation of phenotypes using generative large language models**

235 Severity annotations were gathered from GPT-4 for 16982/18082 (94%) HPO phenotypes in our companion  
236 study<sup>32</sup>. Benchmarking tests of these results using ground-truth HPO branch annotations. For example,

237 phenotypes within the ‘Blindness’ HPO branch (*HP:0000618*) were correctly annotated as causing blindness  
238 by GPT-4. Across all annotations, the recall rate of GPT-4 annotations was 96% (min=89%, max=100%,  
239 SD=4.5) with a mean consistency score of 91% (min=81%, max=97%, SD=5.7) for phenotypes whose  
240 annotation were collected more than once. This clearly demonstrates the ability of GPT-4 to accurately  
241 annotate phenotypes. This allowed us to begin using these annotations to compute systematically collected  
242 severity scores for all phenotypes in the HPO.

243 From these annotations we computed a weighted severity score metric for each phenotype ranging from 0-100  
244 (100 being the theoretical maximum severity of a phenotype that always causes every annotation). Within  
245 our annotations, the most severe phenotype was ‘Atrophy/Degeneration affecting the central nervous system’  
246 (*HP:0007367*) with a severity score of 47, followed by ‘Anencephaly’ (*HP:0002323*) with a severity score of  
247 45. There were 677 phenotypes with a severity score of 0 (e.g. ‘Thin toenail’). The mean severity score  
248 across all phenotypes was 10 (median=9.4, standard deviation=6.4).

#### 249 **Congenital phenotypes are associated with foetal cell types**

250 To further verify the biological relevance of our results, we examined the association of foetal cell types with  
251 phenotypes annotated as congenital in onset. As expected, the frequency of congenital onset with each pheno-  
252 type (as determined by GPT-4 annotations) was strongly predictive with the proportion of significantly asso-  
253 ciated foetal cell types in our results ( $p = 2.03822771861244e-203$ ,  $\chi^2_{Pearson} = 939.821618487545$ ,  $\hat{V}_{Cramer} =$   
254 0.143370676240175). Furthermore, increasing congenital frequency annotation (on an ordinal scale) corre-  
255 sponded to an increase in the proportion of foetal cell types: ‘always’=24% (n=1636 associations), ‘of-  
256 ten’=20% (n=2979 associations), ‘rarely’=12% (n=1956 associations), ‘never’=10% (n=811 associations).  
257 This is consistent with the expected role of foetal cell types in development and the aetiology of congenital  
258 disorders.

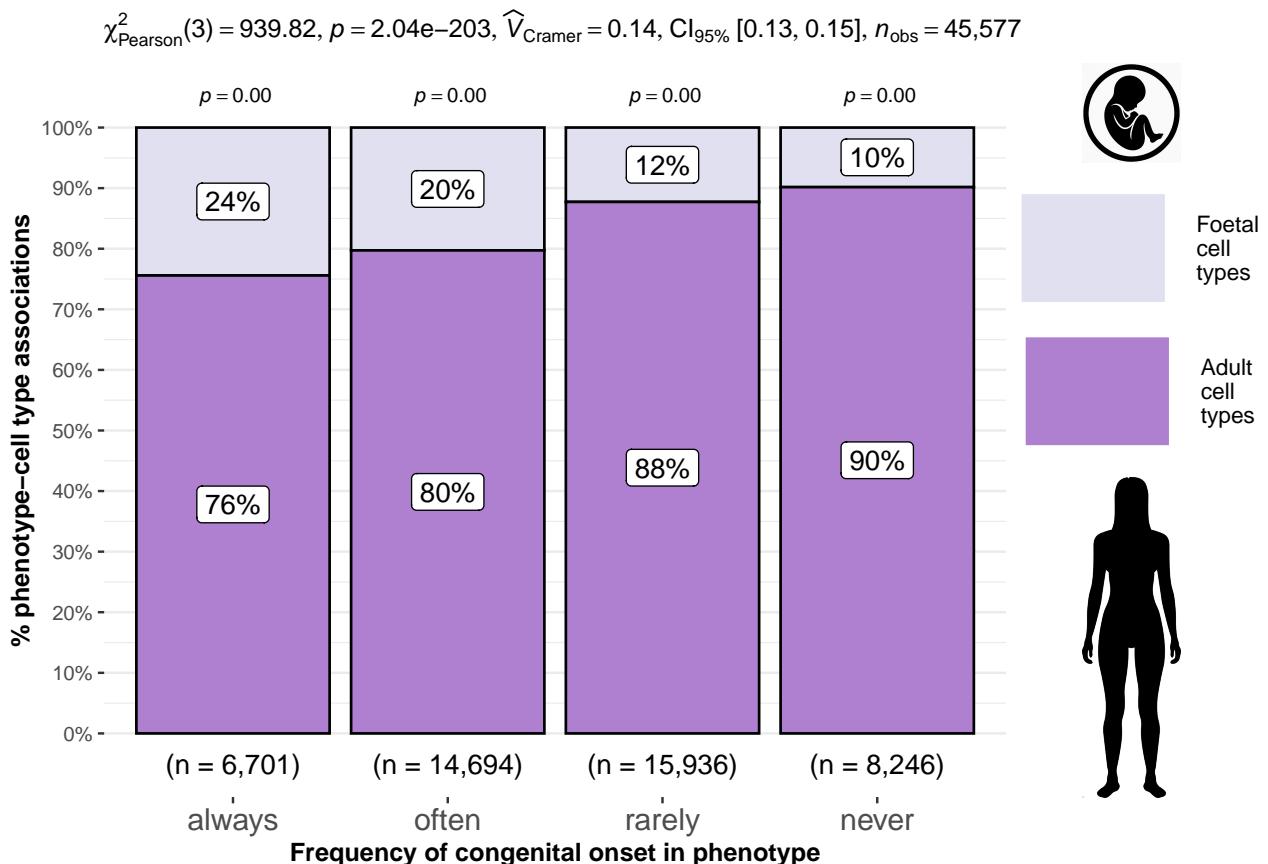


Figure 5: Congenital phenotypes are more often associated with foetal cell types. As a phenotype is more often congenital in nature, the greater proportion of foetal cell types are significantly associated with it. The summary statistics in the plot title are the results of a  $\chi^2$  tests of independence 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  $\chi^2$  tests to determine whether the proportion of foetal vs. non-foetal cell types significantly differ 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.

259 We also found that some branches of the HPO were more commonly enriched in foetal cell types compared  
 260 to others ( $\hat{V}_{\text{Cramer}}=0.22, p<2.2e-308$ ). See The branch with the greatest proportion of fetal cell type enrich-  
 261 ments was ‘Abnormality of limbs’ (35%), followed by ‘Growth abnormality’ (32%) and ‘Abnormality of the  
 262 musculoskeletal system’ (29%). These results align well with the fact that physical malformations tend to  
 263 be developmental in origin.

#### 264 Therapeutic target identification

265 Next, we identified putative cell type-specific gene targets for several severe disease phenotypes. This yielded  
 266 putative therapeutic targets for 5252 phenotypes across 4823 diseases in 201 cell types and 3150 genes  
 267 (Fig. 13). While this constitutes a large number of genes in total, each phenotype was assigned a median of  
 268 2 gene targets (mean=3.3, min=1, max=10). Relative to the number of genes annotations per phenotype

269 in the HPO overall (median=7, mean=62, min=1, max=5003) this represents a substantial decrease in the  
270 number of candidate target genes, even when excluding high-level phenotypes (HPO level>3). It is also  
271 important to note that the phenotypes in the prioritised targets list are ranked by their severity, allowing  
272 us to distinguish between phenotypes with a high medical urgency (e.g. ‘Hydranencephaly’) from those with  
273 lower medical urgency (e.g. ‘Hyperplastic labia majora’). This can be useful for both clinicians, biomedical  
274 scientists, and pharmaceutical manufacturers who wish to focus their research efforts on phenotypes with  
275 the greatest need for intervention.

276 Across all phenotypes, epithelial cell were most commonly implicated (838 phenotypes), followed by stromal  
277 cell (627 phenotypes), stromal cell (627 phenotypes), neuron (475 phenotypes), chondrocyte (383 pheno-  
278 types), and endothelial cell (361 phenotypes). Grouped by higher-order ontology category, ‘Abnormality of  
279 the musculoskeletal system’ had the greatest number of enriched phenotypes (959 phenotypes, 857 genes),  
280 followed by ‘Abnormality of the nervous system’ (733 phenotypes, 1137 genes), ‘Abnormality of head or  
281 neck’ (543 phenotypes, 990 genes), ‘Abnormality of the genitourinary system’ (443 phenotypes, 696 genes),  
282 and ‘Abnormality of the eye’ (377 phenotypes, 548 genes).

### 283 Therapeutic target validation

284 To determine whether the genes prioritised by our therapeutic targets pipeline were plausible, we checked  
285 what percentage of gene therapy targets we recapitulated. Data on therapeutic approval status was gathered  
286 from the Therapeutic Target Database (TTD; release 2024-07-03)<sup>51</sup>. Overall, we prioritised 81% of all non-  
287 failed existing gene therapy targets. A hypergeometric test confirmed that our prioritised targets were signif-  
288 icantly enriched for non-failed gene therapy targets ( $p = 0.0018198077753355$ ). Importantly, we did not pri-  
289 oritise any of the failed therapeutics (0%), defined as having been terminated or withdrawn from the market.  
290 The hypergeometric test for depletion of failed targets did not reach significance ( $p = 0.370503597122302$ ),  
291 but this is to be expected as there was only one failed gene therapy target in the TTD database.

292 Even when considering therapeutics of any kind (Fig. 14), not just gene therapies, we recapitulated 40%  
293 of the non-failed therapeutic targets and 0% of the terminated/withdrawn therapeutic targets (n=1255).  
294 Here we found that our prioritised targets were significantly enriched for non-failed therapeutics ( $p =$   
295  $0.99999999998777$ ), and highly significantly depleted for failed therapeutics ( $p = 3.87720403737775e - 196$ ).  
296 This suggests that our multi-scale evidence-based prioritisation pipeline is capable of selectively identifying  
297 genes that are likely to be effective therapeutic targets.

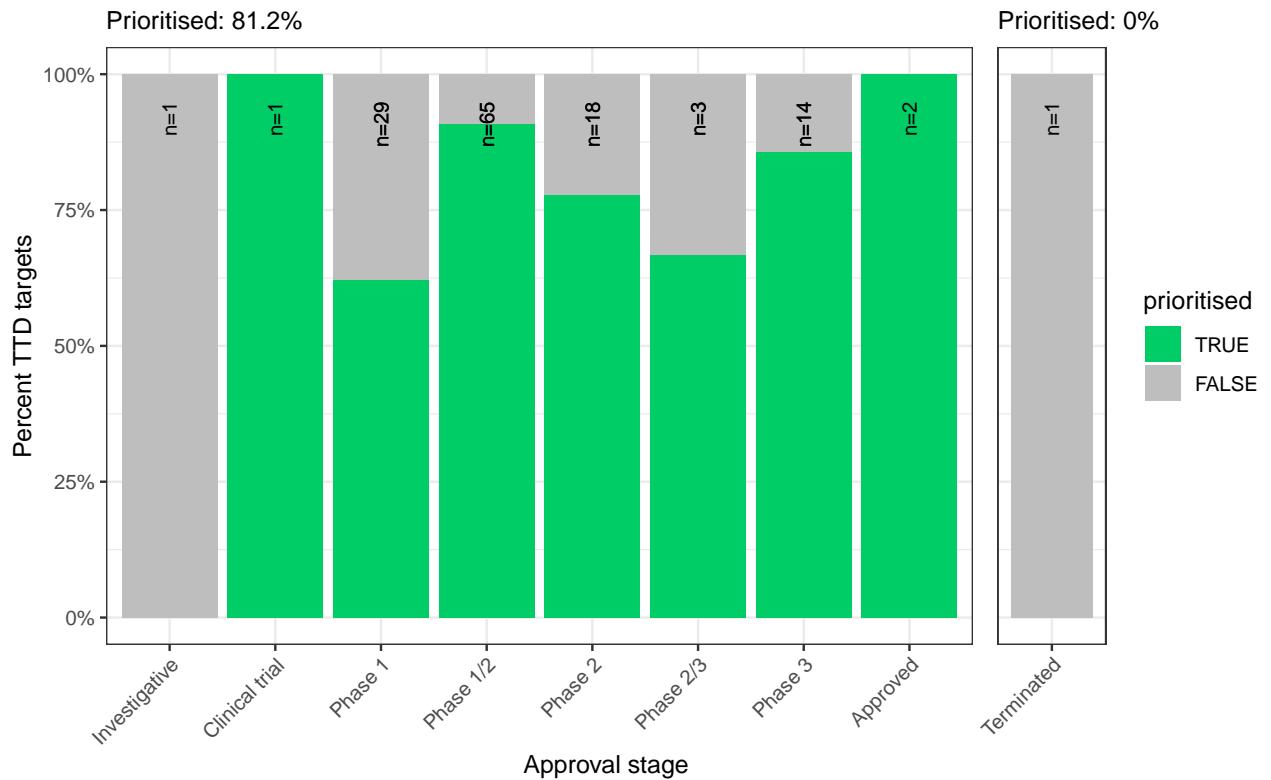


Figure 6: Validation of prioritised therapeutic targets. The proportion of existing gene therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline. Therapeutics are stratified by the stage of clinical development they were at during the time of writing.

## 298 Selected example targets

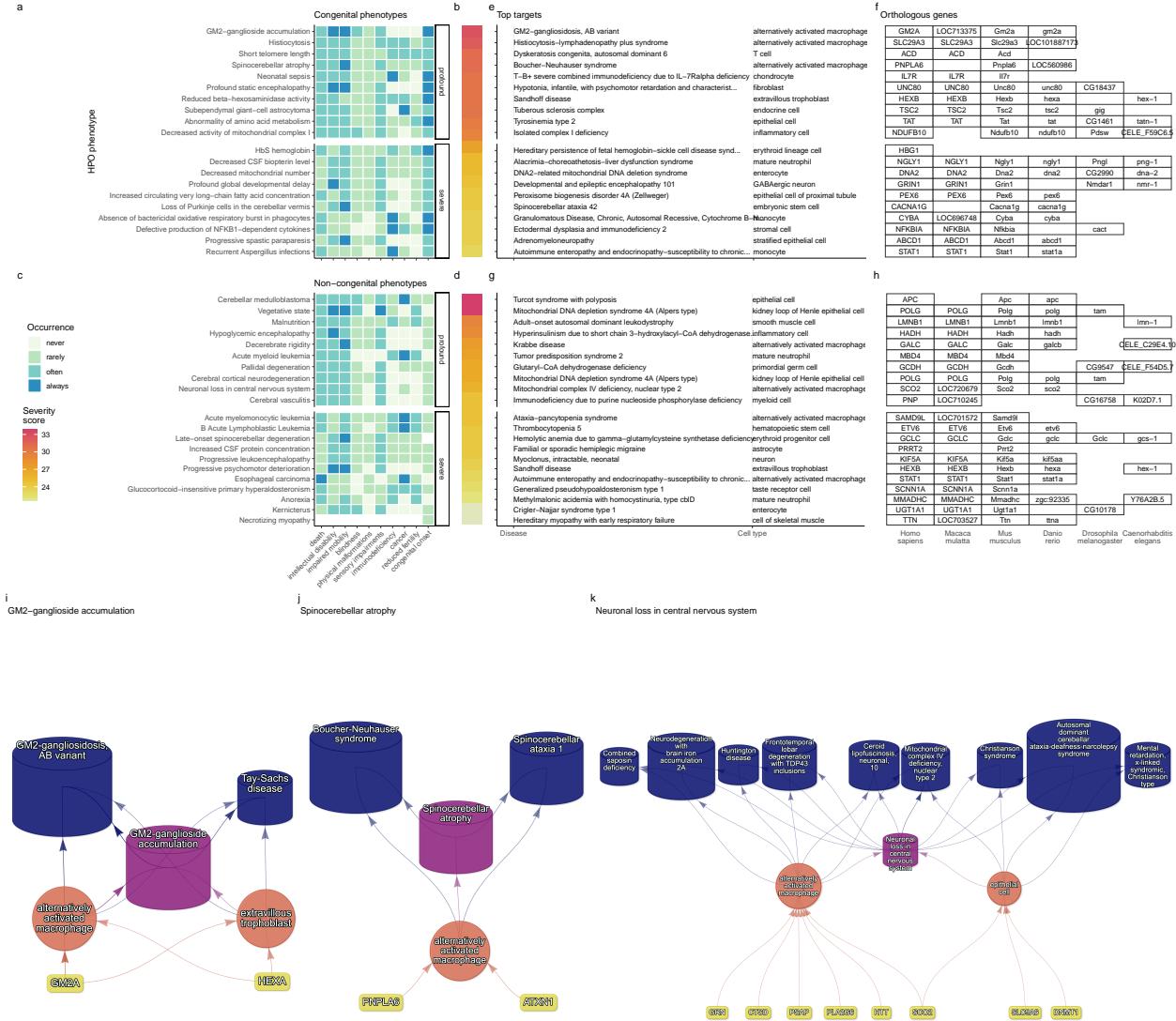


Figure 7: Top 40 prioritised gene therapy targets at multiple biological scales, stratified by congenital (top row) vs. non-congenital phenotypes (bottom row) as well as severity class (“profound” or “severe”). In this plot, only the top 10 most severe phenotypes within a given strata/substrata are shown **a,c**, Severity annotation generated by GPT-4. **b,d**, Composite severity scores computed across all severity metrics. **e,g**, Top mediator disease and cell type-specific target for each phenotype. **f,h** top target gene for each phenotype within humans (*Homo sapiens*). We also include the 1:1 ortholog of each human gene in several commonly used animal models, including monkey (*Macaca mulatta*), mouse (*Mus musculus*), zebrafish (*Danio rerio*), fly (*Drosophila melanogaster*) and nematode (*Caenorhabditis elegans*). Boxes are empty where no 1:1 ortholog is known. **i-k** 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 HPO<sup>11</sup>. Phenotypes are connected to cell types (red circles) via association testing between weighted gene sets (FDR<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 algorithm<sup>50</sup>.

299 From our prioritised targets, we selected the following four sets of phenotypes or diseases as examples:  
300 ‘GM2-ganglioside accumulation’, ‘Spinocerebellar atrophy’, ‘Neuronal loss in central nervous system’. Only  
301 phenotypes with a GPT severity score greater than 15 were considered to avoid overplotting and to focus on  
302 the more clinically relevant phenotypes.

303 Tay-Sachs disease (TSD) is a devastating hereditary condition in which children are born appearing healthy,  
304 which gradually degrades leading to death after 3-5 years. The underlying cause is the toxic accumulation  
305 of gangliosides in the nervous system due to a loss of the enzyme produced by *HEXA*. While this could  
306 in theory be corrected with gene editing technologies, there remain some outstanding challenges. One of  
307 which is identifying which cell types should be targeted to ensure the most effective treatments. Here  
308 we identified alternatively activated macrophages as the cell type most strongly associated with ‘GM2-  
309 ganglioside accumulation’. The role of aberrant macrophage activity in the regulation of ganglioside levels is  
310 supported by observation that gangliosides accumulate within macrophages in TSD<sup>52</sup>, as well as experimental  
311 evidence in rodent models<sup>53,54,55</sup>. Our results not only corroborate these findings, but propose macrophages  
312 as the primary causal cell type in TSD, making it the most promising cell type to target in therapies.

313 Another challenge in TSD is early detection and diagnosis, before irreversible damage has occurred. Our  
314 pipeline implicated extravillous trophoblasts of the placenta in ‘GM2-ganglioside accumulation’. While not  
315 necessarily a target for gene therapy, checking these cells *in utero* for an absence of *HEXA* may serve as  
316 a viable biomarker as these cells normally express the gene at high levels. Early detection of TSD may  
317 lengthen the window of opportunity for therapeutic intervention<sup>56</sup>, especially when genetic sequencing is not  
318 available or variants of unknown significance are found within *HEXA*<sup>57</sup>.

319 Spinocerebellar atrophy is a debilitating and lethal phenotype that occurs in diseases such as Spinocerebellar  
320 ataxia and Boucher-Nenhauser syndrome. These diseases are characterised by progressive degeneration of  
321 the cerebellum and spinal cord, leading to severe motor and cognitive impairments. Our pipeline identi-  
322 fied M2 macrophages as the only causal cell type associated with ‘Spinocerebellar atrophy’. This strongly  
323 suggests that degeneration of cerebellar Purkinje cells are in fact downstream consequences of macrophage  
324 dysfunction, rather than being the primary cause themselves. This is consistent with the known role of  
325 macrophages, especially microglia, in neuroinflammation and other neurodegenerative conditions such as  
326 Alzheimer’s and Parkinsons’ disease<sup>58–60</sup>. While experimental and postmortem observational studies have  
327 implicated microglia in spinocerebellar atrophy previously [<sup>58</sup>], our results provide a statistically-supported  
328 and unbiased genetic link between known risk genes and this cell type. Therefore, targeting M2 microglia in  
329 the treatment of spinocerebellar atrophy may therefore represent a promising therapeutic strategy. This is  
330 aided by the fact that there are mouse models that perturb the ortholog of human spinocerebellar atrophy  
331 risk genes (e.g. *Atxn1*, *Pnpla6*) and reliably recapitulate the effects of this diseases at the cellular (e.g. loss  
332 of Purkinje cells), morphological (e.g. atrophy of the cerebellum, spinal cord, and muscles), and functional  
333 (e.g. ataxia) levels.

334 Next, we investigated the phenotype ‘Neuronal loss in the central nervous system’. Despite the fact that this  
335 is a fairly broad phenotype, we found that it was only significantly associated with 3 cell types (alternatively  
336 activated macrophage, macrophage, epithelial cell), specifically M2 macrophages and sinusoidal endothelial  
337 cells.

338 Skeletal dysplasia is a heterogeneous group of over 450 disorders that affect the growth and development of  
339 bone and cartilage. This phenotype can be lethal when deficient bone growth leads to the constriction of  
340 vital organs such as the lungs. Even after surgical interventions, these complications continue to arise as the  
341 child develops. Pharmacological interventions to treat this condition have largely been ineffective. While  
342 there are various cell types involved in skeletal system development, our pipeline nominated chondrocytes  
343 as the causal cell type underlying the lethal form of this condition (Fig. 22). Assuringly, we found that  
344 the disease ‘Achondrogenesis Type 1B’ is caused by the genes *SLC26A2* and *COL2A1* via chondrocytes.  
345 We also found that ‘Platyspondylic lethal skeletal dysplasia, Torrance type’. Thus, in cases where surgical  
346 intervention is insufficient, targeting these genes within chondrocytes may prove a viable long-term solution  
347 for children suffering from lethal skeletal dysplasia.

348 Alzheimer’s disease (AD) is the most common neurodegenerative condition. It is characterised by a set of  
349 variably penetrant phenotypes including memory loss, cognitive decline, and cerebral proteinopathy. Inter-  
350 estingly, we found that different forms of early onset AD (which are defined by the presence of a specific  
351 disease gene) are each associated with different cell types via different phenotypes (Fig. 22). For example,  
352 AD 3 and AD 4 are primarily associated with cells of the digestive system (‘enterocyte’, ‘gastric goblet  
353 cell’) and are implied to be responsible for the phenotypes ‘Senile plaques’, ‘Alzheimer disease’, ‘Parietal  
354 hypometabolism in FDG PET’. Meanwhile, AD 2 is primarily associated with immune cells (‘alternatively  
355 activated macrophage’) and is implied to be responsible for the phenotypes ‘Neurofibrillary tangles’, ‘Long-  
356 tract signs’. This suggests that different forms of AD may be driven by different cell types and phenotypes,  
357 which may help to explain its variability in onset and clinical presentation.

358 Finally, Parkinson’s disease (PD) is characterised by motor symptoms such as tremor, rigidity, and bradyki-  
359 nesia. However there are a number of additional phenotypes associated with the disease that span multiple  
360 physiological systems. PD 19a and PD 8 seemed to align most closely with the canonical understanding of  
361 PD as a disease of the central nervous system in that they implicated oligodendrocytes and neurons (Fig. 22).  
362 Though the reference datasets being used in this study were not annotated at sufficient resolution to distin-  
363 guish between different subtypes of neurons, in particular dopaminergic neurons. PD 19a/8 also suggested  
364 that risk variants in *LRRK2* mediate their effects on PD through both myeloid cells and oligodendrocytes  
365 by causing gliosis of the substantia nigra. The remaining clusters of PD mechanisms revolved around chon-  
366 drocytes (PD 20), amacrine cells of the eye (hereditary late-onset PD), and the respiratory/immune system  
367 (PD 14). While the diversity in cell type-specific mechanisms is somewhat surprising, it may help to explain  
368 the wide variety of cross-system phenotypes frequently observed in PD.

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

373 **Experimental model translatability**

374 We computed interspecies translatability scores using a combination of both ontological ( $SIM_o$ ) and geno-  
375 typic ( $SIM_g$ ) similarity relative to each homologous human phenotype and its associated genes Fig. 15.  
376 In total, we mapped 278 non-human phenotypes (in *Caenorhabditis elegans*, *Danio rerio*, *Mus musculus*,  
377 *Rattus norvegicus*) to 849 homologous human phenotypes. Amongst the 5252 phenotype within our priori-  
378 tised therapy targets, 354 had viable animal models in at least one non-human species. Per species, the  
379 number of homologous phenotypes was: *Danio rerio* (n=214) *Mus musculus* (n=150) *Caenorhabditis elegans*  
380 (n=35) *Rattus norvegicus* (n=3). Amongst our prioritised targets with a GPT-4 severity score of >10, the  
381 phenotypes with the greatest animal model similarity were ‘Anterior vertebral fusion’ ( $SIM_{o,g} = 0.967$ ),  
382 ‘Disc-like vertebral bodies’ ( $SIM_{o,g} = 0.964$ ), ‘Metaphyseal enchondromatosis’ ( $SIM_{o,g} = 0.946$ ), ‘Periph-  
383 eral retinal avascularization’ ( $SIM_{o,g} = 0.943$ ), ‘Retinal vascular malformation’ ( $SIM_{o,g} = 0.943$ ).

384 **Discussion**

385 Across the 201 cell types and 11,047 RD-associated phenotypes investigated, more than 46,514 significant  
386 phenotype-cell type relationships were discovered. This presents a wealth of opportunities to trace the  
387 mechanisms of rare diseases through multiple biological scales. This in turn enhances our ability to study  
388 and treat causal factors in disease with deeper understanding and greater precision. These results recapitulate  
389 well-known relationships, while providing additional cellular context to many of these known relationships,  
390 and discovering novel relationships.

391 From our target prioritisation pipeline results, we highlight cell type-specific mechanisms for ‘GM2-  
392 ganglioside accumulation’ in Tay-Sachs disease, spinocerebellar atrophy in spinocerebellar ataxia, and  
393 ‘Neuronal loss in central nervous system’ in a variety of diseases (Fig. 7). Of interest, all three of these  
394 neurodegenerative phenotypes involved alternatively activated (M2) macrophages. The role of macrophages  
395 in neurodegeneration is complex, with both neuroprotective and neurotoxic functions, including the  
396 clearance of misfolded proteins, the regulation of the blood-brain barrier, and the modulation of the immune  
397 response<sup>61</sup>. We also recapitulated prior evidence that microglia, the resident macrophages of the nervous  
398 system, are causally implicated in Alzheimer’s disease (AD) (Fig. 22)<sup>62</sup>. An important contribution of our  
399 current study is that we were able to pinpoint the specific phenotypes of AD caused by macrophages to  
400 neurofibrillary tangles and long-tract signs (reflexes that indicate the functioning of spinal long fiber tracts).  
401 Other AD-associated phenotypes were caused by other cell types (e.g. gastric goblet cells, enterocytes).

402 Investigating RDs at the level of phenotypes offers several key advantages. First, the vast majority of  
403 RDs only have one associated gene (7671/8631 diseases = 89%). Aggregating gene sets across diseases into  
404 phenotype-centric “buckets” permits sufficiently well-powered analyses, with an average of ~76 genes per  
405 phenotype (median=7) see Fig. 9. Second, we hypothesise that these phenotype-level gene sets converge  
406 on a limited number of molecular and cellular pathways. Perturbations to these pathways manifest as one  
407 or more phenotypes which, when considered together, tend to be clinically diagnosed as a certain disease.  
408 Third, RDs are often highly heterogeneous in their clinical presentation across individuals, leading to the  
409 creation of an ever increasing number of disease subtypes (some of which only have a single documented case).  
410 In contrast, a phenotype-centric approach enables us to more accurately describe a particular individual’s  
411 version of a disease without relying on the generation of additional disease subcategories. By characterising  
412 an individual’s precise phenotypes over time, we may better understand the underlying biological mecha-  
413 nisms that have caused their condition. However, in order to achieve a truly precision-based approach to  
414 clinical care, we must first characterise the molecular and cellular mechanisms that cause the emergence  
415 of each phenotype. Here, we provide a highly reproducible framework that enables this at the scale of the  
416 entire genome. This presents an opportunity to design basket trials of patients with different diseases  
417 but overlapping phenotypes and cellular mechanisms<sup>63</sup>. It may be especially helpful for complex patients  
418 with diagnostically ambiguous sets of phenotypes who would otherwise be excluded from traditional clinical  
419 trials<sup>64</sup>.

420 It was paramount to the success of this study to ensure our results were anchored in ground-truth bench-  
421 marks, generated falsifiable hypotheses, and rigorously guarded against false-positive associations. Extensive  
422 validation using multiple approaches demonstrated that our methodology consistently recapitulates expected  
423 phenotype-cell type associations (Fig. 1-Fig. 5). This was made possible by the existence of comprehensive,  
424 structured ontologies for all phenotypes (HPO) and cell types (CL), which provide an abundance of clear and  
425 falsifiable hypotheses for which to test our predictions against. Several key examples include 1) strong en-  
426 richment of associations between cell types and phenotypes within the same anatomical systems (Fig. 1b-d),  
427 2) a strong relationship between phenotype-specificity and the strength and number of cell type associations  
428 (Fig. 2), 3) identification of the precise cell subtypes involved in susceptibility to various subtypes of recurrent  
429 bacterial infections (Fig. 3), 4) a strong positive correlation between the frequency of congenital onset of  
430 a phenotype and the proportion of developmental cell types associated with it (Fig. 5)), and 5) consistent  
431 phenotype-cell type associations across multiple independent single-cell datasets (Fig. 10). Having validated  
432 our phenotype-cell type associations, we then went on to demonstrate how these results may be used in  
433 therapeutics development (Fig. 7).

434 Diagnosis is an essential but challenging step in RD patient care. Additional phenotypes that emerge over  
435 time may assist a clinician to reach a more confident disease diagnosis. However many of these phenotypes  
436 can have a serious impact on patient quality of life or survival and avoiding them would be far better for

437 patient outcomes. Often times phenotypes alone cannot clearly pinpoint the disease and thus a diagnosis is  
438 never reached. Having a more complete understanding of the mechanisms underlying observed phenotypes  
439 allows clinicians to far more effectively make predictions about what additional, less obvious phenotypes they  
440 should search for to confirm or reject their hypothesis of disease diagnosis (e.g. with imaging or biomarker  
441 tests).

442 Unfortunately, there are currently only treatments available for less than 5% of RDs<sup>6</sup>. Novel technologies  
443 including CRISPR, prime editing, antisense oligonucleotides, viral vectors, and/or lipid nanoparticles, have  
444 been undergone significant advances in the last several years<sup>65–69</sup> and proven remarkable clinical success in  
445 an increasing number of clinical applications<sup>70–73</sup>. The U.S. Food and Drug Administration (FDA) recently  
446 announced an landmark program aimed towards improving the international regulatory framework to take  
447 advantage of the evolving gene/cell therapy technologies<sup>74</sup> with the aim of bringing dozens more therapies to  
448 patients in a substantially shorter timeframe than traditional pharmaceutical product development (typically  
449 5–20 years with a median of 8.3 years)<sup>75</sup>. While these technologies have the potential to revolutionise RD  
450 medicine, their successful application is dependent on first understanding the mechanisms causing each  
451 disease.

452 To address this critical gap in knowledge, we used our results to create a reproducible and customisable  
453 pipeline to nominate cell type-resolved therapeutic targets (Fig. 13–Fig. 7). Targeting cell type-specific  
454 mechanisms underlying granular RD phenotypes can improve therapeutic effectiveness by treating the causal  
455 root of an individual’s conditions<sup>66,76</sup>. A cell type-specific approach also helps to reduce the number of  
456 harmful side effects caused by unintentionally delivering the therapeutic to off-target tissues/cell types (which  
457 may induce aberrant gene activity), especially when combined with technologies that can target cell surface  
458 antigens (e.g viral vectors)<sup>77</sup>. This has the additional benefit of reducing the minimal effective dose of a  
459 therapeutic, which can be both immunogenic and extremely financially costly<sup>9,10,65,68</sup>. Here, we demonstrate  
460 the utility of a high-throughput evidence-based approach to RD therapeutics discovery by highlighting several  
461 of the most promising therapeutic candidates. Our pipeline takes into account a myriad of factors, including  
462 the strength of the phenotype-cell type associations, symptom-cell type associations, cell type-specificity  
463 of causal genes, the severity and frequency of the phenotypes, suitability for gene therapy delivery systems  
464 (e.g. recombinant adeno-associated viral vectors (rAAV)), as well as a quantitative analysis of phenotypic and  
465 genetic animal model translatability (Fig. 15). We validated these candidates by comparing the proportional  
466 overlap with gene therapies that are presently in the market or undergoing clinical trials, in which we  
467 recovered 81% of all active gene therapies and 0% of failed gene therapies (Fig. 6, Fig. 14). Despite nominating  
468 a large number of putative targets, hypergeometric tests confirmed that our targets were strongly enriched  
469 for targets of existing therapies that are either approved or currently undergoing clinical trials.

470 It should be noted that our study has several key limitations. First, while our cell type datasets are amongst  
471 the most comprehensive human scRNA-seq references currently available, they are nevertheless missing

472 certain tissues, cell types (e.g. spermatocytes, oocytes), and life stages (post-natal childhood, senility). It is  
473 also possible that we have not captured certain cell state signatures that only occur in disease (e.g. disease-  
474 associated microglia<sup>78,79</sup>). Though we reasoned that using only control cell type signatures would mitigate  
475 bias towards any particular disease, and avoid degradation of gene signatures due to loss of function mutations.  
476 Second, the collective knowledge of gene-phenotype and gene-disease associations is far from complete and  
477 we fully anticipate that these annotations will continue to expand and change well into the future. It is  
478 for this reason we designed this study to be easily reproduced within a single containerised script so that  
479 we (or others) may rerun it with updated datasets at any point. Finally, causality is notoriously difficult  
480 to prove definitively from associative testing alone, and our study is not exempt from this rule. Despite  
481 this, there are several reasons to believe that our approach is able to better approximate causal relationships  
482 than traditional approaches. First, we did not intentionally preselect any subset of phenotypes or cell types  
483 to investigate here. Along with a scaling prestep during linear modelling, this means that all the results  
484 are internally consistent and can be directly compared to one another (in stark contrast to literature meta-  
485 analyses). Furthermore, for the phenotype gene signatures we used expert-curated GenCC annotations<sup>80,81</sup>  
486 to weight the current strength of evidence supporting a causal relationship between each gene and phenotype.  
487 This is especially important for phenotypes with large genes lists (thousands of annotations) for which some  
488 of the relationships may be tenuous. Within the cell type references, we deliberately chose to use specificity  
489 scores (rather than raw gene expression) as this normalisation procedure has previously been demonstrated  
490 to better distinguish between signatures of highly similar cell types/subtypes<sup>82</sup>.

491 Common ontology-controlled frameworks like the HPO open a wealth of new opportunities, especially when  
492 addressing RDs. Services such as the Matchmaker Exchange<sup>83,84</sup> have enabled the discovery of hundreds of  
493 underlying genetic etiologies, and led to the diagnosis of many patients. This also opens the possibility of  
494 gathering cohorts of geographically dispersed patients to run clinical trials, the only viable option for treat-  
495 ment in many individuals. To further increase the number of individuals who qualify for these treatments,  
496 as well as the trial sample size, proposals have been made deviate from the traditional single-disease clinical  
497 trial model and instead perform basket trials on groups of RDs with shared molecular etiologies (SaME)<sup>63</sup>.

498 Moving forward, we are now actively seeking industry and academic partnerships to begin experimentally  
499 validating our multi-scale target predictions and exploring their potential for therapeutic translation. Never-  
500 theless, there are more promising therapeutic targets here than our research group could ever hope to pursue  
501 by ourselves. In the interest of accelerating research and ensuring RD patients are able to benefit from this  
502 work as quickly as possible, we have decided to publicly release all of the results described in this study.  
503 These can be accessed in multiple ways, including through a suite of R packages as well as a web app, the  
504 [Rare Disease Celltyping Portal](#). The latter allows our results to be easily queried, filtered, visualised, and  
505 downloaded without any knowledge of programming. Through these resources we aim to make our findings  
506 useful to a wide variety of RD stakeholders including subdomain experts, clinicians, advocacy groups, and

507 patients.

## 508 Conclusions

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

## 517 Methods

### 518 Human Phenotype Ontology

519 The latest version of the HPO (release releases) was downloaded from the EMBL-EBI Ontology Lookup  
520 Service<sup>85</sup> and imported into R using the `HPOExplorer` package. This R object was used to extract ontolog-  
521 ical relationships between phenotypes as well as to assign absolute and relative ontological levels to each  
522 phenotype. The latest version of the HPO phenotype-to-gene mappings and phenotype annotations were  
523 downloaded from the official HPO GitHub repository and imported into R using `HPOExplorer`. This contains  
524 lists of genes associated with phenotypes via particular diseases, formatted as three columns in a table (gene,  
525 phenotype, disease).

526 However, not all genes have equally strong evidence of causality with a disease or phenotype, especially when  
527 considering that the variety of resources used to generate these annotations (OMIM, Orphanet, DECIPHER)  
528 use variable methodologies (e.g. expert-curated review of the medical literature vs. automated text mining  
529 of the literature). Therefore we imported data from the Gene Curation Coalition (GenCC)<sup>80,81</sup>, which (as  
530 of 2024-05-17) 22060 evidence scores across 7259 diseases and 5165 genes. Evidence scores are defined by  
531 GenCC using a standardised ordinal rubric which we then encoded as a semi-quantitative score ranging  
532 from 0 (no evidence of disease-gene relationship) to 6 (strongest evidence of disease-gene relationship) (see  
533 Table 4). As each Disease-Gene pair can have multiple entries (from different studies) with different levels  
534 of evidence, we then summed evidence scores per Disease-Gene pair to generate aggregated Disease-by-Gene  
535 evidence scores. This procedure can be described as follows.

536 Let us denote:

- 537 •  $D$  as diseases.  
538 •  $P$  as phenotypes in the HPO.

- 539 •  $G$  as genes
- 540 •  $S$  as the evidence scores describing the strength of the relationship between each Disease-Gene pair.
- 541 •  $M_{ij}$  as the aggregated Disease-by-Gene evidence score matrix.

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

542 Next, we extracted Disease-Gene-Phenotype relationships from the annotations file distributed by the HPO  
 543 (*phenotype\_to\_genes.txt*). This provides a list of genes associated with phenotypes via particular diseases,  
 544 but does not include any strength of evidence scores.

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

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

547 In order to assign evidence scores to each Phenotype-Gene relationship, we combined the aforementioned  
 548 datasets from GenCC ( $M_{ij}$ ) and HPO ( $A_{ijk}$ ) by merging on the gene and disease ID columns. For each  
 549 phenotype, we then computed the mean of Disease-Gene scores across all diseases for which that phenotype  
 550 is a symptom. This resulted in a final 2D tensor of Phenotype-by-Gene evidence scores ( $L_{ij}$ ):

551

552

553  $L_{ij} = \begin{cases} \frac{\sum_{k=1}^f D_i G_j P_k}{f}, & \text{if } D_i G_j \in A, \\ 1, & \text{if } D_i G_j \notin A \end{cases}$

554

555

556

Tensor of Disease-by-Gene  
evidence scores

Tensor of Phenotype-by-Gene  
evidence scores

Disease-by-Gene-by-Phenotype  
relationships

557 Construction of the tensor of Phenotype-by-Gene evidence scores.

558  
559

560 Histograms of evidence score distributions at each step in processing can be found in Fig. 8.

561 **Single-cell transcriptomic atlases**

562 In this study, the gene by cell type specificity matrix was constructed using the Descartes Human transcriptome atlas of foetal gene expression, which contains a mixture of single-nucleus and single-cell RNA-seq data (collected with sci-RNA-seq3)<sup>28</sup>. This dataset contains 377,456 cells representing 77 distinct cell types across 15 tissues. All 121 human foetal samples ranged from 72 to 129 days in estimated postconceptual age.

563 To independently replicate our findings, we also used the Human Cell Landscape which contains single-cell transcriptomic data (collected with microwell-seq) from embryonic, foetal, and adult human samples across

564 49 tissues<sup>29</sup>.

565 Specificity matrices were generated separately for each transcriptomic atlas using the R package EWCE (v1.11.3)<sup>82</sup>. Within each atlas, cell types were defined using the authors' original freeform annotations in order to preserve the granularity of cell subtypes as well as incorporate expert-identified rare cell types.

566 Cell types were only aligned and aggregated to the level of corresponding Cell Ontology (CL)<sup>34</sup> 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  $F(g, i, c)$ .

567 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:

582

583

**Compute mean expression of each gene per cell type**

---

Gene-by-cell type specificity matrix

$$S_{gc} = \frac{\sum_{i=1}^{|L|} F_{gic}}{\sum_{r=1}^k \left( \frac{\sum_{i=1}^{|L|} F_{gic}}{N_c} \right)}$$

Compute row sums of  
mean gene-by-cell type matrix

---

584

585

586

587 **Phenotype-cell type associations**

588 To test for relationships between each pairwise combination of phenotype (n=11,047) and cell type (n=201)

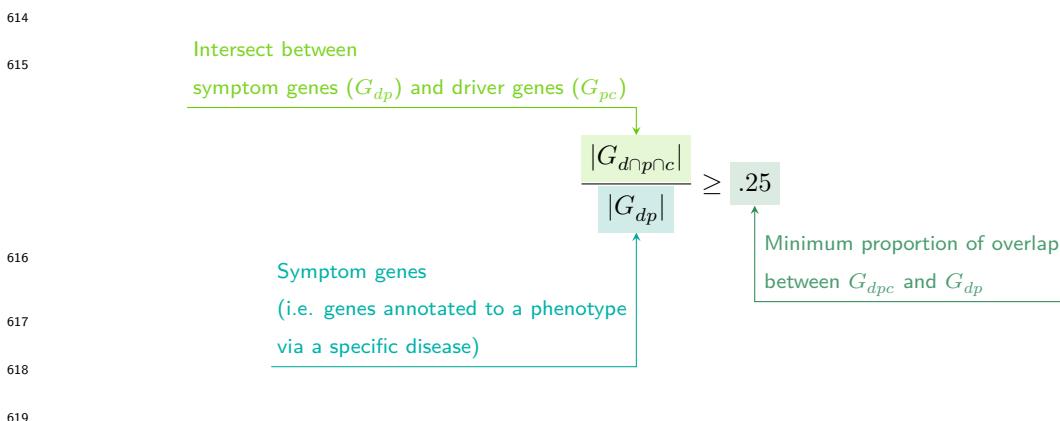
589 we ran a series of univariate generalised linear models implemented via the `stats::glm` function in R. First,

590 we filtered the gene-by-phenotype evidence score matrix ( $L_{ij}$ ) and the gene-by-cell type expression specificity  
 591 matrix ( $S_{gc}$ ) to only include genes present in both matrices (n=4,949 genes in the Descartes Human analyses;  
 592 n=4,653 genes in the Human Cell Landscape analyses). Then, within each matrix any rows or columns with a  
 593 sum of 0 were removed as these were uninformative data points that did not vary. To improve interpretability  
 594 of the results  $\beta$  coefficient estimates across models (i.e. effect size), we performed a scaling prestep on all  
 595 dependent and independent variables. Initial tests showed that this had virtually no impact on the total  
 596 number of significant results or any of the benchmarking metrics based on p-value thresholds Fig. 1. This  
 597 scaling prestep improved our ability to rank cell types by the strength of their association with a given  
 598 phenotype as determined by separate linear models.

599 We repeated the aforementioned procedure separately for each of the single-cell references. Once all results  
 600 were generated using both cell type references (2,206,994 association tests total), we applied Benjamini-  
 601 Hochberg false discovery rate<sup>86</sup> (denoted as  $FDR_{pc}$ ) to account for multiple testing. Of note, we applied  
 602 this correction across all results at once (as opposed to each single-cell reference separately) to ensure the  
 603  $FDR_{pc}$  was stringently controlled for across all tests performed in this study.

#### 604 Symptom-cell type associations

605 Here we define a symptom as a phenotype as it presents within the context of the specific disease. The features  
 606 of a given symptom can be described as the subset of genes annotated to phenotype  $p$  via a particular disease  
 607  $d$ , denoted as  $G_{dp}$  (see Fig. 9). To attribute our phenotype-level cell type enrichment signatures to specific  
 608 diseases, we first identified the gene subset that was most strongly driving the phenotype-cell type association  
 609 by computing the intersect of genes that were both in the phenotype annotation and within the top 25%  
 610 specificity percentile for the associated cell type. We then computed the intersect between symptom genes  
 611 ( $G_{dp}$ ) 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 with 25% or greater  
 612 overlap with the symptom gene subset ( $G_{dp}$ ) were kept. This procedure was repeated for all phenotype-cell  
 613 type-disease triads, which can be summarised as follows:



620 **Validation of expected phenotype-cell type relationships**

621 We first sought to confirm that our tests (across both single-cell references) were able to recover expected  
622 phenotype-cell type relationships across seven high-level branches within the HPO (Fig. 1), including ab-  
623 normalities of the cardiovascular system, endocrine system, eye, immune system, musculoskeletal system,  
624 nervous system, and respiratory system. Within each branch the number of significant tests in a given  
625 cell type were plotted (Fig. 1b). Mappings between freeform annotations (the level at which we performed  
626 our phenotype- cell type association tests) provided by the original atlas authors and their closest CL term  
627 equivalents were provided by CellxGene<sup>26</sup>. CL terms along the *x-axis* of Fig. 1b were assigned colours corre-  
628 sponding to which HPO branch showed the greatest number of enrichments (after normalising within each  
629 branch to account for differences in scale). The normalised colouring allows readers to quickly assess which  
630 HPO branch was most often associated with each cell type, while accounting for differences in the number  
631 of phenotypes across branches. We then ran a series of Analysis of Variance (ANOVA) tests to determine  
632 whether (within a given branch) a given cell type was more often enriched ( $FDR < 0.05$ ) within that branch  
633 relative to all of the other HPO branches of an equivalent level in the ontology (including all branches not  
634 shown in Fig. 1b). After applying Benjamini-Hochberg multiple testing correction<sup>86</sup> (denoted as  $FDR_{b,c}$ ),  
635 we annotated each respective branch-by-cell type bar according to the significance (\*\*\*\* :  $FDR_{b,c} < 1e-04$ ,  
636 \*\*\* :  $FDR_{b,c} < 0.001$ , \*\* :  $FDR_{b,c} < 0.01$ , \* :  $FDR_{b,c} < 0.05$ ). Cell types in Fig. 1a-b were ordered along  
637 the *x-axis* according to a dendrogram derived from the CL ontology (Fig. 1c), which provides ground-truth  
638 semantic relationships between all cell types (e.g. different neuronal subtypes are grouped together).

639 As an additional measure of the accuracy of our phenotype-cell types test results we identified conceptually  
640 matched branches across the HPO and the CL (Fig. 1d and Table 5). For example, ‘Abnormality of the  
641 cardiovascular system’ in the HPO was matched with ‘cardiocytes’ in the CL which includes all cell types  
642 specific to the heart. Analogously, ‘Abnormality of the nervous system’ in the HPO was matched with ‘neural  
643 cell’ in the CL which includes all descendant subtypes of neurons and glia. This cross-ontology matching  
644 was repeated for each HPO branch and can be referred to as on-target cell types. Within each branch, the  
645  $-\log_{10}(FDR_{pc})$  values of on-target cell types were binned by rounding to the nearest integer (*x-axis*) and  
646 the percentage of tests for on-target cell types relative to all cell types were computed at each bin (*y-axis*)  
647 (Fig. 1d). The baseline level (dotted horizontal line) illustrates the percentage of on-target cell types relative  
648 to the total number of observed cell types. Any percentages above this baseline level represent greater than  
649 chance representation of the on-target cell types in the significant tests.

650 **Monarch Knowledge Graph recall**

651 Finally, we gathered known phenotype-cell type relationships from the Monarch Knowledge Graph (MKG),  
652 a comprehensive database of links between many aspects of disease biology<sup>87</sup>. This currently includes 103  
653 links between HPO phenotypes (n=103) and CL cell types (n=79). Of these, we only considered the 82

654 phenotypes that we were able to test given that our ability to generate associations was dependent on  
655 the existence of gene annotations within the HPO. We considered instances where we found a significant  
656 relationship between exactly matching pairs of HPO-CL terms as a hit.

657 However, as the cell types in MKG were not necessarily annotated at the same level as our single-cell refer-  
658 ences, we considered instances where the MKG cell type was an ancestor term of our cell type (e.g. ‘myeloid  
659 cell’ vs. ‘monocyte’), or *vice versa*, as hits. We also adjusted ontological distance by computing the ratio  
660 between the observed ontological distance and the smallest possible ontological distance for that cell type  
661 given the cell type that were available in our references ( $dist_{adjusted} = (\frac{dist_{observed}+1}{dist_{minimum}+1}) - 1$ ). This provides  
662 a way of accurately measuring how dissimilar our identified cell types were for each phenotype-cell type  
663 association (Fig. 11).

#### 664 Annotation of phenotypes using generative large language models

665 Only a small fraction of the the phenotypes in HPO (<1%) have metadata annotations containing informa-  
666 tion on their time course, consequences, and severity. This is due to the time-consuming nature of manually  
667 annotating thousands of phenotypes. To generate such annotations at scale, we previously used Generative  
668 Pre-trained Transformer 4 (GPT-4), a large language model (LLM) as implemented within OpenAI’s Appli-  
669 cation Programming Interface (API)<sup>32</sup>. After extensive prompt engineering and ground-truth benchmarking,  
670 we were able to acquire annotations on how often each phenotype directly causes intellectual disability, death,  
671 impaired mobility, physical malformations, blindness, sensory impairments, immunodeficiency, cancer, re-  
672 duced fertility, or is associated with a congenital onset. These criteria were previously defined in surveys  
673 of medical experts as a means of systematically assessing phenotype severity<sup>88</sup>. Responses for each metric  
674 were provided in a consistent one-word format which could be one of: ‘never’, ‘rarely’, ‘often’, ‘always’. This  
675 procedure was repeated in batches (to avoid exceeding token limits) until annotations were gathered for  
676 16982/18082 HPO phenotypes.

677 We then encoded these responses into a semi-quantitative scoring system (‘never’=0, ‘rarely’=1, ‘often’=2,  
678 ‘always’=3), which were then weighted by multiplying a semi-subjective scoring of the relevance of  
679 each metric to the concept of severity on a scale from 1-6, with 6 being the most severe (‘death’=6,  
680 ‘intellectual\_disability’=5, ‘impaired\_mobility’=4, ‘physical\_malformations’=3, ‘blindness’=4, ‘sen-  
681 sory\_impairments’=3, ‘immunodeficiency’=3, ‘cancer’=3, ‘reduced\_fertility’=1, ‘congenital\_onset’=1).  
682 Finally, the product of the score was normalised to a quantitative severity score ranging from 0-100, where  
683 100 is the theoretical maximum severity score. This phenotype severity scoring procedure can be expressed  
684 as follows.

685 Let us denote:

- 686 •  $p$  : a phenotype in the HPO.

- 687 •  $j$  : the identity of a given annotation metric (i.e. clinical characteristic, such as ‘intellectual disability’  
 688 or ‘congenital onset’).
- 689 •  $W_j$ : the assigned weight of metric  $j$ .
- 690 •  $F_j$ : the maximum possible value for metric  $j$  (equivalent across all  $j$ ).
- 691 •  $F_{pj}$  : the numerically encoded value of annotation metric  $j$  for phenotype  $p$ .
- 692 •  $NSS_p$ : the final composite severity score for phenotype  $p$  after applying normalisation to align values  
 693 to a 0-100 scale and ensure equivalent meaning regardless of which other phenotypes are being analysed  
 694 in addition to  $p$ . This allows for direct comparability of severity scores across studies with different  
 695 sets of phenotypes.

696                                  Sum of weighted annotation values  
 697                                  across all metrics  
 698                                  Numerically encoded annotation value  
   of metric  $j$  for phenotype  $p$   
 699                                  Weight for metric  $j$   
 700                                  Theoretical maximum severity score  
 701                                  Normalised Severity Score  
   for each phenotype

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

## 702 Congenital phenotypes are associated with foetal cell types

703 The GPT-4 annotations also enabled us to assess whether foetal cell types were more often significantly  
 704 associated with congenital phenotypes in our Human Cell Landscape results as this single-cell reference  
 705 contained both adult and foetal versions of cell types (Fig. 5). To do this, we performed a chi-squared ( $\chi^2$ )  
 706 test on the proportion of significantly associated cell types containing any of the substrings ‘fetal’, ‘fetus’,  
 707 ‘primordial’, ‘hESC’ or ‘embryonic’ (within cell types annotations from the original Human Cell Landscape  
 708 authors<sup>29</sup>) vs. those associated without, stratified by how often the corresponding phenotype had a congenital  
 709 onset according to the GPT phenotype annotations (including ‘never’, ‘rarely’, ‘often’, ‘always’). In addition,  
 710 a series of  $\chi^2$  tests were performed within each congenital onset frequency strata, to determine whether the  
 711 observed proportion of foetal cell types vs. non-foetal cell types significantly deviated from the proportions  
 712 expected by chance.

713 We next tested whether the proportion of tests with significant associations with foetal cell types varied  
 714 across the major HPO branches using a  $\chi^2$  test. We also performed separate  $\chi^2$  test within each branch to  
 715 determine whether the proportion of significant associations with foetal cell types was significantly different

716 from chance.

## 717 Therapeutic target identification

718 We developed a systematic and automated strategy for identifying putative cell type-specific gene targets  
719 for each phenotype based on a series of filters at phenotype, cell type, and gene levels. The entire target  
720 prioritisation procedure can be replicated with a single function: `MSTExplorer::prioritise_targets`. This  
721 function automates all of the reference data gathering (e.g. phenotype metadata, cell type metadata, cell  
722 type signature reference, gene lengths, severity tiers) and takes a variety of arguments at each step for greater  
723 customisability. Each step is described in detail in Table 3. Phenotypes that often or always caused physical  
724 malformations (according to the GPT-4 annotations) were also removed from the final prioritised targets  
725 list, as these were unlikely to be amenable to gene therapy interventions. Finally, phenotypes were sorted  
726 by their composite severity scores such that the most severe phenotypes were ranked the highest.

## 727 Therapeutic target validation

728 To assess whether our prioritised therapeutic targets were likely to be viable, we computed the overlap  
729 between our gene targets and those of existing gene therapies at various stages of clinical development  
730 (Fig. 6). Gene targets were obtained for each therapy from the Therapeutic Target Database (TTD; release  
731 2024-07-03) and mapped onto standardised HUGO Gene Nomenclature Committee (HGNC) gene symbols  
732 using the `orthogene` R package. We stratified our overlap metrics according to whether the therapies had  
733 failed (unsuccessful clinical trials or withdrawn), or were non-failed (successful or ongoing clinical trials).  
734 We then conducted hypergeometric tests to determine whether the observed overlap between our prioritised  
735 targets and the non-failed therapy targets was significantly greater than expected by chance (i.e. enrichment).  
736 We also conducted a second hypergeometric test to determine whether the observed overlap between our  
737 prioritised targets and the failed therapy targets was significantly less than expected by chance (i.e. depletion).  
738 Finally, we repeated the analysis against all therapeutic targets, not just those of gene therapies, to determine  
739 whether our prioritised targets had relevance to other therapeutic modalities.

## 740 Experimental model translatability

741 To improve the likelihood of successful translation between preclinical animal models and human patients,  
742 we created an interspecies translatability prediction tool for each phenotype nominated by our gene therapy  
743 prioritised pipeline (Fig. 15). First, we extracted ontological similarity scores of homologous phenotypes  
744 across species from the MKG<sup>87</sup>. Briefly, the ontological similarity scores ( $SIM_o$ ) are computed for each  
745 homologous pair of phenotypes across two ontologies by calculating the overlap in homologous phenotypes  
746 that are ancestors or descendants of the target phenotype. Next, we generated genotypic similarity scores  
747 ( $SIM_g$ ) for each homologous phenotype pair by computing the proportion of 1:1 orthologous genes using  
748 gene annotation from their respective ontologies. Interspecies orthologs were also obtained from the MKG.

749 Finally, both scores are multiplied together to yield a unified ontological-genotypic similarity score ( $SIM_{o,g}$ ).

750 **Novel R packages**

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

759 **Rare Disease Celltyping Portal**

760 To further increase the ease of access for stakeholders in the RD community without the need for program-  
761 matic experience, we developed a series of web apps to interactively explore, visualise, and download the  
762 results from our study. Collectively, these web apps are called the Rare Disease Celltyping Portal. The  
763 landing page for the website was made using HTML, CSS, and javascript and the web apps were created  
764 using the Shiny Web application framework for R and deployed on the shinyapps.io server. The website  
765 can be accessed at [https://neurogenomics.github.io/rare\\_disease\\_celltyping\\_apps/home](https://neurogenomics.github.io/rare_disease_celltyping_apps/home). All code used to  
766 generate the website can be found at [https://github.com/neurogenomics/rare\\_disease\\_celltyping\\_apps](https://github.com/neurogenomics/rare_disease_celltyping_apps).

<sub>767</sub> **Tables**

Table 2: 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                 | 19,929         | 26,585             | 46,514    |
| tests                             | 848,078        | 1,358,916          | 2,206,994 |
| 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            | 7,340          | 9,049              | 9,575     |
| phenotypes tested                 | 11,014         | 10,959             | 11,028    |
| phenotypes                        | 11,047         | 11,047             | 11,047    |
| phenotypes significant (%)        | 66.4           | 81.9               | 86.7      |
| 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           | 2.43               | 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)   | 259            | 214                | 231       |
| phenotypes per cell type (median) | 252            | 200                | 209       |
| phenotypes per cell type (min)    | 71             | 57                 | 57        |
| phenotypes per cell type (max)    | 696            | 735                | 735       |

Table 3: 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.  |

Table 3: 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   |
|-------|-----------|---|
| 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  |

768 **Data Availability**

769 All data is publicly available through the following resources: - Human Phenotype Ontology (<https://hpo.jax.org>)  
770 - GenCC (<https://thegencc.org/>) - Descartes Human scRNA-seq atlas (<https://cellxgene.cziscience.com/collections/c114c20f-1ef4-49a5-9c2e-d965787fb90c>) - Human Cell Landscape scRNA-seq atlas (<https://cellxgene.cziscience.com/collections/38833785-fac5-48fd-944a-0f62a4c23ed1>) - Rare Disease Celltyping Port  
771 tal ([https://neurogenomics.github.io/rare\\_disease\\_celltyping\\_apps/home](https://neurogenomics.github.io/rare_disease_celltyping_apps/home))  
772  
773

774 **Code Availability**

775 All code is made freely available through the following GitHub repositories:

- 776 • KGExplorer (<https://github.com/neurogenomics/KGExplorer>)
- 777 • HPOExplorer (<https://github.com/neurogenomics/HPOExplorer>)
- 778 • MSTExplorer (<https://github.com/neurogenomics/MSTExplorer>)
- 779 • Code to replicate analyses ([https://github.com/neurogenomics/rare\\_disease\\_celltyping](https://github.com/neurogenomics/rare_disease_celltyping))
- 780 • Cell type-specific gene target prioritisation ([https://neurogenomics.github.io/RareDiseasePrioritisation/reports/prioritise\\_targets](https://neurogenomics.github.io/RareDiseasePrioritisation/reports/prioritise_targets))
- 781 • Complement system gene list (<https://www.genenames.org/data/genegroup/#!/group/492>)
- 782

783 **Acknowledgements**

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

787 **Funding**

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

791 **References**

- 792 1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A* **179**, 885–892 (2019).
- 793 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).
- 794 3. Rare diseases BioResource.
- 795 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).

- 796 5. Molster, C. *et al.* Survey of healthcare experiences of australian adults living with rare diseases. *Orphanet J. Rare Dis.* **11**, 30 (2016).
- 797 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).
- 798 7. Institute of Medicine (US) Committee on Accelerating Rare Diseases Research and Orphan Product Development, Field, M. J. & Boat, T. F. *Coverage and Reimbursement: Incentives and Disincentives for Product Development*. (National Academies Press (US), 2010).
- 799 8. Yates, N. & Hinkel, J. The economics of moonshots: Value in rare disease drug development. *Clin. Transl. Sci.* **15**, 809–812 (2022).
- 800 9. Nuijten, M. Pricing zolgensma - the world's most expensive drug. *J Mark Access Health Policy* **10**, 2022353 (2022).
- 801 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).
- 802 11. Gargano, M. A. *et al.* The human phenotype ontology in 2024: Phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
- 803 12. Köhler, S. *et al.* Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
- 804 13. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
- 805 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).
- 806 15. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: Analysis of the orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
- 807 16. Rare diseases, common challenges. *Nat. Genet.* **54**, 215 (2022).
- 808 17. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
- 809 18. 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).
- 810 19. McKusick, V. A. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
- 811 20. Maiella, S., Rath, A., Angin, C., Mousson, F. & Kremp, O. [Orphanet and its consortium: Where to find expert-validated information on rare diseases]. *Rev. Neurol.* **169 Suppl 1**, S3–8 (2013).

- 812 21. Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E. en & Cornel, M. C. [Orphanet: A european  
database for rare diseases]. *Ned. Tijdschr. Geneesk.* **152**, 518–519 (2008).
- 813 22. Firth, H. V. *et al.* DECIPHER: Database of chromosomal imbalance and phenotype in humans using  
ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
- 814 23. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell  
multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
- 815 24. 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).
- 816 25. Qi, R. & Zou, Q. Trends and potential of machine learning and deep learning in drug study at  
Single-Cell level. *Research* **6**, 0050 (2023).
- 817 26. 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).
- 818 27. Svensson, V., Veiga Beltrame, E. da & Pachter, L. A curated database reveals trends in single-cell  
transcriptomics. *Database* **2020**, (2020).
- 819 28. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
- 820 29. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
- 821 30. Kawabata, H. *et al.* Improving cell-specific recombination using AAV vectors in the murine CNS by  
capsid and expression cassette optimization. *Molecular Therapy Methods & Clinical Development* **32**,  
(2024).
- 822 31. O’Carroll, S. J., Cook, W. H. & Young, D. AAV targeting of glial cell types in the central and  
peripheral nervous system and relevance to human gene therapy. *Frontiers in Molecular Neuroscience*  
**13**, (2021).
- 823 32. Murphy, K., Schilder, B. M. & Skene, N. G. Harnessing generative AI to annotate the severity of all  
phenotypic abnormalities within the Human Phenotype Ontology. doi:[10.1101/2024.06.10.24308475](https://doi.org/10.1101/2024.06.10.24308475).
- 824 33. DiStefano, M. T. *et al.* The gene curation coalition: A global effort to harmonize gene–disease evidence  
resources. *Genetics in Medicine* **24**, 1732–1742 (2022).
- 825 34. Diehl, A. D. *et al.* The cell ontology 2016: Enhanced content, modularization, and ontology interop-  
erability. *J. Biomed. Semantics* **7**, 44 (2016).
- 826 35. Heim, C. E. *et al.* Myeloid-derived suppressor cells contribute to staphylococcus aureus orthopedic  
biofilm infection. *J. Immunol.* **192**, 3778–3792 (2014).
- 827 36. 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).
- 828 37. Stoll, H. *et al.* Staphylococcal enterotoxins Dose-Dependently modulate the generation of Myeloid-  
Derived suppressor cells. *Front. Cell. Infect. Microbiol.* **8**, 321 (2018).

- 829 38. 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).
- 830 39. Zhou, Z., Xu, M.-J. & Gao, B. Hepatocytes: A key cell type for innate immunity. *Cell. Mol. Immunol.*  
**13**, 301–315 (2016).
- 831 40. Dixon, L. J., Barnes, M., Tang, H., Pritchard, M. T. & Nagy, L. E. Kupffer cells in the liver. *Compr.  
Physiol.* **3**, 785–797 (2013).
- 832 41. Ladhami, S. N. *et al.* Invasive meningococcal disease in patients with complement deficiencies: A case  
series (2008-2017). *BMC Infect. Dis.* **19**, 522 (2019).
- 833 42. Rosain, J. *et al.* Strains responsible for invasive meningococcal disease in patients with terminal  
complement pathway deficiencies. *J. Infect. Dis.* **215**, 1331–1338 (2017).
- 834 43. 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).
- 835 44. Lung, T. *et al.* The complement system in liver diseases: Evidence-based approach and therapeutic  
options. *J Transl Autoimmun* **2**, 100017 (2019).
- 836 45. Reis, E. S. *et al.* Applying complement therapeutics to rare diseases. *Clin. Immunol.* **161**, 225–240  
(2015).
- 837 46. Seal, R. L. *et al.* Genenames.org: The HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009  
(2023).
- 838 47. Al-Hamoudi, W. K. Severe autoimmune hepatitis triggered by varicella zoster infection. *World J.  
Gastroenterol.* **15**, 1004–1006 (2009).
- 839 48. Brewer, E. C. & Hunter, L. Acute liver failure due to disseminated varicella zoster infection. *Case  
Reports Hepatol* **2018**, 1269340 (2018).
- 840 49. Eshchar, J., Reif, L., Waron, M. & Alkan, W. J. Hepatic lesion in chickenpox. A case report. *Gas-  
troenterology* **64**, 462–466 (1973).
- 841 50. Sugiyama, K., Tagawa, S. & Toda, M. Methods for visual understanding of hierarchical system struc-  
tures. *IEEE Trans. Syst. Man Cybern.* **11**, 109–125 (1981).
- 842 51. 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).
- 843 52. 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).
- 844 53. 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).

- 845 54. 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).
- 846 55. Demir, S. A., Timur, Z. K., Ateş, N., Martínez, L. A. & Seyrantepe, 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).
- 847 56. Solovyeva, V. V. *et al.* **New approaches to tay-sachs disease therapy.** *Frontiers in Physiology* **9**, (2018).
- 848 57. 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).
- 849 58. Ferro, A., Sheeler, C., Rosa, J.-G. & Cvetanovic, M. **Role of microglia in ataxias.** *Journal of molecular biology* **431**, 1792–1804 (2019).
- 850 59. Hol, E. M. & Pasterkamp, R. J. Microglial transcriptomics meets genetics: New disease leads. *Nature Reviews Neurology* **1–2** (2022) doi:[10.1038/s41582-022-00633-w](https://doi.org/10.1038/s41582-022-00633-w).
- 851 60. Lopes, K. de P. *et al.* **Atlas of genetic effects in human microglia transcriptome across brain regions, aging and disease pathologies.** *bioRxiv* 2020.10.27.356113 (2020) doi:[10.1101/2020.10.27.356113](https://doi.org/10.1101/2020.10.27.356113).
- 852 61. Gao, C., Jiang, J., Tan, Y. & Chen, S. **Microglia in neurodegenerative diseases: mechanism and potential therapeutic targets.** *Signal Transduction and Targeted Therapy* **8**, 1–37 (2023).
- 853 62. McQuade, A. & Blurton-jones, M. **Microglia in alzheimer's disease : Exploring how genetics and phenotype influence risk.** *Journal of Molecular Biology* **1–13** (2019) doi:[10.1016/j.jmb.2019.01.045](https://doi.org/10.1016/j.jmb.2019.01.045).
- 854 63. Zanello, G. *et al.* **Targeting shared molecular etiologies to accelerate drug development for rare diseases.** *EMBO Mol. Med.* **15**, e17159 (2023).
- 855 64. 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).
- 856 65. 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).
- 857 66. Bulaklak, K. & Gersbach, C. A. **The once and future gene therapy.** *Nat. Commun.* **11**, 5820 (2020).
- 858 67. Godbout, K. & Tremblay, J. P. **Prime editing for human gene therapy: Where are we now?** *Cells* **12**, (2023).
- 859 68. Kohn, D. B., Chen, Y. Y. & Spencer, M. J. **Successes and challenges in clinical gene therapy.** *Gene Ther.* **30**, 738–746 (2023).
- 860 69. Zhao, Z., Shang, P., Mohanraju, P. & Geijsen, N. **Prime editing: Advances and therapeutic applications.** *Trends Biotechnol.* **41**, 1000–1012 (2023).
- 861 70. Darrow, J. J. **Luxturna: FDA documents reveal the value of a costly gene therapy.** *Drug Discov. Today* **24**, 949–954 (2019).

- 862 71. Mendell, J. R. *et al.* Single-Dose Gene-Replacement therapy for spinal muscular atrophy. *N. Engl. J. Med.* **377**, 1713–1722 (2017).
- 863 72. 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).
- 864 73. Russell, S. *et al.* Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: A randomised, controlled, open-label, phase 3 trial. *Lancet* **390**, 849–860 (2017).
- 865 74. Lu, C.-F. FDA takes first step toward international regulation of gene therapies to treat rare diseases. (2024).
- 866 75. Brown, D. G., Wobst, H. J., Kapoor, A., Kenna, L. A. & Southall, N. Clinical development times for innovative drugs. *Nat. Rev. Drug Discov.* **21**, 793–794 (2022).
- 867 76. 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).
- 868 77. Zhou, Q. & Buchholz, C. J. Cell type specific gene delivery by lentiviral vectors: New options in immunotherapy. *Oncoimmunology* **2**, e22566 (2013).
- 869 78. Keren-shaul, H. *et al.* A unique microglia type associated with restricting development of alzheimer 's disease. *Cell* **169**, 1276–1290.e17 (2017).
- 870 79. Deczkowska, A. *et al.* Disease-associated microglia: A universal immune sensor of neurodegeneration. *Cell* **173**, 1073–1081 (2018).
- 871 80. DiStefano, M. T. *et al.* The gene curation coalition: A global effort to harmonize gene-disease evidence resources. *Genet. Med.* **24**, 1732–1742 (2022).
- 872 81. DiStefano, M. *et al.* P451: The gene curation coalition works to resolve discrepancies in gene-disease validity assertions. *Genetics in Medicine Open* **1**, 100498 (2023).
- 873 82. 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).
- 874 83. Osmond, M. *et al.* Outcome of over 1500 matches through the matchmaker exchange for rare disease gene discovery: The 2-year experience of Care4Rare canada. *Genet. Med.* **24**, 100–108 (2022).
- 875 84. Philippakis, A. A. *et al.* The matchmaker exchange: A platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).
- 876 85. Côté, R. *et al.* The ontology lookup service: Bigger and better. *Nucleic Acids Res.* **38**, W155–60 (2010).
- 877 86. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).

- 878 87. Putman, T. E. *et al.* The monarch initiative in 2024: An analytic platform integrating phenotypes,  
genes and diseases across species. *Nucleic Acids Res.* **52**, D938–D949 (2024).
- 879 88. Lazarin, G. A. *et al.* Systematic classification of disease severity for evaluation of expanded carrier  
screening panels. *PLoS One* **9**, e114391 (2014).

880

881

882 **Supplementary Materials**

883 **Supplementary Figures**

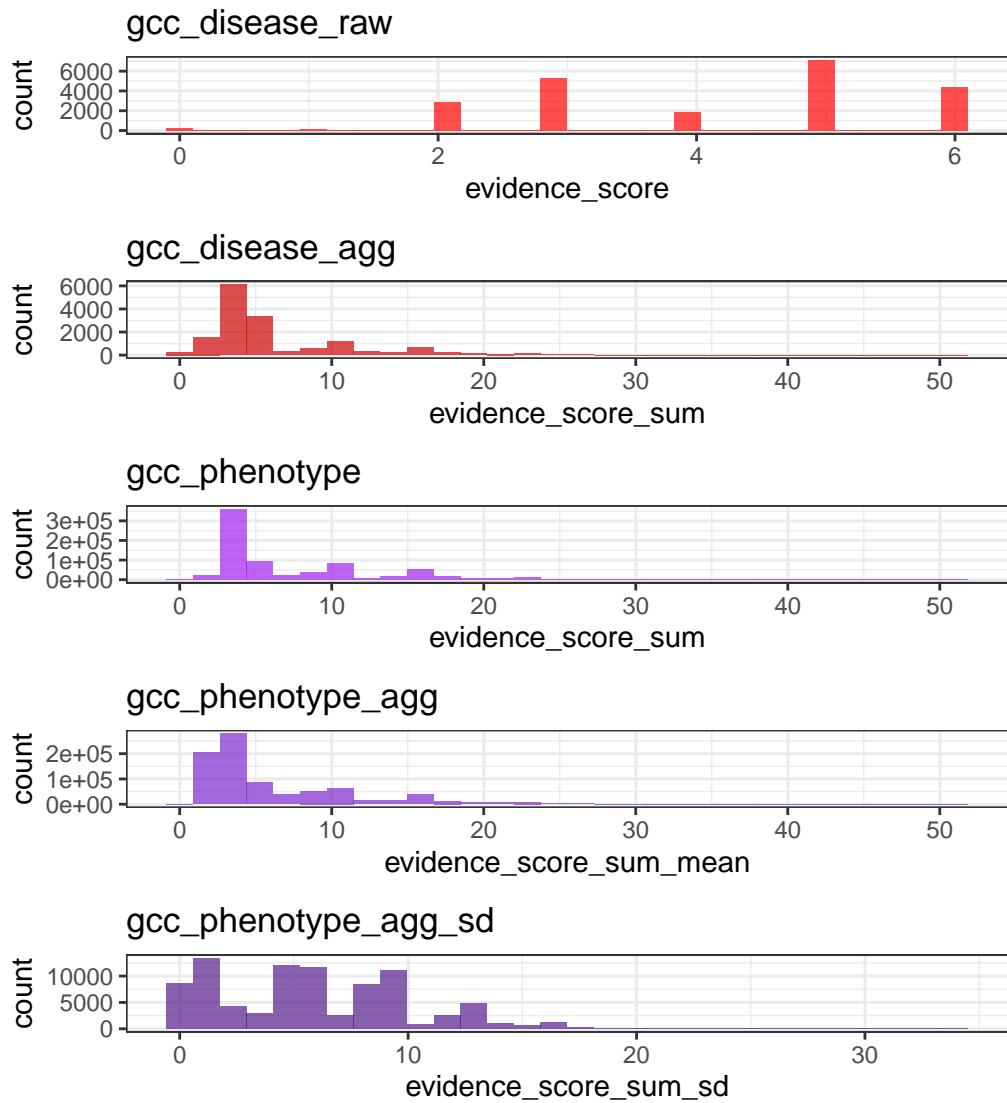


Figure 8: Distribution of evidence scores at each processing step.

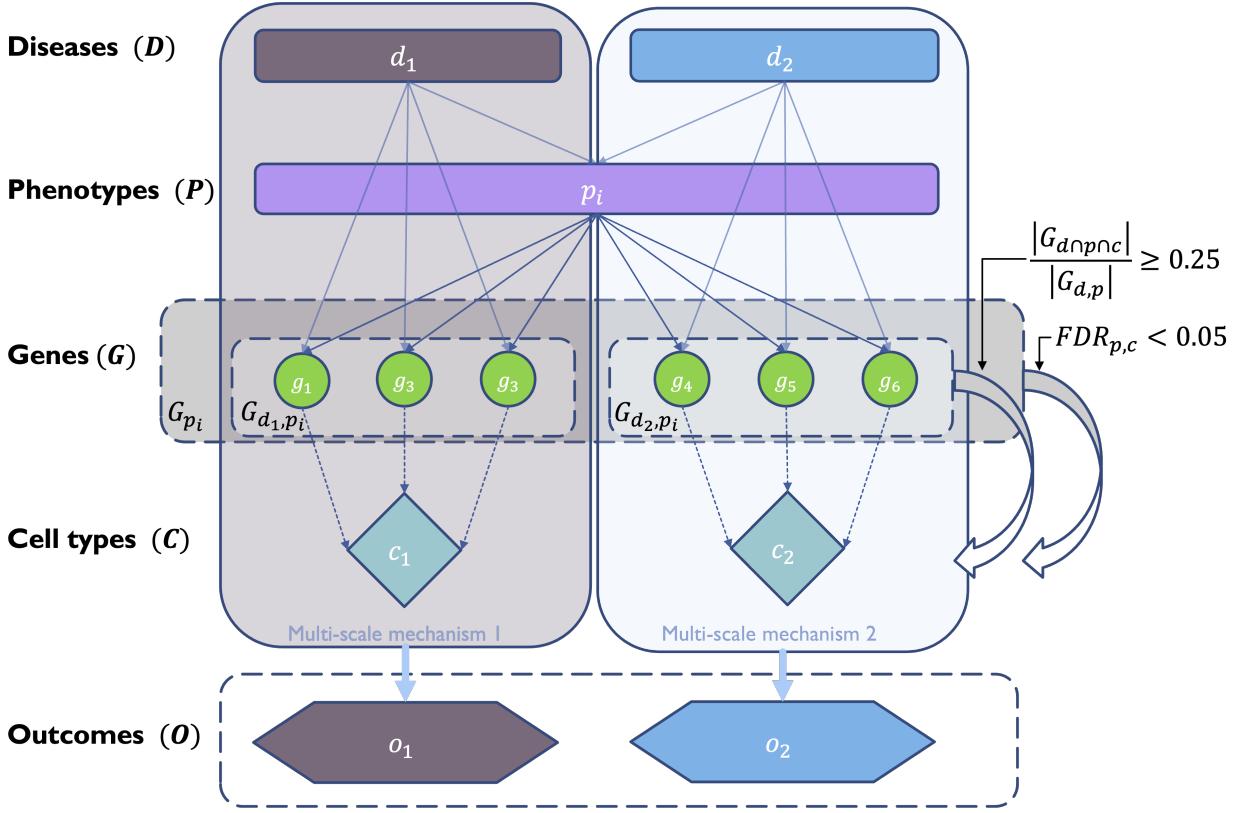


Figure 9: Diagrammatic overview of multi-scale disease investigation strategy. Here we provide an abstract example of differential disease aetiology across multiple scales: diseases ( $D$ ), phenotypes ( $P$ ), cell types ( $C$ ), genes ( $G$ ), and clinical outcomes ( $O$ ). In the HPO, genes are assigned 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 ( $FDR < 0.05$ ). 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%.

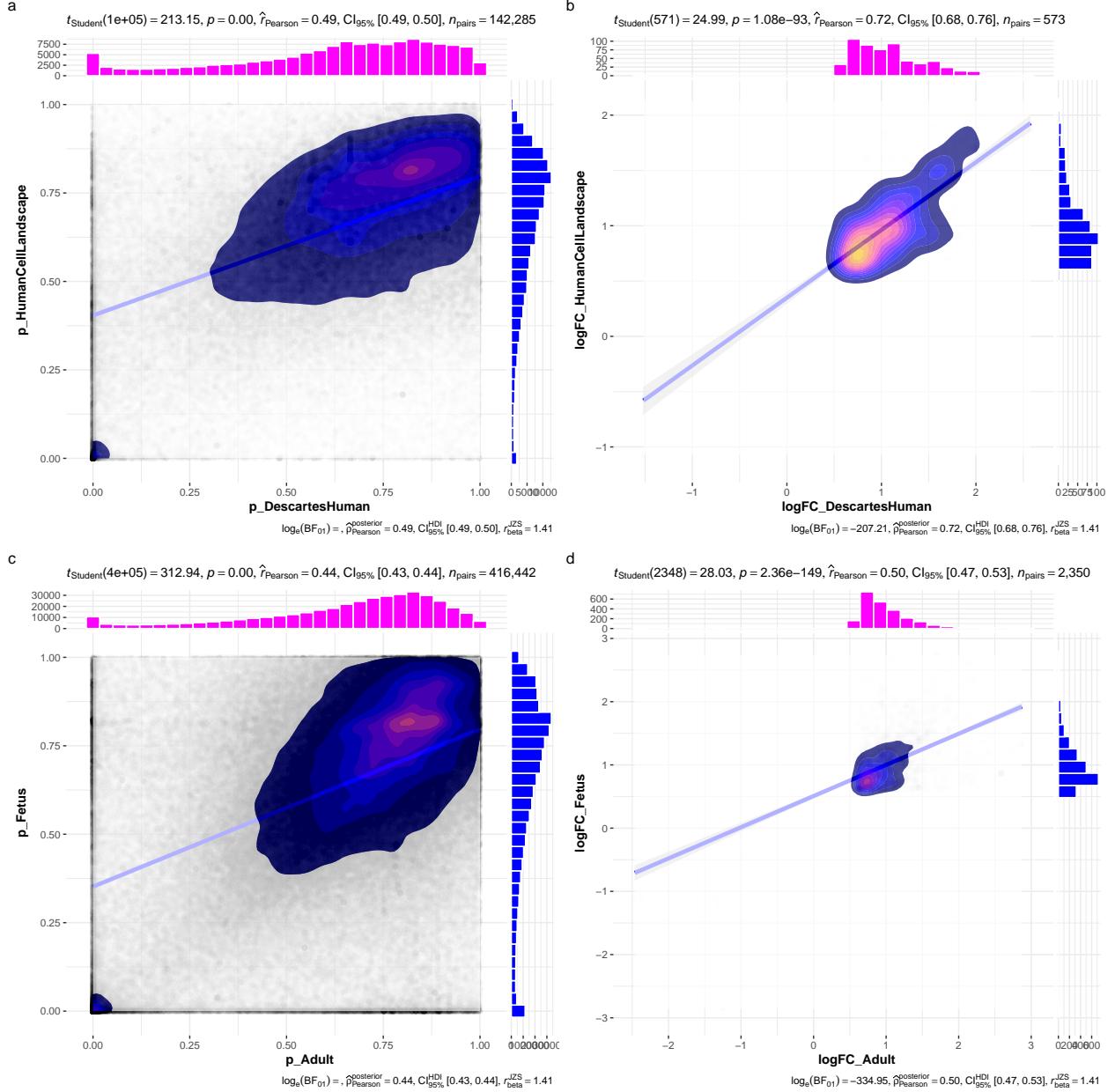


Figure 10: Inter- and intra-dataset validation across the different CellTypeDataset (CTD) and developmental stages. Correlations are computed using Pearson correlation coefficient. Point density is plotted using a 2D kernel density estimate. **a** Correlation between the uncorrected p-values from all 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<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<0.05) using the Human Cell Landscape fetal samples vs. Human Cell Landscape adult samples.

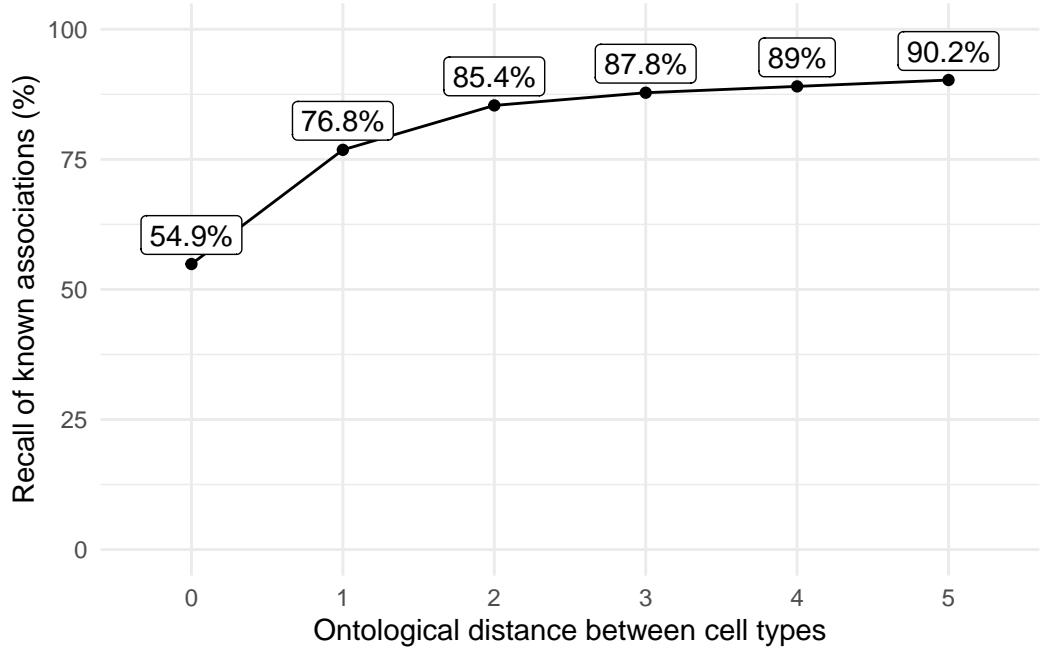


Figure 11: Recall of ground-truth Monarch Knowledge Graph phenotype-cell type relationships at each ontological distance between cell types according to the Cell Ontology.

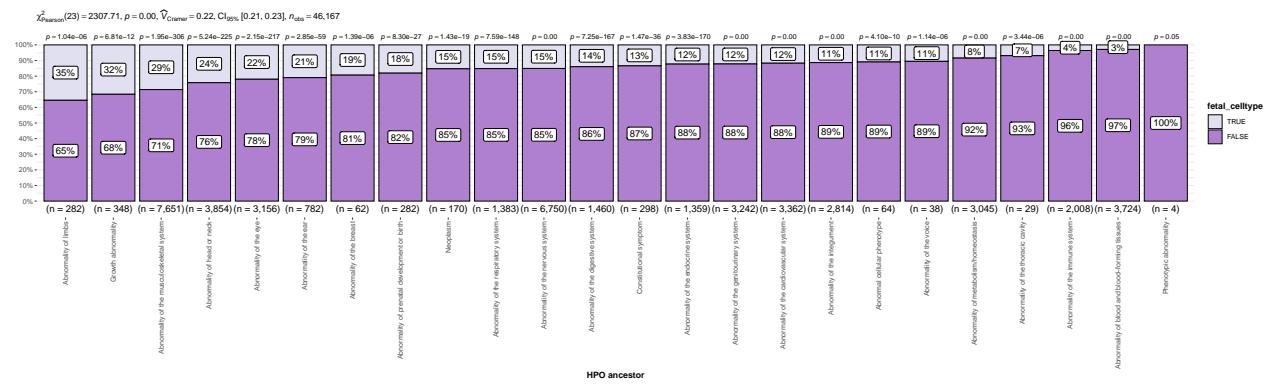


Figure 12: The proportion of cell type-phenotype association tests that are enriched for foetal cell types within each HPO branch.

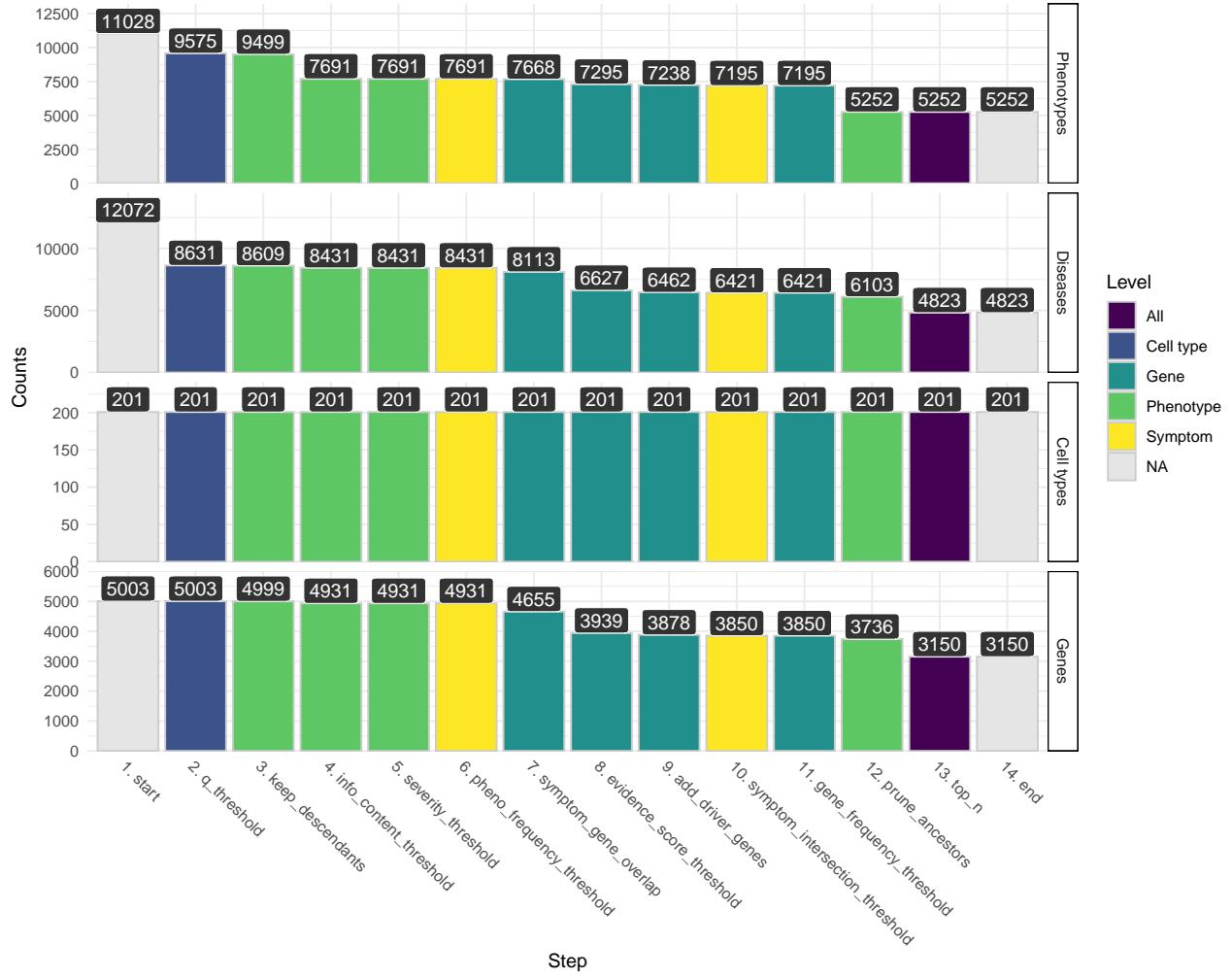


Figure 13: Prioritised target filtering steps. This plot visualises the number of unique phenotype-cell type associations, cell types, genes, and phenotypes (*y-axis*) at each filtering step (*x-axis*) within the multi-scale therapeutic target prioritisation pipeline. Each step in the pipeline can be easily adjusted according to user preference and use case. See Table 3 for descriptions and criterion of each filtering step.

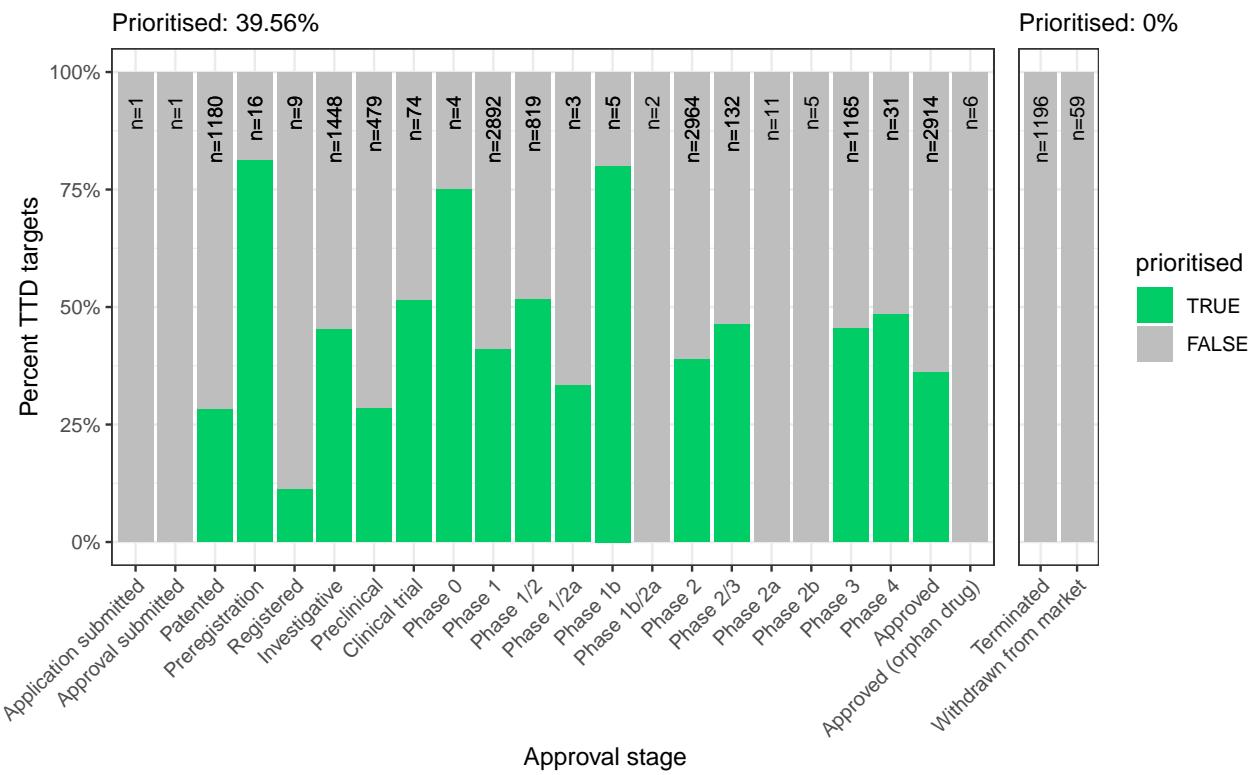


Figure 14: Therapeutics - Validation of prioritised therapeutic targets. Proportion of existing all therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline.

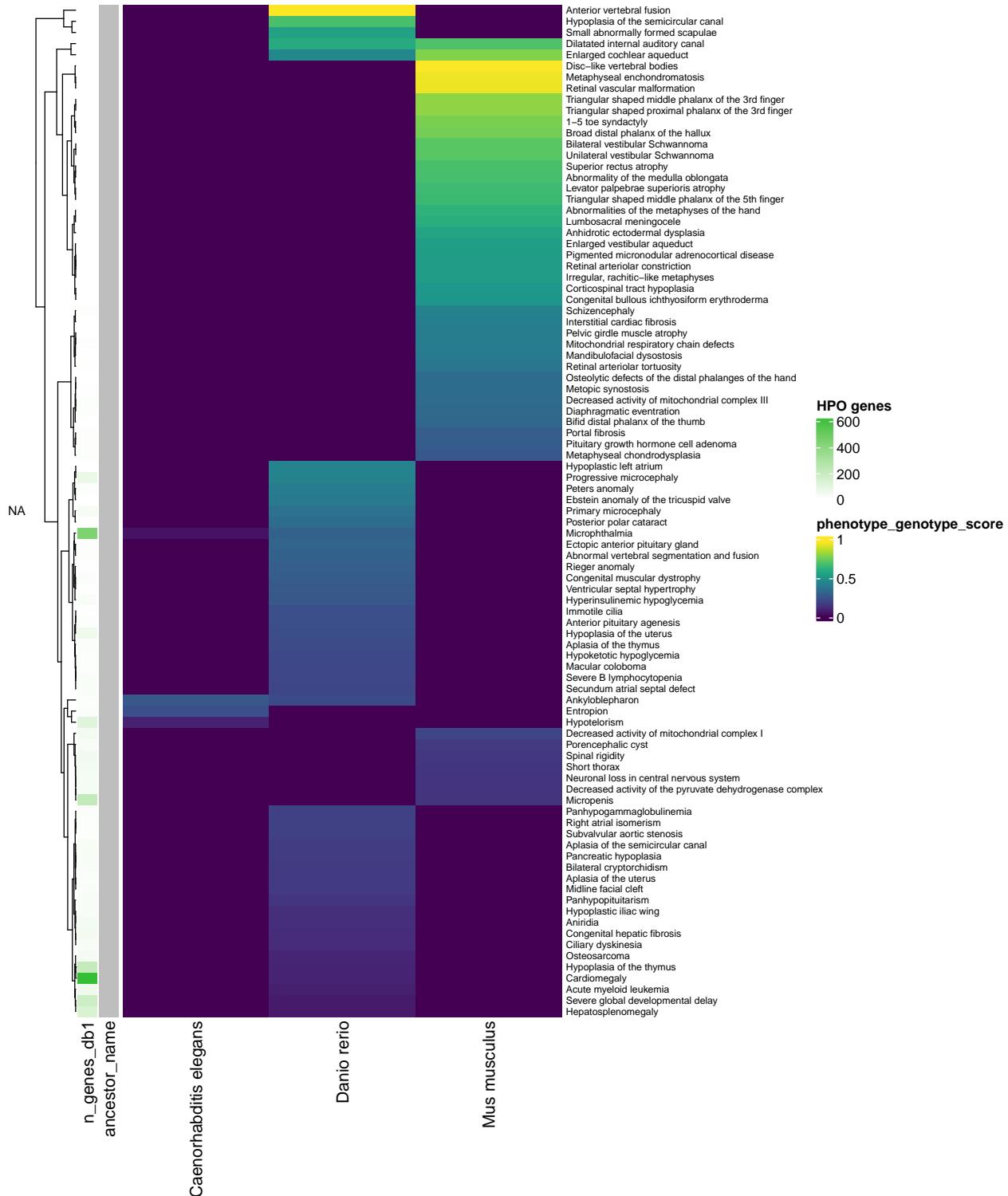
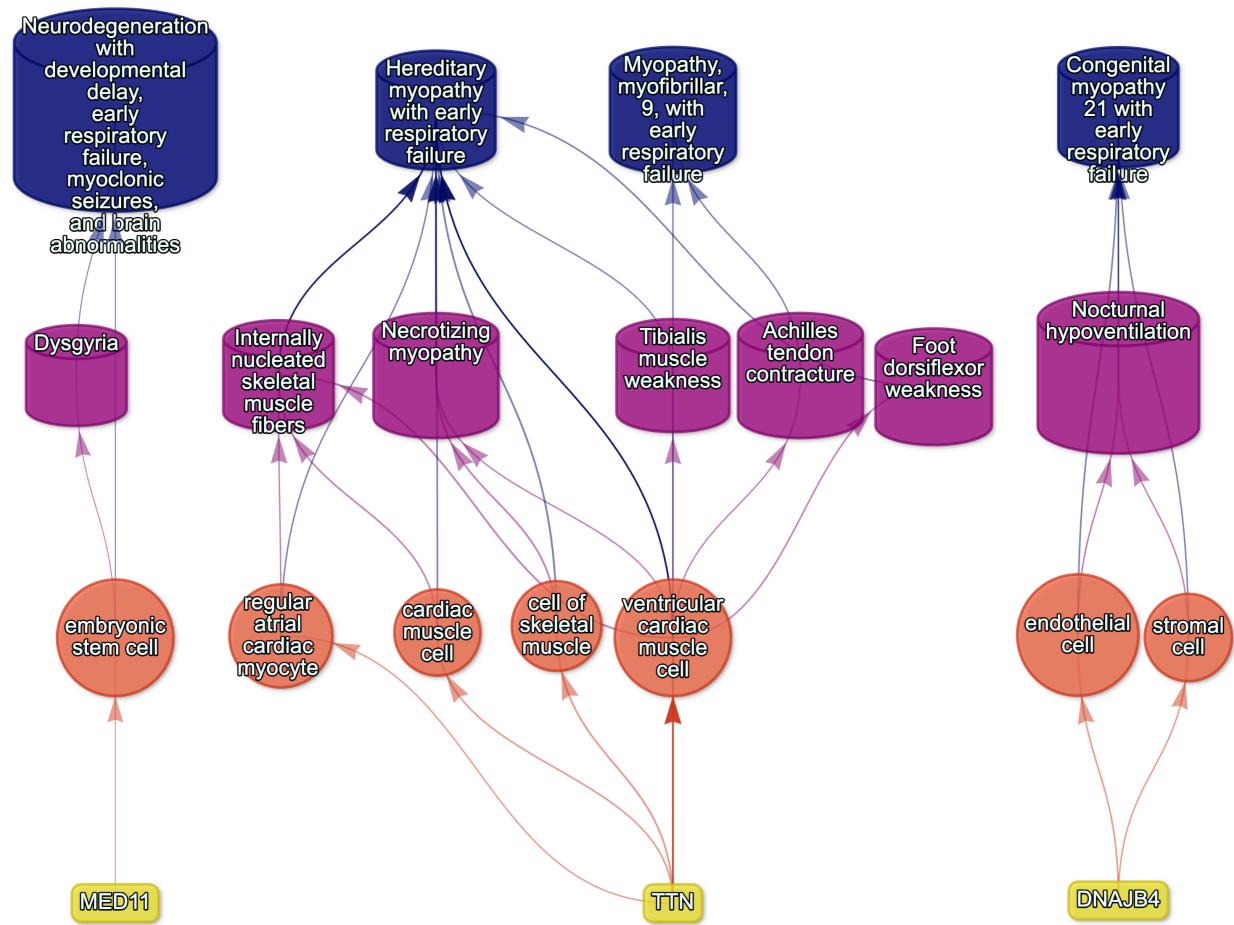


Figure 15: Identification of translatable experimental models. Interspecies translatability of human phenotypes nominated by the gene therapy prioritised pipeline. Above, the combined ontological-genotypic similarity score ( $SIM_{o,g}$ ) is displayed as the heatmap fill colour stratified by the model organism ( $x$ -axis). An additional column ("n\_genes\_db1" on the far left) displays the total number of unique genes annotated to the phenotypic within the HPO. Phenotypes are clustered according to their ontological similarity in the HPO ( $y$ -axis).

884 **Supplementary Tables**

Table 4: 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        |



(a) Respiratory failure

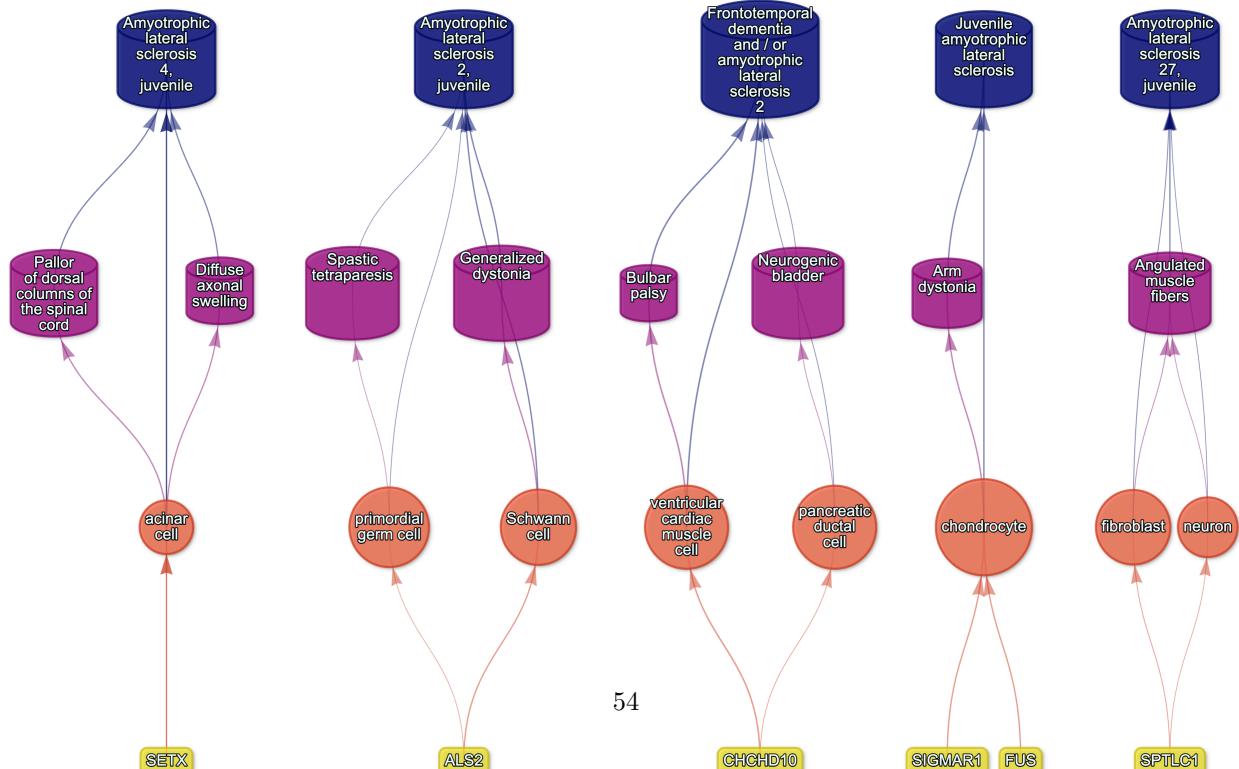


Table 5: On-target cell types for each HPO ancestral branch.

| 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 |

Table 5: On-target cell types for each HPO ancestral branch.

| hpo_branch                                | cl_branch                             | cl_name                            | cl_id      |
|---|---------------------------------------|------------------------------------|------------|
| 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 |
| 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 |

Table 5: On-target cell types for each HPO ancestral branch.

| hpo_branch                        | cl_branch   | cl_name                         | cl_id      |
|-----------------------------------|-------------|---------------------------------|------------|
| 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 5: 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 6: 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        |