

<sup>1</sup> Cell type-specific contextualisation of the human phenome: towards  
<sup>2</sup> the systematic treatment of all rare diseases

<sup>3</sup> Brian M. Schilder      Kitty B. Murphy      Hiranyamaya Dash      Yichun Zhang  
<sup>4</sup> Robert Gordon-Smith      Jai Chapman      Momoko Otani      Nathan G. Skene

<sup>5</sup> 2026-02-24

## **6 Abstract**

7 Rare diseases (RDs) are a highly heterogeneous and underserved group of conditions. Most RDs have a  
8 strong genetic basis but their causal pathophysiological mechanisms remain poorly understood. We therefore  
9 systematically characterised the cell type–specific mechanisms for all genetically defined RD phenotypes by  
10 integrating the Human Phenotype Ontology with whole-body single-cell transcriptomic atlases from embryonic,  
11 foetal, and adult samples. This revealed significant associations between 201 cell types and 9,575/11,028  
12 (86.7%) phenotypes across 8,628 RDs, substantially expanding knowledge of phenotype–cell type links. We  
13 then prioritised phenotypes for clinical impact based on severity (e.g. lethality, motor/mental impairment)  
14 and gene therapy compatibility (e.g. cell type specificity, postnatal treatability). All results are reproducible  
15 and freely available, including via an interactive web portal (<https://neurogenomics-ukdri.dsi.ic.ac.uk/>),  
16 representing a major advance toward treating patients across a broad spectrum of serious RDs.

## **17 Introduction**

18 Rare diseases (RDs) are individually uncommon but collectively this class of over 10,000 conditions affects  
19 300–400 million people worldwide (1 in 10–20 individuals)<sup>1,2</sup>, 75% of RD patients are children, with a 30%  
20 rate of mortality by age five<sup>3</sup>. Diagnosis is challenging due to highly variable presentations, averaging five  
21 years<sup>4</sup>, with ~46% misdiagnosed and >75% never diagnosed<sup>5</sup>. Prognosis is similarly difficult. Treatments  
22 exist for <5% of RDs<sup>6</sup> and high development costs for small patient populations deter investment<sup>7,8</sup>, making  
23 these therapies among the world’s most expensive<sup>9,10</sup>. High-throughput therapeutic discovery could lower  
24 costs and accelerate delivery.

25 A major barrier in research and clinical care of diseases is inconsistent medical terminology. The Human  
26 Phenotype Ontology (HPO) provides a unified, hierarchical framework of 18,082 phenotypes spanning 10,300  
27 RDs<sup>11–13</sup>, integrated into diagnostics and linked to other ontologies (e.g. SNOMED CT, UMLS, ICD). Over  
28 80% of RDs have known genetic causes<sup>14</sup>, with HPO gene annotations curated from OMIM, Orphanet,  
29 DECIPHER, and case reports. Yet gene lists alone lack the tissue and cell type context essential for  
30 understanding pathogenesis and improving diagnosis, prognosis, and treatment.

31 Single-cell RNA-seq (scRNA-seq) now enables transcriptome-wide profiling at cellular resolution<sup>15–17</sup>. Com-  
32 prehensive atlases such as Descartes Human<sup>18</sup> and Human Cell Landscape<sup>19</sup> cover embryonic to adult stages  
33 across tissues, providing gene signatures for hundreds of cell subtypes. Integrating RD gene annotations with  
34 these profiles reveals the specific cell types through which genes act, including understudied cell types.

35 Cell type-specific mechanisms are critical for guiding the development of effective therapeutics, especially  
36 virally-mediated gene therapies<sup>20,21</sup>. Knowledge of the specific causal cell types can enhance efficacy and  
37 improve safety by avoiding off-target effects. To facilitate these key insights, we developed a high-throughput  
38 pipeline to nominate cell type–resolved gene therapy targets across thousands of RD phenotypes, ranked by

39 composite phenotype severity scores<sup>22</sup>. This work expands knowledge of the cell types, organ systems, and  
40 life stages underlying RDs, with direct applications to precision therapeutic development.

41 **Results**

42 **Phenotype-cell type associations**

43 We systematically investigated cell types underlying HPO phenotypes, hypothesising that genes with cell  
44 type-specific expression are most relevant to those cell types, and that disrupting such genes will have  
45 variable effects across cellular contexts. More precisely, for each phenotype we created a list of associated  
46 genes weighted by the strength of the evidence supporting those associations, imported from the Gene  
47 Curation Coalition (GenCC)<sup>23</sup>. Analogously, we created mean gene expression profiles for each cell type  
48 using scRNA-seq atlases and then normalised them to compute cell type specificity of gene expression (see  
49 **Methods** subsection *Single-cell transcriptomic atlases* for further details).

50 For comprehensiveness, we used two pan-tissue scRNA-seq atlases: Descartes Human (~4 million single-nuclei  
51 and single-cells from 15 foetal tissues)<sup>18</sup> and Human Cell Landscape (~703,000 single-cells from 49 embryonic,  
52 foetal and adult tissues)<sup>19</sup>. For every unique combination of phenotype and cell type, we trained a generalised  
53 linear regression model to test for association between the respective gene-phenotype association scores and  
54 gene-cell type expression specificity scores (Fig. 1). We then applied stringent multiple testing correction to  
55 control the false discovery rate (FDR) across all tests, and significant phenotype–cell type associations were  
56 identified at FDR<0.05.

57 In Descartes Human, 19,929 / 848,078 (2.35%) tests were significant across 77 cell types and 7,340 phenotypes.  
58 In Human Cell Landscape, the corresponding values were 1.96% significant tests, 124 cell types, and 9,049  
59 phenotypes, with more phenotypes linked to at least one cell type due to greater cell type diversity and  
60 life-stage coverage. Across both atlases, the median number of significant cell types per phenotype was 3,  
61 indicating specificity of associations. Overall, 8,628/8,631 (>99.96%) of diseases had significant cell type  
62 associations for at least one phenotype. Full stratified results are provided in Table S2.

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

64 We intuit that organ system-specific abnormalities are often driven by cell types within that system. The  
65 HPO’s high-level categories allow systematic testing; for example, heart phenotypes should typically involve  
66 cardiocytes, and nervous system abnormalities should involve neural cells. All cell types in our single-cell  
67 atlases were mapped to the Cell Ontology (CL), a hierarchical vocabulary of cell types.

68 A cell type was considered *on-target* for an HPO branch if it belonged to a matching CL branch (Table S4).  
69 For each HPO branch (Fig. 2b), we tested whether cell types were more often associated with phenotypes in  
70 that branch compared to all others, and identified those overrepresented at FDR<0.05. All 7 HPO branches



**Figure 1: Multi-modal data fusion reveals the cell types underlying thousands of human phenotypes.** Schematic overview of study design in which we numerically encoded the strength of evidence linking each gene and each phenotype (using the Human Phenotype Ontology and GenCC databases). We then created gene signature profiles for all cell types in the Descartes Human and Human Cell Landscape scRNA-seq atlases. Finally, we iteratively ran generalised linear regression tests between all pairwise combinations of phenotype gene signatures and cell type gene signatures. The resulting associations were then used to nominate cell type-resolved gene therapy targets for thousands of rare diseases.

<sup>71</sup> showed disproportionate associations with on-target cell types from their respective organ systems.

<sup>72</sup> We hypothesised that more strongly significant phenotype–cell type associations are more likely to be on-target.

<sup>73</sup> Grouping  $-\log_{10}(\text{p-values})$  into six bins, we calculated the proportion of on-target cell types per HPO–CL

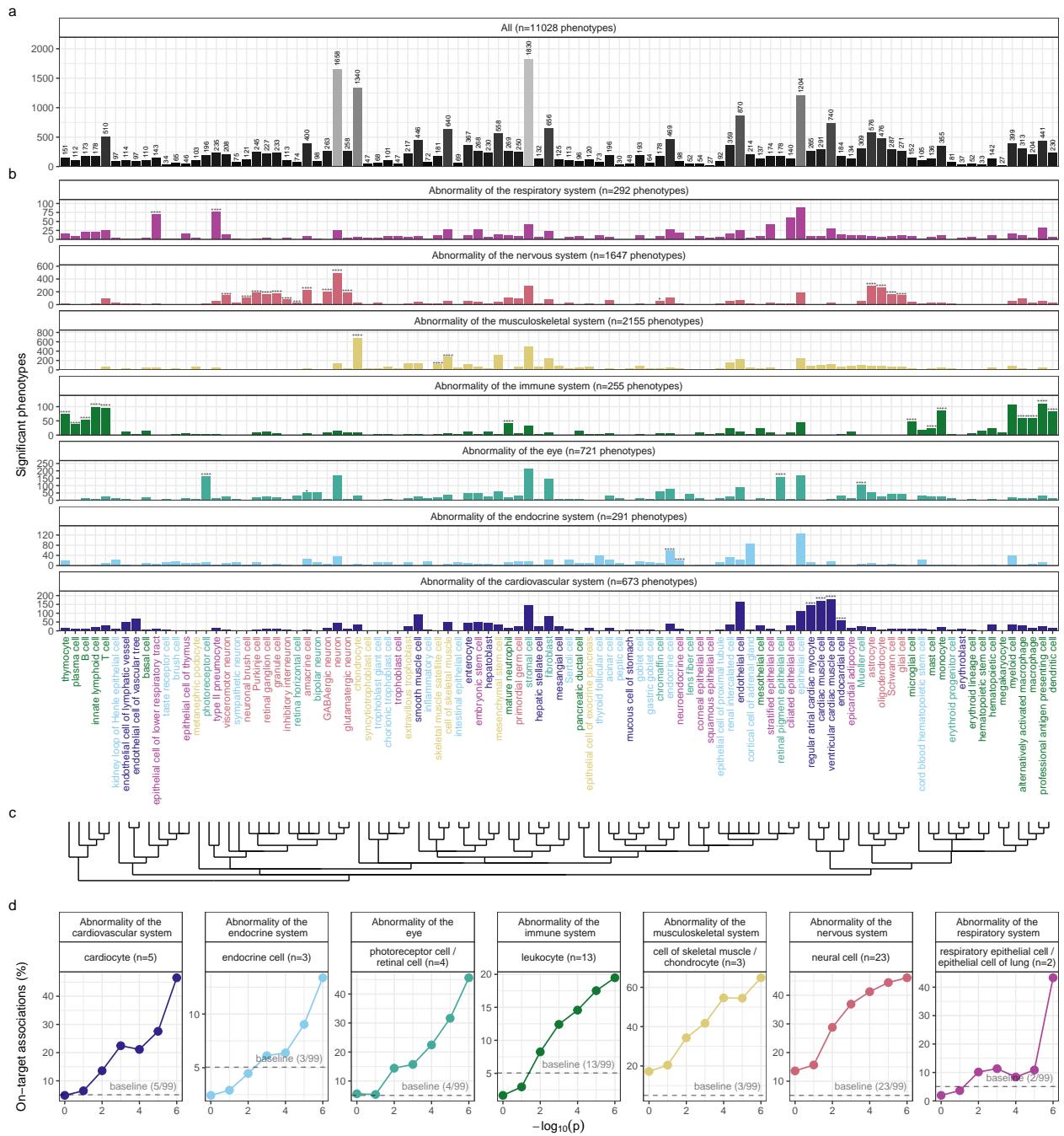
<sup>74</sup> branch pairing. Indeed, this proportion consistently increased with association significance ( $r_{Pearson} = 0.63$ ,

<sup>75</sup>  $p = 1.1 \times 10^{-6}$ ). For example, in nervous system abnormalities neural cells constituted only 23% of all tested

<sup>76</sup> cell types, yet made up 46% of associations within the  $-\log_{10}(\text{p-values}) \geq 6$  bin (which corresponds to

<sup>77</sup>  $p \leq 10^{-7}$ ). This confirms that stronger associations are more likely to involve on-target cell types, confirming

<sup>78</sup> our association strategy captures real relationships.



**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 associations 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>24</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.

Figure 2

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

80 If our methodology works, it should yield consistent phenotype-cell type associations across different datasets.  
81 We therefore tested for the consistency of our results across the two single-cell reference datasets (Descartes  
82 Human vs. Human Cell Landscape) for the subset of overlapping cell types Fig. S3. In total there were  
83 142,285 phenotype-cell type associations to compare across the two datasets (across 10,945 phenotypes and  
84 13 cell types annotated to the exact same CL term. We found that the correlation between association  
85 effect sizes across the two datasets was significant ( $r_{Pearson}=0.58$ ,  $p=1.4 \times 10^{-95}$ ). Within the subset of  
86 phenotype-cell type associations that were significant in both single-cell datasets ( $FDR<0.05$ ), we found that  
87 degree of correlation between the association effect sizes across datasets was even stronger ( $r_{Pearson}=0.73$ ,  
88  $p=1.4 \times 10^{-95}$ ). The symmetric replication rate, i.e. the mean percentage of associations that were significant  
89 in both datasets at  $FDR<0.05$ , was 39%.

90 While cell types change their transcriptional signatures over the course of development, we would nevertheless  
91 expect there to be some degree of correlation between the developing and mature versions of the same  
92 cell types. Therefore, we also checked for the intra-dataset consistency between the p-values of the foetal  
93 and adult samples in the Human Cell Landscape, showing a very similar degree of correlation as the inter-  
94 dataset comparison for full results estimates ( $r_{Pearson}=0.44$ ,  $p=2.2 \times 10^{-308}$ ) and the significant subset  
95 ( $r_{Pearson}=0.40$ ,  $p=4.5 \times 10^{-92}$ ). The symmetric replication rate was 36%.

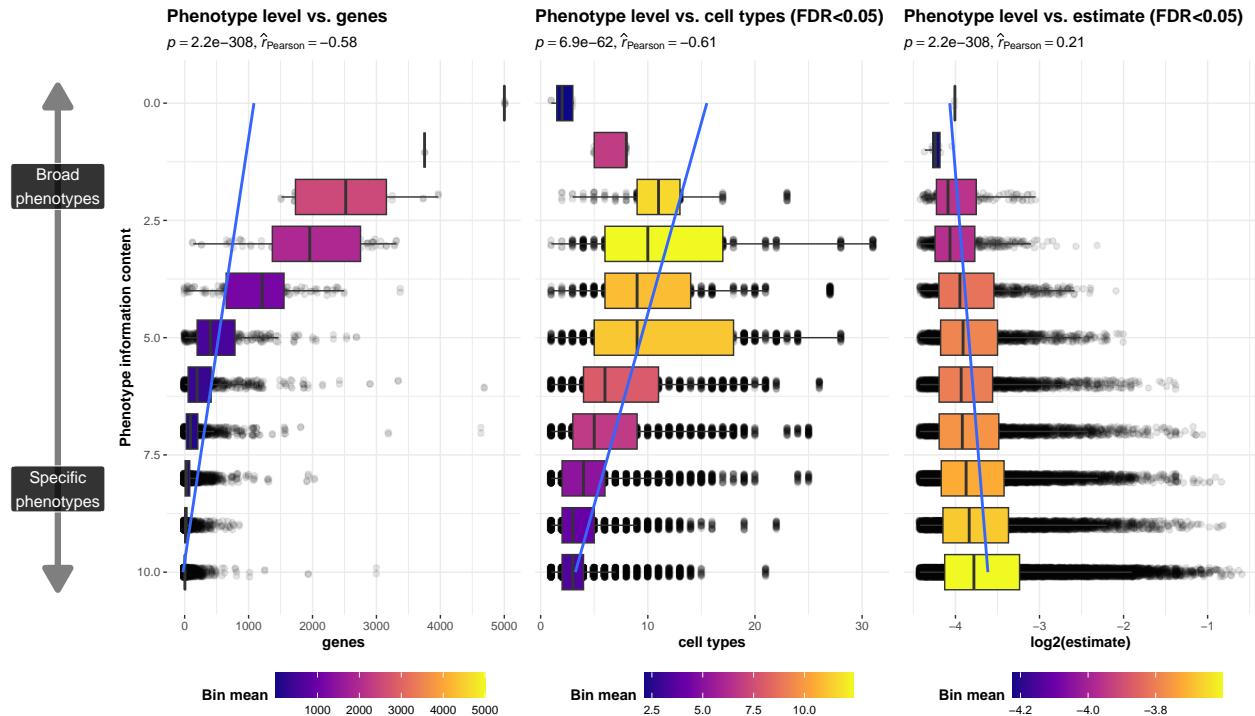
96 All inter- and intra-dataset correlation tests were confirmed with permutation testing (1,000 permutations,  
97 Table S8). Together, these results suggest that our approach to identifying phenotype-cell type associations  
98 is both replicable and generalisable to new datasets.

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

100 Higher levels of the ontology represent broad classes of phenotype (e.g. ‘Abnormality of the nervous system’)  
101 while the lower levels can get very detailed (e.g. ‘Spinocerebellar atrophy’). The higher level phenotypes  
102 inherit all genes associated with lower level phenotypes, so naturally they have more genes than the lower  
103 level phenotypes (Fig. 3a;  $r_{Pearson}=-0.58$ ,  $p<2.2 \times 10^{-308}$ ).

104 Next, we reasoned that the more detailed and specific a phenotype is, the more likely it is to be driven by  
105 fewer cell types. For example, while ‘Neurodevelopmental abnormality’ could plausibly be driven by any/all  
106 cell types in the brain, it is more likely that ‘Impaired visuospatial constructive cognition’ is driven by a  
107 select subset of cell types. This was indeed the case, as we observed a strongly significant negative correlation  
108 between phenotype level and the number of significantly associated cell types (Fig. 3b;  $r_{Pearson}=-0.61$ ,  
109  $p<6.9 \times 10^{-62}$ ). We also found that the phenotype-cell type association effect size increased with greater  
110 phenotype specificity, showing that more specific phenotypes are strongly associated with a refined subset of  
111 cell types (Fig. 3c;  $r_{Pearson}=0.21$ ,  $p<2.2 \times 10^{-308}$ ). In all cases, we confirmed the robustness of these results  
112 by iteratively comparing the observed correlations against permuted baseline correlations (1,000 permutations,

113  $p < 2.2 \times 10^{-308}$ ; Table S7).



**More specific phenotypes are associated with fewer, more specific genes and cell types.** Information content (IC), is a normalised measure of ontology term specificity. Terms with lower IC represent the broadest HPO terms (e.g. ‘All’), while terms with higher IC indicate progressively more specific HPO terms (e.g. ‘Contracture of proximal interphalangeal joints of 2nd-5th fingers’). Box plots show the relationship between HPO phenotype IC and **a**, the number of genes annotated to each phenotype, **b**, the number of significantly enriched cell types, **c**, the effect sizes (absolute model  $R^2$  estimates after log-transformation) of significant phenotype-cell type association tests. IC was binned only for the purposes of visualisation; all statistics were performed on the unbinned data. Boxes are coloured by the mean *x-axis* value within each IC bin.

Figure 3

#### 114 Validation of phenotype-cell type associations using biomedical knowledge graphs

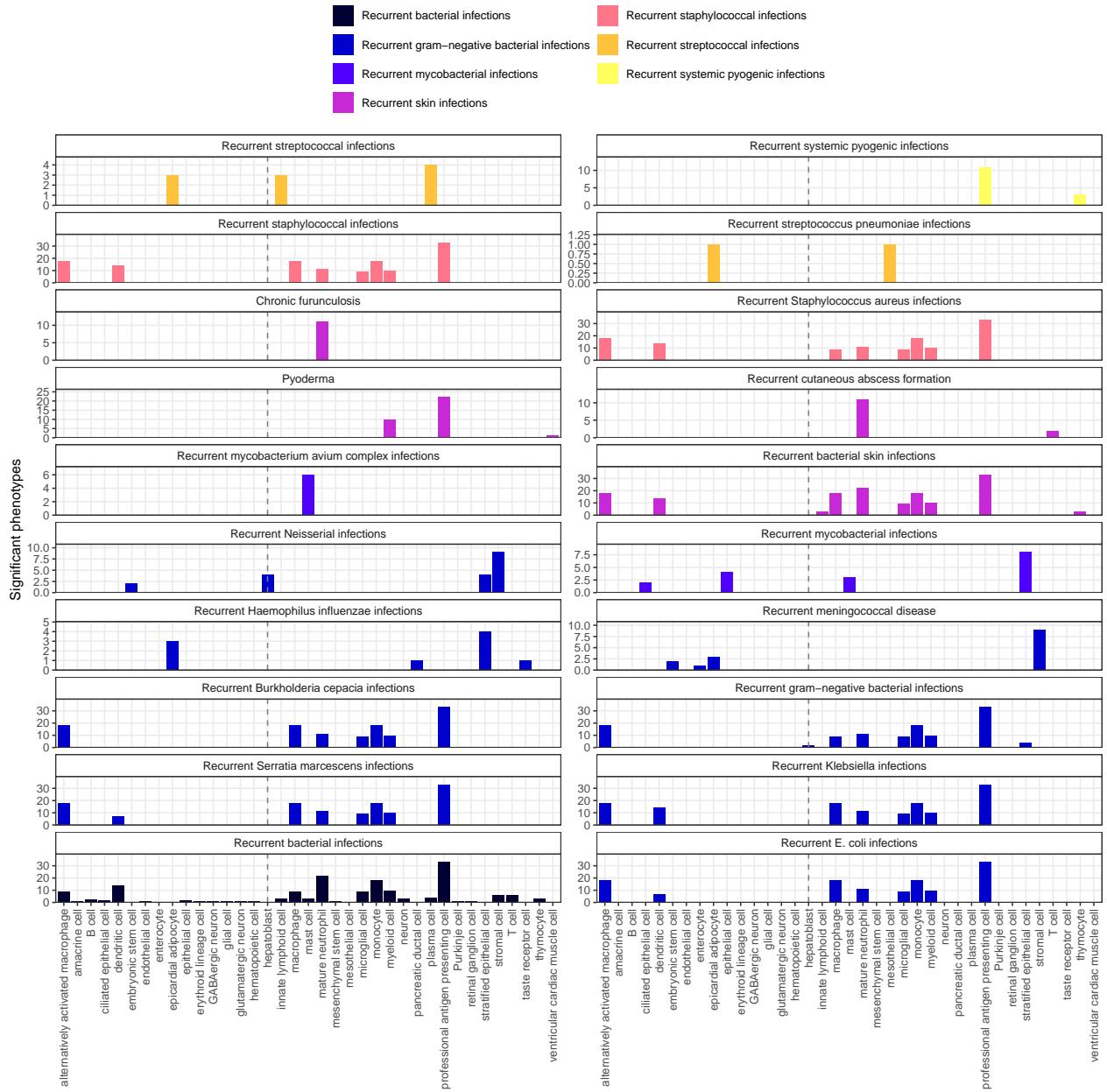
115 To validate phenotype–cell type associations without literature bias, we used the Monarch Knowledge Graph  
 116 (MKG), a curated database of biomedical concepts and relationships containing 103 known associations<sup>25</sup>.  
 117 The MKG served as a benchmark for the field’s current knowledge. For each MKG association, we calculated  
 118 the proportion of cell types recovered in our results at different ontological distances in the Cell Ontology.  
 119 Distance 0 indicates the closest possible match (e.g. “monocyte” vs. “monocyte”), with greater distances  
 120 reflecting progressively broader matches (e.g. distance 1: “monocyte” vs. “classical monocyte”). The  
 121 theoretical maximum recall was capped by the percentage of MKG phenotypes for which we identified at  
 122 least one significant association ( $FDR_{pc}$ ). In other words, since our results only contain significant cell  
 123 type associations for 90% of the phenotypes in the MKG’s phenotype-cell type associations, our maximum  
 124 achievable performance was 90% recall.

125 Our results included at least one significant cell type for 90% of MKG phenotypes. At distance 0, we recalled

126 57% of associations; at distance 1, recall rose to 77%, reaching a maximum of 90% at the largest allowed  
127 distance. Precision could not be computed, as MKG lists only true positives. Overall, these benchmarks show  
128 that our approach recovers most known phenotype–cell type associations while generating many novel ones.

129 **Phenome-wide analyses discover novel phenotype-cell type associations**

130 Having confirmed many phenotype-cell type associations match prior expectations, we explored novel links  
131 for undercharacterised phenotypes. ‘Recurrent bacterial infections’ (19 descendants, e.g. staphylococcal,  
132 streptococcal, Neisserial) mostly associated with immune cells (e.g. macrophages, dendritic cells, T cells,  
133 monocytes, neutrophils) (Fig. 4). Known links include ‘Recurrent staphylococcal infections’ with myeloid  
134 cells<sup>26–29</sup>, where monocytes were most strongly associated ( $FDR=1.0 \times 10^{-30}$ ,  $\beta=0.18$ ). Notably, amongst the  
135 recurrent bacterial infection phenotypes, hepatoblasts were exclusively associated with ‘Recurrent Neisserial  
136 infections’, hereafter RNI (Descartes Human:  $FDR=1.1 \times 10^{-6}$ ,  $\beta=8.2 \times 10^{-2}$ ).



**Recurrent bacterial infection subtypes are associated with different sets of immune and non-immune cell types.** 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 associated with ‘Recurrent Neisserial 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’).

Figure 4

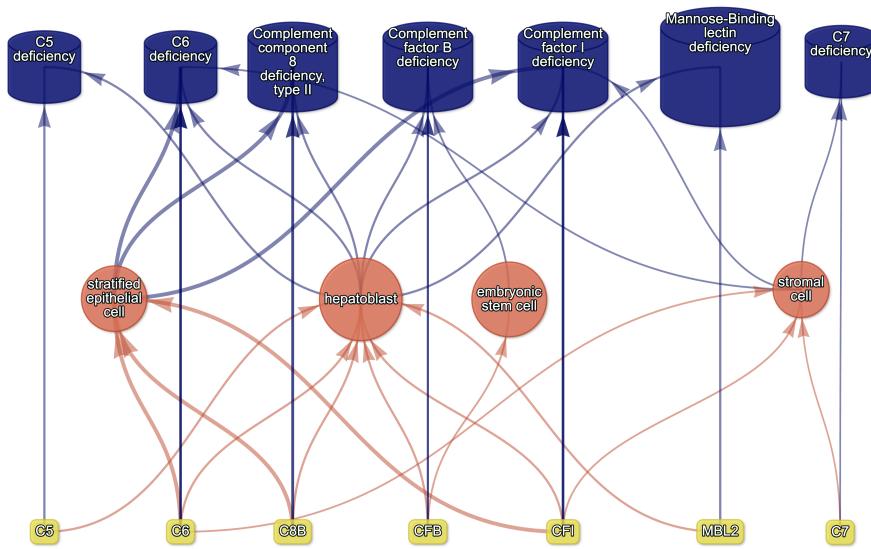
- 137 To better understand the multi-scale mechanisms underlying RNI susceptibility, we visualised the putative  
 138 causal relationships between genes, cell types and diseases associated with RNI as a network (Fig. 5). The  
 139 phenotype RNI was connected to cell types through the aforementioned association test results (FDR<0.05).  
 140 Genes that were primarily driving these associations (i.e. genes both strongly linked with RNI and highly

141 specifically expressed in the given cell type) were designated as *driver genes* and retained for plotting  
142 (see **Methods** for details). While a gene being non-specifically expressed does not necessarily mean it is  
143 unimportant for the function of a cell (many ubiquitously expressed genes are essential for cell function),  
144 gene expression specificity is nevertheless a useful metric to unambiguously link cell types to other biological  
145 entities (e.g. phenotypes). Diseases were then connected to phenotypes and gene nodes based on HPO  
146 annotations. Finally, diseases were also connected to cell types as a transitive property of being connected  
147 to phenotypes, but only if the disease gene set overlapped with 25% or more of the driver genes for that  
148 particular phenotype-cell type relationship (see **Methods** subsection *Symptom-cell type associations* and  
149 Fig. S2 for further explanation).

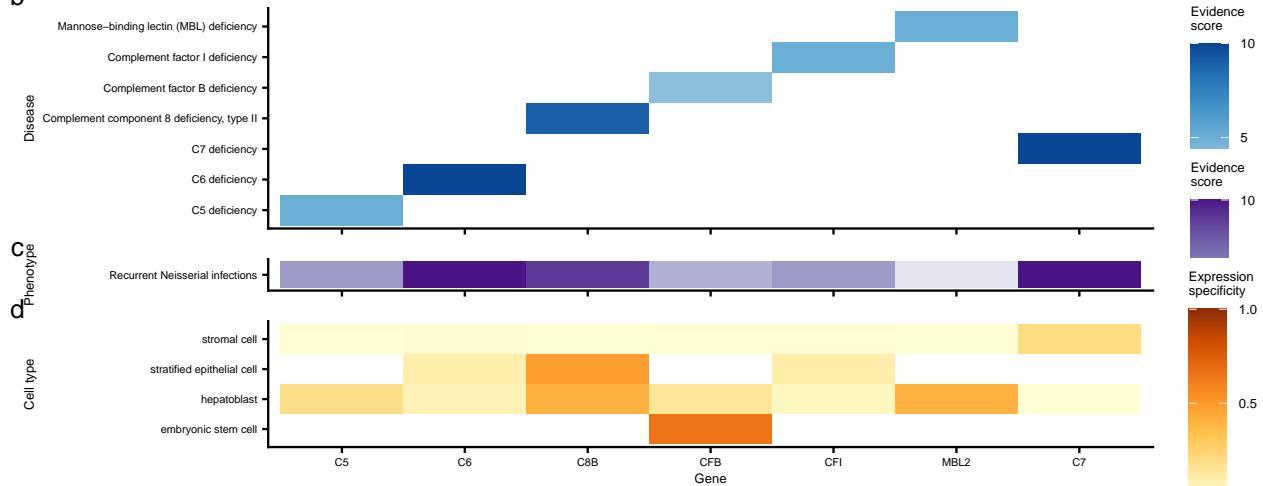
150 Using this network-based approach, we found that RNI (a phenotype of 7 diseases: ‘C5 deficiency’, ‘C6  
151 deficiency’, ‘C7 deficiency’, ‘Complement component 8 deficiency, type II’, ‘Complement factor B deficiency’,  
152 ‘Complement factor I deficiency’, ‘Mannose-Binding lectin deficiency’) was also linked to stromal cells  
153 ( $FDR=4.6 \times 10^{-6}$ ,  $\beta=7.9 \times 10^{-2}$ ), stratified epithelial cells ( $FDR=1.7 \times 10^{-23}$ ,  $\beta=0.15$ ), and embryonic stem  
154 cells ( $FDR=5.4 \times 10^{-5}$ ,  $\beta=7.4 \times 10^{-2}$ ). All of the gene implicated in this causal network are part of the  
155 complement system (*C5*, *C6*, *C8B*, *CFB*, *CFI*, *MBL2*, and *C7*). Complement deficiencies are known to cause  
156 a marked susceptibility to infection<sup>30,31</sup>. However, the complement system comprises more than 56 genes  
157 and can be expressed in a wide variety of cell types<sup>32</sup>. Our analysis was able to decompose the multiple  
158 mechanisms underlying RNI into subsets of complement proteins and identify the specific cell types they each  
159 affect. For example, disruptions in complement genes *C5*, *C8*, and *C7* cause RNI via hepatoblasts, stratified  
160 epithelial cells, and stromal cells, respectively (Fig. 5). These four cell types may represent disease subtypes  
161 with distinct clinical courses or biomarkers, allowing us to begin resolving RNI-related disease mechanisms at  
162 cell-type resolution.

163 In particular, we were intrigued by the finding that RNI was associated with hepatoblasts but not their mature  
164 counterpart, hepatocytes. Mature hepatocytes are well recognised as the principal source of complement  
165 proteins<sup>33</sup>. However, emerging evidence indicates that foetal hepatoblasts also express detectable amounts of  
166 complement genes<sup>35</sup>, which we confirmed by querying the CellxGene browser<sup>36</sup> (see **Data Availability**).  
167 Upon further inspection, we found that these were in fact circulating hepatoblast-like cells found in liver,  
168 placenta and spleen, according to the authors of the Descartes scRNA-seq atlas<sup>37</sup>. These cells express  
169 high levels of alpha fetoprotein (AFB), serum albumin (ALB), and apolipoproteins (including APOE). Our  
170 results support this hypothesis as these AFB+/ALB+ cells were significantly associated with 12 liver-related  
171 phenotypes, as well as 58 blood-related phenotypes. Together, these findings suggest that these hepatoblast-  
172 like cell subpopulations play a causal role in complement-related disorders during development. However,  
173 this relatively novel cell type must be better characterised before drawing any firm conclusions.

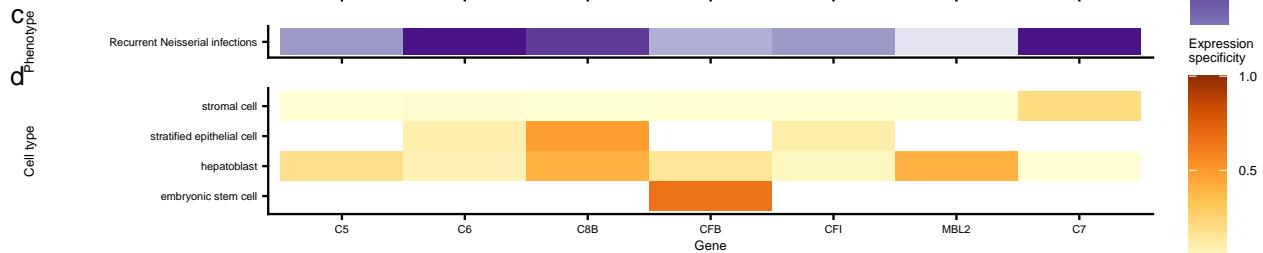
a



b



c



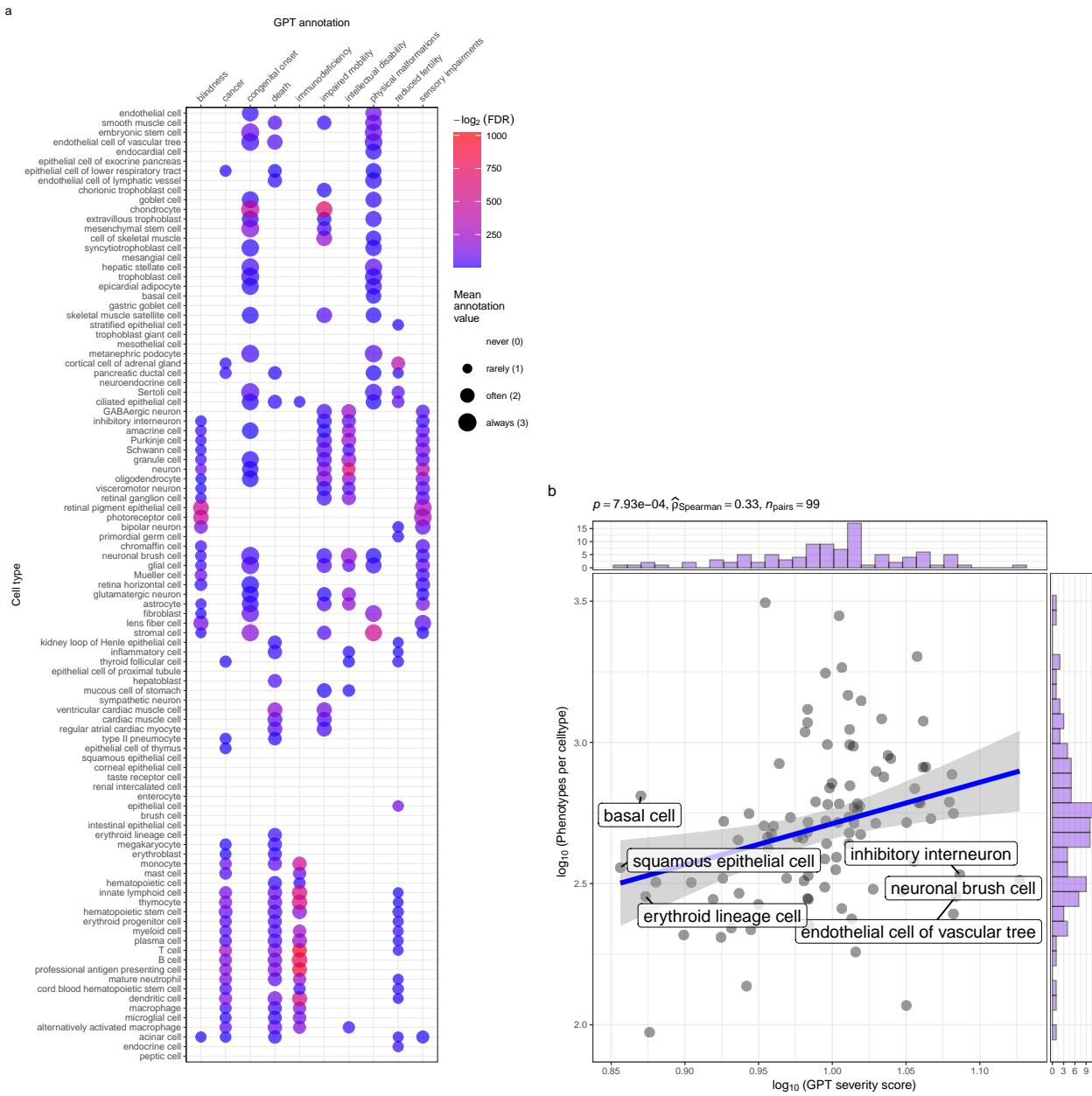
**Causal network of recurrent Neisserial infections (RNI) reveals multi-scale disease etiology.** RNI is a phenotype in seven different monogenic diseases caused by disruptions to specific complement system genes. Four cell types were significantly associated with RNI. **a**, One can trace how genes causal for RNI (yellow boxes, bottom) mediate their effects through cell types (orange circles, middle) and diseases (blue cylinders, top). Cell types are connected to RNI via association testing ( $FDR < 0.05$ ). Genes shown here have both strong evidence for a causal role in RNI and high expression specificity in the associated cell type. Cell types can be linked to monogenic diseases via the genes specifically expressed in those cell types (i.e. are in the top 25% of cell type specificity expression quantiles). Nodes are arranged using the Sugiyama algorithm<sup>38</sup>. **b** Expression specificity quantiles (1-40 scale) of each driver gene in each cell type (darker = greater specificity). **c** GenCC-derived eevidence scores between the RNI phenotype and each gene. **d** Expression specificity (0 = least specific, 1 = most specific) of each gene in each cell type.

Figure 5

<sup>174</sup> **Prioritising phenotypes based on severity**

<sup>175</sup> Some phenotypes are more severe than others and thus could be prioritised for treatment (e.g. ‘Leukonychia’  
<sup>176</sup> is far less severe than ‘Leukodystrophy’). To systematically rank phenotypes, we used GPT-4 to annotate  
<sup>177</sup> severity for 16,982/18,082 (94%) HPO phenotypes<sup>22</sup>. Benchmarking against ground-truth HPO branch  
<sup>178</sup> annotations showed high accuracy (recall=96%, min=89%, max=100%, SD=4.5%) and strong consistency  
<sup>179</sup> (91%). From these, we computed weighted severity scores (0–100) for all phenotypes. The most severe  
<sup>180</sup> was ‘Atrophy/Degeneration affecting the central nervous system’ (*HP:0007367*, score=47), followed by  
<sup>181</sup> ‘Anencephaly’ (*HP:0002323*, score=45). There were 677 phenotypes with score 0 (e.g. ‘Thin toenail’),  
<sup>182</sup> mean=10 (median=9.4).

<sup>183</sup> Merging severity scores with significant (FDR<0.05) phenotype–cell type associations revealed that neuronal  
<sup>184</sup> brush cells had the highest average severity, followed by Mueller cells and glial cells, while megakaryocytes  
<sup>185</sup> had the lowest. Numerically encoding GPT annotations (0–3) and applying Wilcoxon tests confirmed  
<sup>186</sup> expected links, e.g. retinal pigment epithelial cells with blindness, ventricular cardiac muscle cells with death,  
<sup>187</sup> and analogous patterns for reduced fertility, immunodeficiency, impaired mobility, and cancer. Finally, we  
<sup>188</sup> found that cell types associated with more phenotypes also tended to have higher mean composite severity  
<sup>189</sup> ( $p = 7.9 \times 10^{-4}$ ,  $\hat{\rho}_{Spearman} = 0.33$ ), supporting the idea that broadly involved cell types perform critical  
<sup>190</sup> physiological functions whose disruption causes more severe disease.

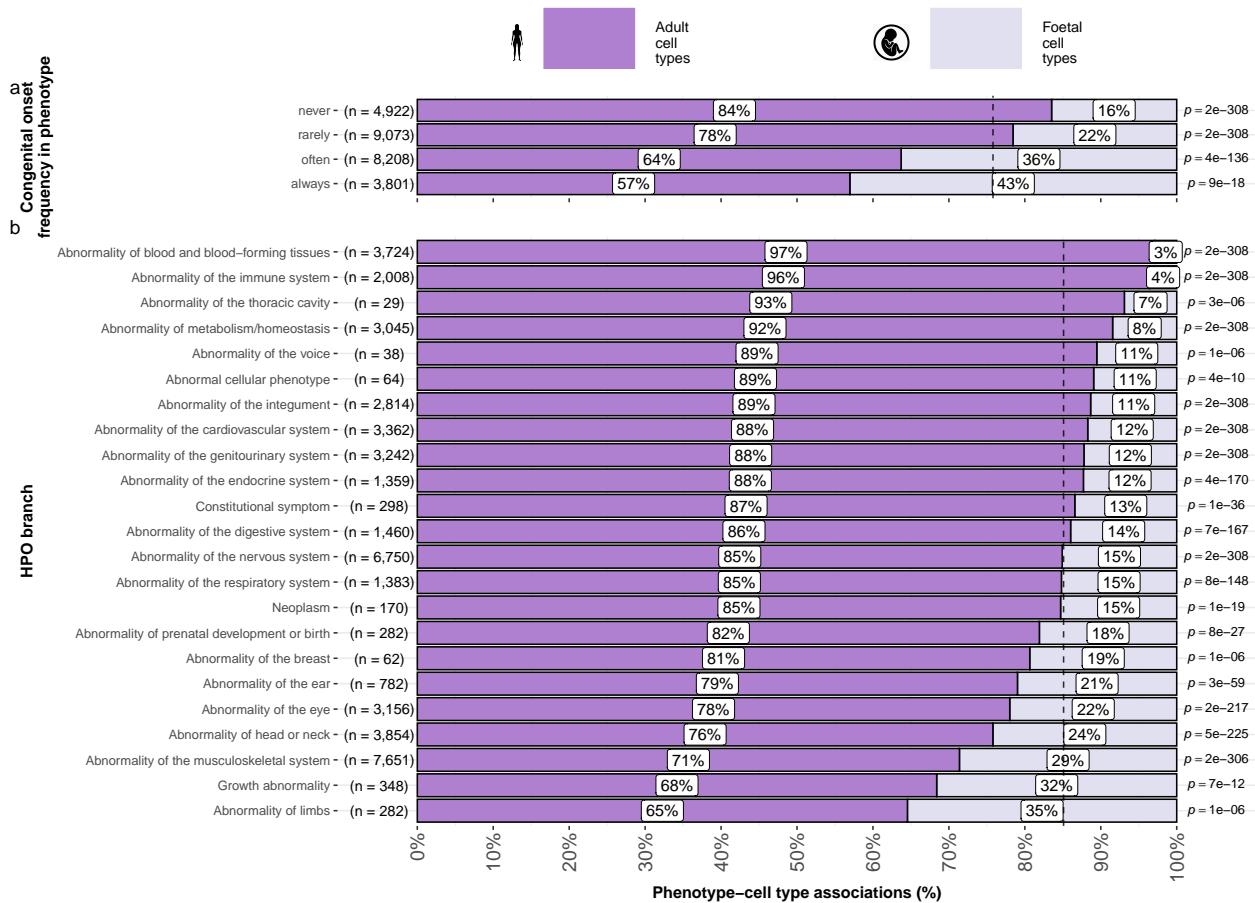


**Genetic disruptions to some cell types cause more clinically severe phenotypes than others.** **a**, Different cell types are associated with different aspects of phenotypic severity. The dot plot shows the mean encoded frequency value for a given severity annotation (0=“never”, 1=“rarely”, 2=“often”, 3=“always”; shown as dot size), aggregated by the associated cell type. One-sided Wilcoxon rank-sum tests were performed for each cell type (within each GPT annotation) to determine which cell types more frequently caused severe phenotypes than all other cell types. Dots are colored by  $-\log_2(FDR)$  when Wilcoxon test FDR values were less than 0.05. All dots with non-significant Wilcoxon tests are not shown. Cell types (rows) are clustered according to the p-values of the Wilcoxon tests. **b**, Cell types that affect more phenotypes tend to have more clinically severe consequences. Specifically, the number of phenotypes each cell type is significantly associated with, and the mean composite severity score of each cell type. The cell types with the top/bottom three x/y axis values are labeled to illustrate the cell types that cause the most/least phenotypic disruption when dysfunctional. Side histograms show the density of data points along each axis. Summary statistics for the linear regression are shown in the title ( $p$  = p-value,  $\hat{\rho}_{Spearman}$  = Spearman rank correlation coefficient,  $n_{pairs}$  = number of observed data pairs).

Figure 6

191 **Congenital phenotypes are associated with foetal cell types**

192 The life stage at which a phenotype manifests affects treatment options, as some interventions (e.g. gene  
 193 therapies) may be ineffective once developmental defects occur. In the DescartesHuman dataset all cells  
 194 were foetal, while the Human Cell Landscape included both embryonic/foetal (29% of cell types), and  
 195 adult tissues (71% of cell types). Some cell types exist in both stages (e.g. chondrocytes), while others are  
 196 foetal-specific (e.g. neural crest cells). Congenital phenotypes (according to our severity annotations) were  
 197 strongly associated with foetal cell types ( $p = 4.7 \times 10^{-261}$ ,  $\chi^2 = 1.2 \times 10^3$ ), consistent with their developmental  
 198 origins (Fig. 7a).



**Foetal vs. adult cell type references provide development context to phenotype etiology.** **a**, 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. **b**, The proportion of phenotype-cell type association tests that are enriched for foetal cell types within each HPO branch. The p-values to the right of 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 dashed vertical line). The foetal silhouette was generated with DALL-E<sup>39</sup>. The adult silhouette is from phylopic.org and is freely available via CC0 1.0 Universal Public Domain Dedication.

Figure 7

199 HPO branches varied significantly in the proportion of their significant associations with foetal cell types

<sup>200</sup> ( $\hat{V}_{Cramer}=0.22$ ,  $p < 2.2 \times 10^{-308}$ , Fig. 7b). Branches with the most disproportionate number of foetal cell  
<sup>201</sup> type associations were ‘Abnormality of limbs’ (35%), ‘Growth abnormality’ (32%), and ‘Abnormality of the  
<sup>202</sup> musculoskeletal system’ (29%). The most adult-biased branches were ‘Abnormality of blood and blood-forming  
<sup>203</sup> tissues’ (97%) and ‘Abnormality of the immune system’ (96%).

<sup>204</sup> Some phenotypes involve only foetal or only adult versions of a cell type. We quantified bias by comparing  
<sup>205</sup> association p-values between foetal and adult versions of the same type (metric range: 1=foetal-only, -  
<sup>206</sup> 1=adult-only). The top 50 foetal-biased phenotypes revealed were enriched for the HPO branches ‘Abnormal  
<sup>207</sup> nasal morphology’ ( $p = 2.4 \times 10^{-7}$ ) and ‘Abnormal external nose morphology’ ( $p = 2.5 \times 10^{-6}$ ), which included  
<sup>208</sup> specific phenotypes such as ‘Abnormal morphology of the nasal alae’ (see Table S9, and Table S10). Adult-  
<sup>209</sup> biased phenotypes were instead enriched for the branches ‘Abnormal elasticity of skin’ ( $p = 3.6 \times 10^{-7}$ ) and  
<sup>210</sup> ‘Abnormality of the cardiovascular system’ ( $p = 4.7 \times 10^{-3}$ ), with examples like ‘Excessive wrinkled skin’ and  
<sup>211</sup> ‘Paroxysmal supraventricular tachycardia’. These align with known developmental and age-related processes,  
<sup>212</sup> supporting our approach for linking phenotypes to causal cell types.

### <sup>213</sup> Therapeutic target identification

<sup>214</sup> In the above sections, we demonstrated how gene association databases can be used to investigate the cell  
<sup>215</sup> types underlying disease phenotypes at scale. While these associations are informative on their own, we  
<sup>216</sup> wished to take these results further in order to have a more translational impact. Knowledge of the causal  
<sup>217</sup> cell types underlying each phenotype can be highly informative for scientists and clinicians in their quest to  
<sup>218</sup> study and treat them. Therapeutic targets with supportive genetic evidence have 2.6x higher success rates  
<sup>219</sup> in clinical trials<sup>40–42</sup>. Furthermore, knowing which cell types to target with gene therapy can maximise the  
<sup>220</sup> efficacy of highly expensive payloads, and minimise side effects (e.g. immune reaction to viral vectors). Recent  
<sup>221</sup> biotechnological advances have greatly enhanced our ability to target specific cell types with gene therapy,  
<sup>222</sup> making specific and accurate knowledge the correct underlying cell types more pertinent than ever<sup>20,21</sup>.

<sup>223</sup> Rather than consider phenotypes in isolation, or even phenotype associations with particular cell types, we  
<sup>224</sup> sought to identify multi-scale therapeutic targets. That is, specific genes to target in specific cell types in  
<sup>225</sup> specific phenotypes as they present in particular diseases. The confluence of these pieces of information  
<sup>226</sup> is crucial for clinical utility, as it provides the much-needed context to develop effective therapeutics for  
<sup>227</sup> real-world patient populations. For example, the same phenotype may be caused by disruptions to one of  
<sup>228</sup> several cell types, each of which are in turn caused by mutations to particular genes. Networks are a natural  
<sup>229</sup> way to visualize the complex relationships between these various biological entities, and using sensible filters  
<sup>230</sup> (i.e. pruning) keeps the networks small enough to gain meaningful insights through visual exploration.

<sup>231</sup> Towards this objective, we developed an automated pipeline to identify putative cell type-specific gene  
<sup>232</sup> targets for each phenotype by integrating phenotype-cell type association results with primary resources  
<sup>233</sup> such as GenCC gene-disease relationships and scRNA-seq atlas datasets, producing a table where each row

234 represented a disease–phenotype–cell type–gene tetrad. We applied sequential filters to retain only significant  
235 phenotype–cell type associations ( $FDR < 0.05$ ), phenotype–gene pairs with strong causal evidence (GenCC  
236 score  $> 3$ ), phenotypes with high specificity ( $IC > 8$ ), and gene–cell type links in the top 25% expression  
237 specificity quantile, and further required a symptom intersection  $> 0.25$  when linking cell types to diseases  
238 via phenotypes. The filtered results were ranked by GPT-4 composite severity scores, with only the top 10  
239 tetrads retained per phenotype, yielding compact, high-confidence networks suitable for manual inspection  
240 and visualization.

241 This yielded putative therapeutic targets for 5,252 phenotypes across 4,819 diseases in 201 cell types and  
242 3,148 genes (Fig. S6). While this constitutes a large number of genes in total, each phenotype was assigned a  
243 median of 2 gene targets (mean=3.3, min=1, max=10). Relative to the number of genes annotations per  
244 phenotype in the HPO overall (median=7.0, mean=62, min=1, max=5,003) this represents a substantial  
245 decrease in the number of candidate target genes, even when excluding high-level phenotypes (HPO level $>3$ ).  
246 It is also important to note that the phenotypes in the prioritised targets list are ranked by their severity,  
247 allowing us to distinguish between phenotypes with a high medical urgency (e.g. ‘Hydranencephaly’) from  
248 those with lower medical urgency (e.g. ‘Fair hair’). This can be useful for clinicians, biomedical scientists,  
249 and pharmaceutical manufacturers who wish to focus their research efforts on phenotypes with the greatest  
250 need for intervention.

251 Across all phenotypes, epithelial cells were most commonly implicated (838 phenotypes), followed by stromal  
252 cells (626 phenotypes), neurons (475 phenotypes), chondrocytes (383 phenotypes), endothelial cells (361  
253 phenotypes), and fibroblasts (348 phenotypes). Grouped by higher-order ontology category, ‘Abnormality of  
254 the musculoskeletal system’ had the greatest number of enriched phenotypes (959 phenotypes, 857 genes),  
255 followed by ‘Abnormality of the nervous system’ (733 phenotypes, 1,138 genes), ‘Abnormality of head or  
256 neck’ (543 phenotypes, 986 genes), ‘Abnormality of the genitourinary system’ (443 phenotypes, 695 genes),  
257 and ‘Abnormality of the eye’ (377 phenotypes, 545 genes).

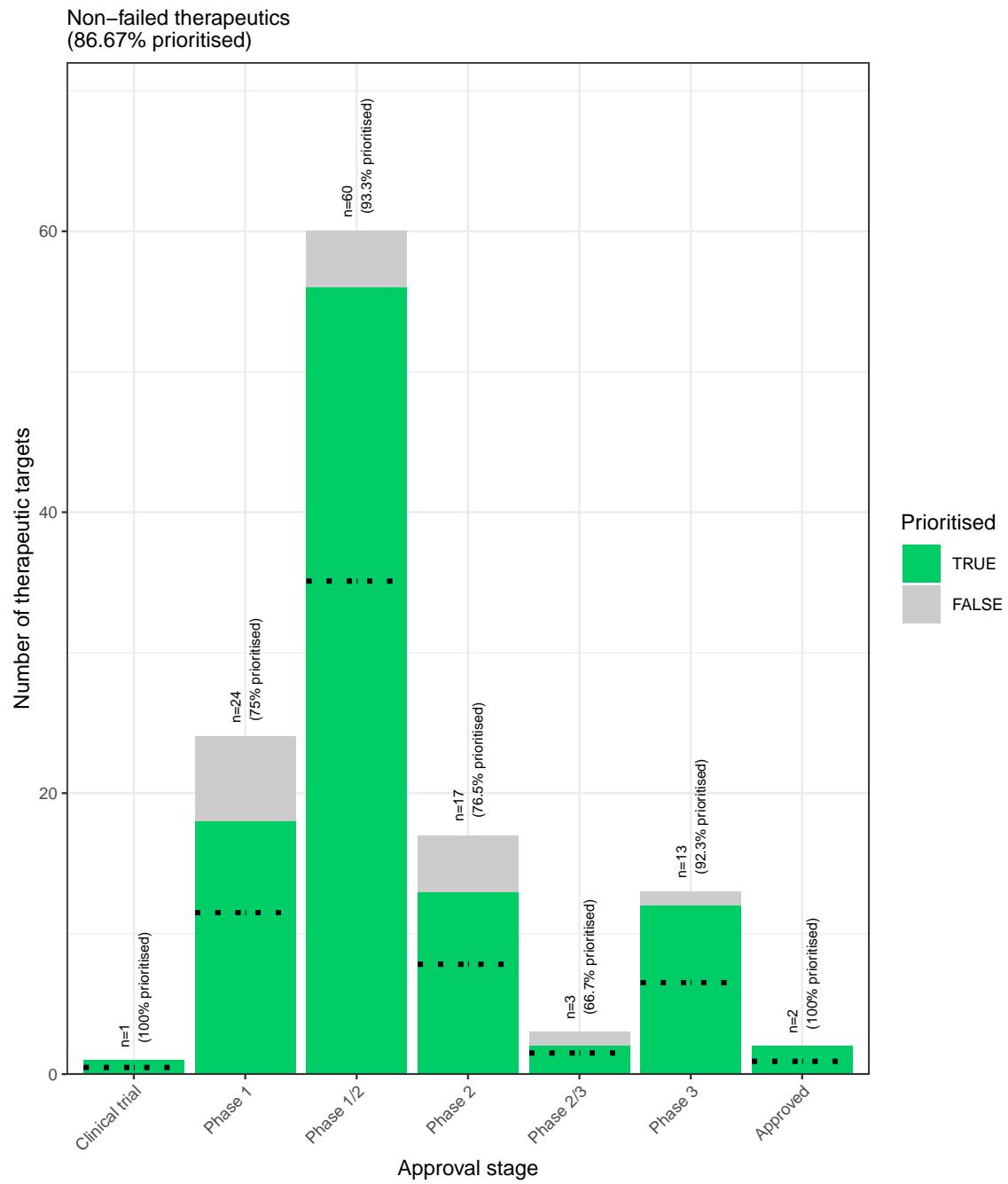
## 258 Therapeutic target validation

259 To determine whether the genes prioritised by our therapeutic targets pipeline were plausible, we checked  
260 what percentage of existing gene therapy targets we recapitulated. Data on therapeutic approval status  
261 was gathered from the Therapeutic Target Database (TTD; release 2026-02-25)<sup>43</sup>. Overall, we prioritised  
262 87% (120 total) of all non-failed existing gene therapy targets (ie. those which are currently approved,  
263 investigative, or undergoing clinical trials). A hypergeometric test confirmed that our prioritised targets were  
264 significantly enriched for non-failed gene therapy targets ( $p = 5.6 \times 10^{-5}$ , odds ratio=3.0, sensitivity=0.83,  
265 specificity=0.38, Table S11). For these hypergeometric tests, the background gene set was composed of the  
266 union of all phenotype-associated genes in the HPO and all gene therapy targets listed in TTD.

267 Even when considering therapeutics of any kind (Fig. S7), not just gene therapies, we recapitulated 40%

268 of the non-failed therapeutic targets and 0% of the terminated/withdrawn therapeutic targets (n=1,255).  
269 Here we found that our prioritised targets were significantly depleted for failed therapeutics ( $p = 4.4 \times 10^{-23}$ ,  
270 odds ratio=0.36, sensitivity=0.27, specificity=0.49). This suggests that our multi-scale evidence-based  
271 prioritisation pipeline is capable of selectively identifying genes that are likely to be effective therapeutic  
272 targets.

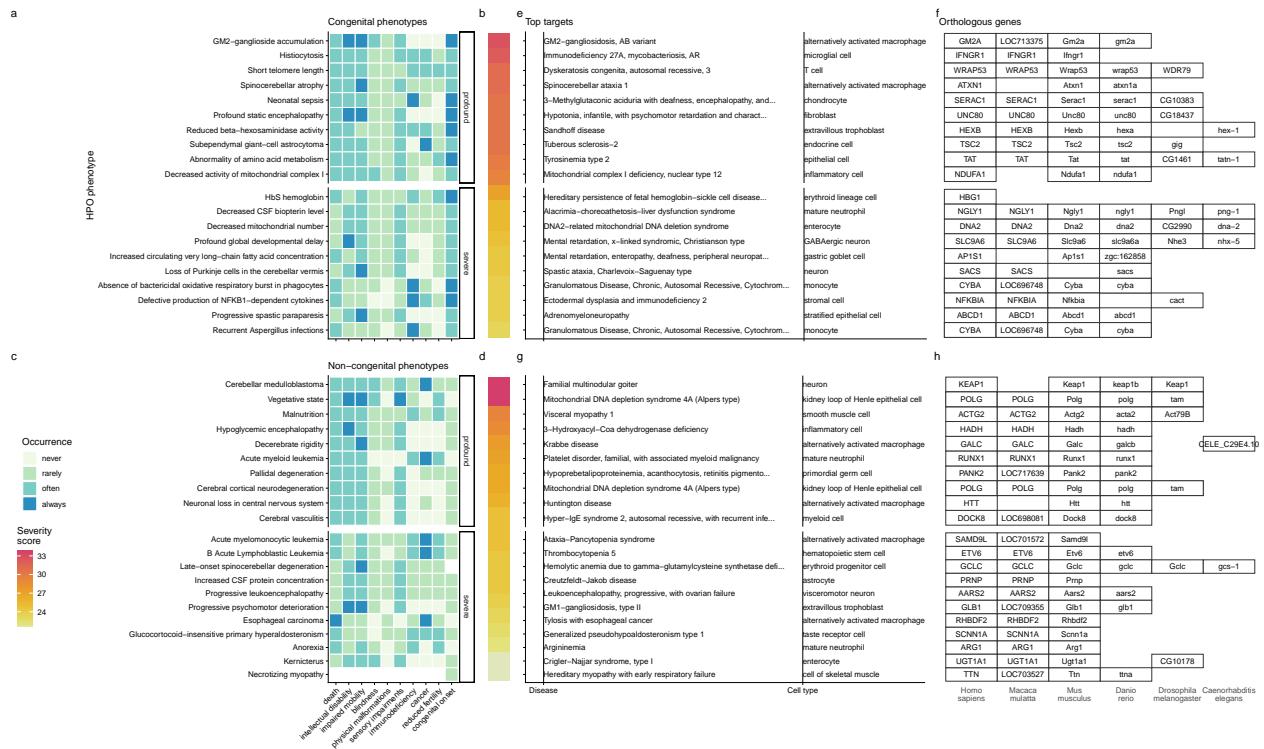
273 In addition to aggregate enrichment results, we also provide specific examples of successful gene therapies  
274 whose cell type-specific mechanism were recapitulated by our phenotype-cell associations. In particular, our  
275 pipeline nominated the gene *RPE65* within ‘retinal pigment epithelial cells’ as the top target for ‘Fundus  
276 atrophy’ vision-related phenotypes that are hallmarks of ‘Leber congenital amaurosis, type II’ and ‘Severe  
277 early-childhood-onset retinal dystrophy’. Indeed, gene therapies targeting *RPE65* within the retina of patients  
278 with these rare genetic conditions are some of the most successful clinical applications of this technology  
279 to date, able to restore vision in many cases<sup>44</sup>. In other cases, a tissue (e.g. liver) may be known to be  
280 causally involved in disease genesis, but the precise causal cell types within that tissue remain unknown  
281 (e.g. hepatocytes, Kupffer cells, Cholangiocytes, Hepatic stellate cells, Natural killer cells, etc.). Tissue-level  
282 investigations (e.g. using bulk transcriptomics or epigenomics) would be dominated by hepatocytes, which  
283 comprise 75% of the liver. Our prioritized gene therapy targets can aid in such scenarios by providing the  
284 cell type-resolution context most likely to be causal for a given phenotype or set of phenotypes.



**Prioritised targets recapitulate existing gene therapy 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. While our prioritized targets did not include any failed ('Terminated') therapies, the fact that only one such therapy exists in the dataset preclude us from making any conclusions about depletion of failed gene therapy targets in our prioritised targets list. Black dashed lines indicate the number of prioritised targets expected by chance.

Figure 8

285 **Selected example targets**



**Evidence-based pipeline nominates causal mechanisms to target for gene therapy.** Shown here are the 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. See supplement Fig. S9 for network plots of cell type-specific gene therapy targets for several severe phenotypes and their associated diseases.

Figure 9

- 286 From our prioritised targets, we selected four phenotype or disease examples: ‘GM2-ganglioside accumulation’,  
 287 ‘Spinocerebellar atrophy’, ‘Neuronal loss in central nervous system’. To focus on clinically relevant phenotypes  
 288 and reduce overplotting, we limited selection to those with GPT severity scores above 15 (Fig. S9). Selection  
 289 was based on severity and network simplicity to allow compact visualisation.
- 290 Tay-Sachs disease (TSD) is a fatal neurodegenerative condition caused by *HEXA* deficiency and ganglioside  
 291 buildup. We identified alternatively activated macrophages as the cell type most associated with ‘GM2-  
 292 ganglioside accumulation’ (Fig. S9). This aligns with prior findings of ganglioside accumulation in TSD  
 293 macrophages<sup>45,46,47,48</sup>. Our results support macrophages as causal in TSD and the most promising therapeutic  
 294 target.
- 295 Spinocerebellar atrophy is a progressive neurodegenerative phenotype in disorders like Spinocerebellar ataxia.

296 Our pipeline implicates M2 macrophages ('Alternatively activated macrophages') as the only causal cell  
297 type (Fig. S9). This suggests Purkinje cell loss is downstream of macrophage dysfunction, consistent with  
298 microglial roles in neurodegeneration<sup>49–51</sup>. Our findings provide the first statistically supported link between  
299 risk genes and this cell type, which is supported by relevant mouse models (e.g. *Atxn1*, *Pnpla6*) that replicate  
300 cellular and behavioural disease phenotypes.

301 'Neuronal loss in the central nervous system' is a phenotype by multiple serious diseases (e.g. Huntington  
302 disease, frontotemporal lobar degeneration, and certain mitochondrial disorders). Across all of these diverse  
303 conditions with varying genetic causes (>8 genes), these conditions converge on just 2 cell types: M2  
304 macrophages and epithelial cells.

305 Additional examples of therapeutics targets include; cardiac muscle and endothelial cells in phenotypes  
306 associated with respiratory failure (Fig. S10a), microglia in frontal lobe dementia (Fig. S11), chondrocytes in  
307 lethal skeletal dysplasia (Fig. S12), endothelial cells in small vessel disease (Fig. S13), oligodendrocytes and  
308 neurons in Parkinson's disease (Fig. S14). and multiple gastrointestinal and immune cell types in Alzheimer's  
309 disease (Fig. S15). For a more extensive discussion of these phenotypes, please refer to the Supplementary  
310 Results.

### 311 **Mappings**

312 Mappings from HPO phenotypes and other commonly used medical ontologies (SNOMED, UMLS, ICD-9,  
313 and ICD-10) were gathered using the Ontology Xref Service (Oxo; <https://www.ebi.ac.uk/spot/oxo/>) to  
314 facilitate others using our results in future work. Direct mappings, with a cross-ontology distance of 1, are  
315 the most precise and reliable. Counts of mappings at each distance are shown in Table S1. In total, there  
316 were 15,105 direct mappings between the HPO and other ontologies, with the largest number of mappings  
317 coming from the UMLS ontology (12,898 UMLS terms).

### 318 **Discussion**

319 Investigating rare diseases (RDs) at the phenotype level offers advantages in research and clinical medicine.  
320 Most RDs have a single known causal gene (7,671/8,631 = 89%), prohibiting traditional gene set enrichment-  
321 based analyses. Instead, aggregating genes into phenotype-based sets enables well-powered analyses (mean  
322 ~76 genes/phenotype). Phenotypes often converge on shared molecular pathways, and a phenotype-centric  
323 approach captures interindividual variation in disease presentations. This requires mapping the molecular  
324 and cellular mechanisms behind each phenotype, which we achieve here at genome scale.

325 Across 201 cell types and 11,047 phenotypes, we found >46,514 significant phenotype–cell type relationships,  
326 enabling multi-scale mechanistic tracing of disease biology. Results replicate known links, add cellular  
327 context, and uncover novel associations. Extensive benchmarking confirmed expected associations, aided by  
328 comprehensive phenotype and cell type ontologies. Key findings include enrichment of anatomically matched

329 associations, correlation of phenotype specificity with association strength, precise subtypes for recurrent  
330 infections, and links between congenital onset frequency and developmental cell types.

331 It is a matter of ongoing scientific debate as to whether the rare, more monogenic versions of diseases and  
332 phenotypes share the same genetic etiology as their common, polygenic counterparts<sup>53</sup>. Nevertheless, we  
333 identified several examples where these two disparate data sources converge on the same cell types. These  
334 include the implication of smooth muscle cells and endothelial cells in small vessel disease<sup>56</sup>, and the consistent  
335 association between macrophages and neurodegenerative diseases such as Alzheimer's and Parkinson's<sup>54</sup>.  
336 Additionally, we recover genome-wide association study (GWAS) -derived associations between macular  
337 degeneration and specific cell types of the eye (photoreceptor cells, retinal pigment cells, Mueller cells),  
338 immune system (dendritic cells, macrophages), and vascular system (endothelial cells)<sup>64</sup>.

339 Despite our growing knowledge of RD genetics, less than 5% of RDs have treatments<sup>6</sup>. However, advances in  
340 CRISPR, prime editing, antisense oligonucleotides, viral/lipid delivery<sup>65–68</sup> are accelerating. The FDA's new  
341 program<sup>69</sup> aims to expand gene/cell therapy approvals in years rather than decades<sup>70</sup>, but success depends  
342 on understanding the causal mechanisms of each RD. Here, we built a reproducible pipeline for nominating  
343 cell type-resolved therapeutic targets (Fig. 9), factoring in association strength, gene specificity, severity,  
344 therapy delivery suitability, and model translatability. We recovered 87% of active gene therapies, confirming  
345 strong enrichment. Highlighted cases include macrophage-driven phenotypes in Tay-Sachs, spinocerebellar  
346 ataxia, and Alzheimer's disease, pinpointing links between specific phenotypes (e.g. neurofibrillary tangles)  
347 and causal cell types.

348 Current limitations of our study include missing certain rare cell subtypes and states (e.g. immune cell  
349 responses, diseased states, aging) and incomplete knowledge of gene–phenotype associations. A related  
350 concern is that our statistical power to detect cell type associations may be diminished in phenotypes with very  
351 few gene associations. Fewer genes decreases variance explained with respect to the denser cell type specificity  
352 vectors, in turn decreasing power. The remaining associations for the most specific phenotypes therefore  
353 reflect cases where the effect size is sufficiently strong to overcome the reduced variance (Fig. 3). Therefore,  
354 weaker cell type associations with highly monogenic phenotypes may be missed by our analyses. With the  
355 expectation that data will continue to improve over time, our pipeline is fully containerised and documented  
356 for end-to-end reproducibility. Comprehensive, ontology-aligned frameworks such as ours enable discovery,  
357 diagnosis, and basket trial design for shared molecular etiologies across many diseases<sup>71</sup>. Furthermore, we  
358 invite collaborations to validate and translate these predictions, and have publicly released all results via  
359 R packages and the Rare Disease Celltyping Portal (<https://neurogenomics-ukdri.dsi.ic.ac.uk/>) to support  
360 broad access for researchers, clinicians, and patients.

361 In summary, we present a scalable, cost-effective, and reproducible method for genome-wide, cell type-  
362 specific mechanism prediction in RDs. With advances in gene therapy and supportive regulatory changes,  
363 this approach can help realise the promise of genomic medicine for the global RD community.

364 **Methods**

365 **Human Phenotype Ontology**

366 The latest version of the HPO (release 2024-02-08) was downloaded from the EMBL-EBI Ontology Lookup  
367 Service<sup>72</sup> and imported into R using the `HPOExplorer` package. This R object was used to extract ontological

368 relationships between phenotypes as well as to assign absolute and relative ontological levels to each phenotype.

369 The latest version of the HPO phenotype-to-gene mappings and phenotype annotations were downloaded  
370 from the official HPO GitHub repository and imported into R using `HPOExplorer`. This contains lists of genes

371 associated with phenotypes via particular diseases, formatted as three columns in a table (gene, phenotype,  
372 disease).

373 However, not all genes have equally strong evidence of causality with a disease or phenotype, especially when  
374 considering that the variety of resources used to generate these annotations (OMIM, Orphanet, DECIPHER)  
375 use variable methodologies (e.g. expert-curated review of the medical literature vs. automated text mining of  
376 the literature). Because of this, formalizing phenotypes (or any biological entity) as unweighted gene sets can  
377 lead to loss of information about the most relevant signals. This is especially true when gene sets become  
378 large and poorly supported, essentially decreasing the signal-to-noise ratio<sup>75</sup>.

379 Therefore we imported data from the Gene Curation Coalition (GenCC)<sup>76,77</sup>, which (as of 2025-11-30)  
380 24,124 evidence scores across 7,566 diseases and 5,533 genes. Evidence scores are defined by GenCC using a  
381 standardised ordinal rubric which we then encoded as a semi-quantitative score ranging from 0 (no evidence  
382 of disease-gene relationship) to 6 (strongest evidence of disease-gene relationship) (see Table S5). As each  
383 Disease-Gene pair can have multiple entries (from different studies) with different levels of evidence, we then  
384 summed evidence scores per Disease-Gene pair to generate aggregated Disease-by-Gene evidence scores. This  
385 procedure can be described as follows.

386 Let us denote:

387 •  $D$  as diseases.

388 •  $P$  as phenotypes in the HPO.

389 •  $G$  as genes

390 •  $S$  as the evidence scores describing the strength of the relationship between each Disease-Gene pair.

391 •  $M_{ij}$  as the aggregated Disease-by-Gene evidence score matrix.

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

392 Next, we extracted Disease-Gene-Phenotype relationships from the annotations file distributed by the HPO

393 (*phenotype\_to\_genes.txt*). This provides a list of genes associated with phenotypes via particular diseases,  
 394 but does not include any strength of evidence scores.

395 Here we define:

- 396 •  $A_{ijk}$  as the Disease-Gene-Phenotype relationships.  
 397 •  $D_i$  as the  $i$ th disease.  
 398 •  $G_j$  as the  $j$ th gene.  
 399 •  $P_k$  as the  $k$ th phenotype.

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

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

404

405

406

Tensor of Disease-by-Gene  
evidence scores

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

407

408

#### 409 Construction of the tensor of Phenotype-by-Gene evidence scores.

410 Histograms of evidence score distributions at each step in processing can be found in Fig. S1.

#### 411 Single-cell transcriptomic atlases

412 In this study, the gene by cell type specificity matrix was constructed using the Descartes Human transcriptome  
 413 atlas of foetal gene expression, which contains a mixture of single-nucleus and single-cell RNA-seq data  
 414 (collected with sci-RNA-seq3)<sup>18</sup>. This dataset contains 377,456 cells representing 77 distinct cell types across  
 415 15 tissues. All 121 human foetal samples ranged from 72 to 129 days in estimated postconceptual age. To  
 416 independently replicate our findings, we also used the Human Cell Landscape which contains single-cell

417 transcriptomic data (collected with microwell-seq) from embryonic, foetal, and adult human samples across  
418 49 tissues<sup>19</sup>.

419 Specificity matrices were generated separately for each transcriptomic atlas using the R package EWCE  
420 (v1.11.3)<sup>78</sup>. Within each atlas, cell types were defined using the authors' original freeform annotations in order  
421 to preserve the granularity of cell subtypes as well as incorporate expert-identified rare cell types. Cell types  
422 were only aligned and aggregated to the level of corresponding Cell Ontology (CL)<sup>24</sup> annotations afterwards  
423 when generating summary figures and performing cross-atlas analyses. Using the original gene-by-cell count  
424 matrices from each single-cell atlas, we computed gene-by-cell type expression specificity matrices as follows.  
425 Genes with very no expression across any cell types were considered to be uninformative and were therefore  
426 removed from the input gene-by-cell matrix  $F(g, i, c)$ .

427 Next, we calculated the mean expression per cell type and normalised the resulting matrix to transform it  
428 into a gene-by-cell type expression specificity matrix ( $S_{g,c}$ ). In other words, each gene in each cell type had a  
429 0-1 score where 1 indicated the gene was mostly specifically expressed in that particular cell type relative to  
430 all other cell types. This procedure was repeated separately for each of the single-cell atlases and can be  
431 summarised as:

432

433

Compute mean expression of each gene per cell type

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

Gene-by-cell type specificity matrix

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

↓

$\sum_{i=1}^{|L|} F_{gic}$

$N_c$

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

434

435

#### 436 Phenotype-cell type associations

437 To test for relationships between each pairwise combination of phenotype (n=11,047) and cell type (n=201)  
438 we ran a series of univariate generalised linear models implemented via the `MSTExplorer::run_phenomix`  
439 function in R (which uses `stats::glm` internally).

440 First, we filtered the gene-by-phenotype evidence score matrix ( $L_{ij}$ ) and the gene-by-cell type expression  
441 specificity matrix ( $S_{gc}$ ) to only include genes present in both matrices (n=4,949 genes in the Descartes Human  
442 analyses; n=4,653 genes in the Human Cell Landscape analyses). Then, within each matrix any rows or  
443 columns with a sum of 0 were removed as these were uninformative data points that did not vary. To improve  
444 interpretability of the results  $\beta$  coefficient estimates across models (i.e. effect size), we performed a scaling  
445 prestep on all dependent and independent variables. Initial tests showed that this had virtually no impact

446 on the total number of significant results or any of the benchmarking metrics based on p-value thresholds  
447 (Fig. 2). This scaling prestep improved our ability to rank cell types by the strength of their association with  
448 a given phenotype as determined by separate linear models.

449 We repeated the aforementioned procedure separately for each of the single-cell references. Once all results  
450 were generated using both cell type references (2,206,994 association tests total), we applied Benjamini-  
451 Hochberg false discovery rate<sup>79</sup> (denoted as  $FDR_{pc}$ ) to account for multiple testing. Of note, we applied this  
452 correction across all results at once (as opposed to each single-cell reference separately) to ensure the  $FDR_{pc}$   
453 was stringently controlled for across all tests performed in this study.

454 For the significant phenotype-cell type associations, we also wished to identify which genes were most strongly  
455 driving this association. We therefore designed a heuristic to consistently extract such *driver genes*. For a  
456 given phenotype-cell type pair, *driver genes* were defined as the intersect of genes that had a phenotype  
457 evidence score  $>0$  and were within the top 75th expression specificity percentile (quantiles 30-40 out of 40)  
458 for the associated cell type. This be described as follows.

459 Let

- 460 •  $G$  be the full set of genes,  
461 •  $s_i$  be the phenotype evidence score for gene  $i$ ,  
462 •  $e_i$  be the expression specificity of gene  $i$  in the associated cell type,  
463 •  $q(e_i)$  be the empirical specificity percentile of  $e_i$  (ranging from 1 to 40).

464 We define two subsets:

$$A = \{ i \in G : s_i > 0 \},$$

$$B = \{ i \in G : q(e_i) \geq 30 \}.$$

466 The set of driver genes for that phenotype-cell type pair is then

$$D = A \cap B.$$

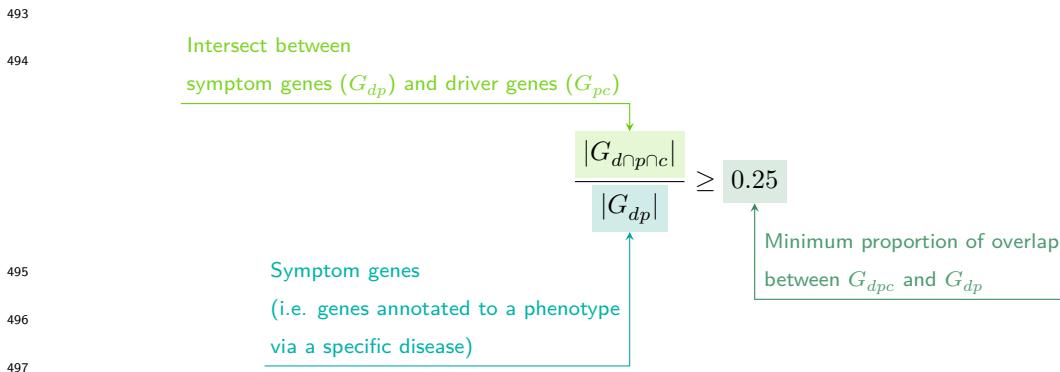
467 Framing association testing as a regression problem rather than a gene set enrichment problem has additional  
468 benefits, including speed. This approach allowed us to complete all tests within ~30 min on a MacBook Pro  
469 with 8 CPUs, while early tests indicated that gene set enrichment-based methods like EWCE<sup>80</sup> would take days to  
470 weeks on a high-performance computing cluster to complete the same number of tests due to its computationally  
471 expensive bootstrapping procedure. These tests also showed that EWCE was unable to recover nearly as many  
472 true positive results (as indicated by our MKG benchmark) with a comparable degree of precision (see related  
473 Issue for further details: [https://github.com/neurogenomics/rare\\_disease\\_celltyping/issues/51](https://github.com/neurogenomics/rare_disease_celltyping/issues/51)).

474 A programmatic implementation of this heuristic can be found in the `MSTExplorer::add_driver_genes` R

475 function.

## 476 Symptom-cell type associations

477 In the HPO, a phenotype can be associated with multiple diseases via different subsets of genes (Fig. S2).  
478 Here we define a symptom as a phenotype as it presents within the context of the specific disease. A  
479 symptom's gene set is therefore the subset of genes connecting a phenotype to a specific disease. For example,  
480 the phenotype 'Headache' (*HP:0002315*) is associated with 346 genes across different 192 diseases. Our  
481 association tests described above use evidence scores from all 346 of these genes when testing for a relationship  
482 between 'Headache' and various cell types. However, the disease 'Erythrocytosis, familial, 1' (*OMIM:133100*)  
483 is linked to the 'Headache' phenotype via a single gene; *EPOR*. Thus, the symptom 'Heachache due to  
484 Erythrocytosis, familial, 1' would have only *EPOR* as its gene list. Another disease, 'Migraine with or  
485 without aura, susceptibility to, 1' (*OMIM:157300*) is linked with 'Headache' via 3 different genes (*ESR1*,  
486 *TNF*, *EDNRA*), so the symptom 'Heachache' in the presence of this disease would consist of these 3 genes.  
487 More formally, the features of a given symptom can be described as the subset of genes annotated to phenotype  
488  $p$  via a particular disease  $d$ , denoted as  $G_{dp}$  (see Fig. S2). We then computed the intersect between symptom  
489 genes ( $G_{dp}$ ) and driver genes (described in the previous section, and denoted as  $G_{pc}$ ), resulting in the gene  
490 subset  $G_{d \cap p \cap c}$ . Only  $G_{d \cap p \cap c}$  gene sets with 25% or greater overlap with the symptom gene subset ( $G_{dp}$ ) were  
491 kept. This procedure was repeated for all phenotype-cell type-disease triads, which can be summarised as  
492 follows:



## 498 Validation of expected phenotype-cell type relationships

499 We first sought to confirm that our tests (across both single-cell references) were able to recover expected  
500 phenotype-cell type relationships across seven high-level branches within the HPO (Fig. 2), including  
501 abnormalities of the cardiovascular system, endocrine system, eye, immune system, musculoskeletal system,  
502 nervous system, and respiratory system. Within each branch the number of significant tests in a given cell  
503 type were plotted (Fig. 2b). Mappings between freeform annotations (the level at which we performed our  
504 phenotype-cell type association tests) provided by the original atlas authors and their closest CL term

505 equivalents were provided by CellxGene<sup>36</sup>. CL terms along the *x-axis* of Fig. 2b were assigned colours  
506 corresponding to which HPO branch showed the greatest number of enrichments (after normalising within  
507 each branch to account for differences in scale). The normalised colouring allows readers to quickly assess  
508 which HPO branch was most often associated with each cell type, while accounting for differences in the  
509 number of phenotypes across branches. We then ran a series of Analysis of Variance (ANOVA) tests to  
510 determine whether (within a given branch) a given cell type was more often enriched ( $FDR < 0.05$ ) within that  
511 branch relative to all of the other HPO branches of an equivalent level in the ontology (including all branches  
512 not shown in Fig. 2b). After applying Benjamini-Hochberg multiple testing correction<sup>79</sup> (denoted as  $FDR_{b,c}$ ),  
513 we annotated each respective branch-by-cell type bar according to the significance (\*\*\*\* :  $FDR_{b,c} < 1e - 04$ ,  
514 \*\*\* :  $FDR_{b,c} < 0.001$ , \*\* :  $FDR_{b,c} < 0.01$ , \* :  $FDR_{b,c} < 0.05$ ). Cell types in Fig. 2a-b were ordered along  
515 the *x-axis* according to a dendrogram derived from the CL ontology (Fig. 2c), which provides ground-truth  
516 semantic relationships between all cell types (e.g. different neuronal subtypes are grouped together).

517 As an additional measure of the accuracy of our phenotype-cell types test results we identified conceptually  
518 matched branches across the HPO and the CL (Fig. 2d and Table S6). For example, ‘Abnormality of the  
519 cardiovascular system’ in the HPO was matched with ‘cardiocytes’ in the CL which includes all cell types  
520 specific to the heart. Analogously, ‘Abnormality of the nervous system’ in the HPO was matched with ‘neural  
521 cell’ in the CL which includes all descendant subtypes of neurons and glia. This cross-ontology matching  
522 was repeated for each HPO branch and can be referred to as on-target cell types. Within each branch, the  
523  $-\log_{10}(FDR_{pc})$  values of on-target cell types were binned by rounding to the nearest integer (*x-axis*) and  
524 the percentage of tests for on-target cell types relative to all cell types were computed at each bin (*y-axis*)  
525 (Fig. 2d). The baseline level (dotted horizontal line) illustrates the percentage of on-target cell types relative  
526 to the total number of observed cell types. Any percentages above this baseline level represent greater than  
527 chance representation of the on-target cell types in the significant tests.

## 528 Validation of inter- and intra-dataset consistency

529 We tested for inter-dataset consistency of our phenotype-cell type association results across different single-  
530 cell reference datasets (Descartes Human and Human Cell Landscape). For association tests with exactly  
531 matching Cell Ontology ID across the two references, we tested for a relationship between the effect  
532 sizes (each GLM model’s  $R^2$  estimates) generated with each of the references using Pearson correlation  
533 (`stats::cor.test(method="pearson")`). We repeated this correlation analysis using only the phenotype-cell  
534 type associations that were significant ( $FDR < 0.05$ ) in both reference datasets.

535 We also tested for intra-dataset consistency within the Human Cell Landscape by running additional  
536 correlations between the phenotype-cell type association test statistics of the foetal and the adult samples  
537 (again using estimates from all results, and significant-only results at  $FDR < 0.05$ ).

538 To confirm the correlation findings, we repeated these analyses using a permutation procedure (1,000

539 permutations) in which we generate a null distribution by randomly shuffling the IC values across phenotypes.  
540 We then derive a two-sided empirical p-value by calculating the proportion of the null distribution's correlation  
541 coefficient that were greater than or equal to the observed correlation coefficient. This procedure is implemented  
542 in the function `MSTExplorer::run_permutation_test_cor`.

543 Finally, we compute the symmetric replication rate within each of the comparison described above (between  
544 scRNA-seq references, and between developmental stages within the Human Cell Landscape). This was  
545 defined as the proportion of significant results ( $FDR < 0.05$ ) in dataset A that were also significant ( $FDR <$   
546  $0.05$ ) in dataset B, and vice versa, averaged across both directions. This can be described as follows.

547 **Variable descriptions**

- 548 •  $S_A$ : The set of phenotype–cell-type associations that pass the significance threshold (e.g.,  $FDR < 0.05$ )  
549 in dataset  $A$ .
- 550 •  $S_B$ : The set of phenotype–cell-type associations that pass the same significance threshold in dataset  $B$ .
- 551 •  $S_A \cap S_B$ : The set of associations that are significant in *both* datasets.
- 552 •  $RR$ : Replication rate.

553 **Directional replication rate ( $A \rightarrow B$ )**

$$RR_{A \rightarrow B} = \frac{|S_A \cap S_B|}{|S_A|}$$

554 **Directional replication rate ( $B \rightarrow A$ )**

$$RR_{B \rightarrow A} = \frac{|S_A \cap S_B|}{|S_B|}$$

555 **Symmetric replication rate**

$$RR_{sym} = \frac{(RR_{A \rightarrow B} + RR_{B \rightarrow A})}{2}$$

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

557 To explore the relationship between HPO phenotype specificity and various metrics from our results, we  
558 computed the information content (IC) scores for each term in the HPO. IC is a measure of how much  
559 specific information a term within an ontology contains. In general, terms deeper in an ontology (closer to the  
560 leaves) are more specific, and thus informative, than terms at the very root of the ontology (e.g. ‘Phenotypic  
561 abnormality’). Where  $k$  denotes the number of offspring terms (including the term itself) and  $N$  denotes the  
562 total number of terms in the ontology, IC can be calculated as:

$$IC = -\log\left(\frac{k}{N}\right)$$

563 Next, IC scores were quantised into 10 bins using the `ceiling` R function to improve visualisation. We  
 564 then performed a series of linear regressions between phenotype binned IC scores and: 1) number of genes  
 565 annotated per HPO phenotype, 2) the number of significantly associated cell types per HPO phenotype, and  
 566 3) the model estimate of each significant phenotype-cell type associations (at FDR < 0.05) after taking the  
 567 log of the absolute value ( $\log_2(|estimate|)$ ).

## 568 Monarch Knowledge Graph recall

569 Finally, we gathered known phenotype-cell type relationships from the Monarch Knowledge Graph (MKG), a  
 570 comprehensive database of links between many aspects of disease biology<sup>25</sup>. This currently includes 103 links  
 571 between HPO phenotypes (n=103) and CL cell types (n=79). Of these, we only considered the 82 phenotypes  
 572 that we were able to test given that our ability to generate associations was dependent on the existence of  
 573 gene annotations within the HPO. We considered instances where we found a significant relationship between  
 574 exactly matching pairs of HPO-CL terms as a hit.

575 However, as the cell types in MKG were not necessarily annotated at the same level as our single-cell references,  
 576 we considered instances where the MKG cell type was an ancestor term of our cell type (e.g. ‘myeloid cell’  
 577 vs. ‘monocyte’), or *vice versa*, as hits. We also adjusted ontological distance by computing the ratio between  
 578 the observed ontological distance and the smallest possible ontological distance for that cell type given the  
 579 cell type that were available in our references ( $dist_{adjusted} = \left(\frac{dist_{observed}+1}{dist_{minimum}+1}\right) - 1$ ). This provides a way of  
 580 accurately measuring how dissimilar our identified cell types were for each phenotype-cell type association  
 581 (Fig. S4).

## 582 Prioritising phenotypes based on severity

583 Only a small fraction of the the phenotypes in HPO (<1%) have metadata annotations containing information  
 584 on their time course, consequences, and severity. This is due to the time-consuming nature of manually  
 585 annotating thousands of phenotypes. To generate such annotations at scale, we previously used Generative Pre-  
 586 trained Transformer 4 (GPT-4), a large language model (LLM) as implemented within OpenAI’s Application  
 587 Programming Interface (API)<sup>22</sup>. After extensive prompt engineering and ground-truth benchmarking, we  
 588 were able to acquire annotations on how often each phenotype directly causes intellectual disability, death,  
 589 impaired mobility, physical malformations, blindness, sensory impairments, immunodeficiency, cancer, reduced  
 590 fertility, or is associated with a congenital onset. These criteria were previously defined in surveys of medical  
 591 experts as a means of systematically assessing phenotype severity<sup>81</sup>. Responses for each metric were provided  
 592 in a consistent one-word format which could be one of: ‘never’, ‘rarely’, ‘often’, ‘always’. This procedure was  
 593 repeated in batches (to avoid exceeding token limits) until annotations were gathered for 16,982/18,082 HPO

594 phenotypes.

595 We then encoded these responses into a semi-quantitative scoring system ('never'=0, 'rarely'=1, 'often'=2,  
596 'always'=3), which were then weighted by multiplying a semi-subjective scoring of the relevance of each metric  
597 to the concept of severity on a scale from 1-6, with 6 being the most severe ('death'=6, 'intellectual\_dis-  
598 ability'=5, 'impaired\_mobility'=4, 'physical\_malformations'=3, 'blindness'=4, 'sensory\_impairments'=3,  
599 'immunodeficiency'=3, 'cancer'=3, 'reduced\_fertility'=1, 'congenital\_onset'=1). Finally, the product of  
600 the score was normalised to a quantitative severity score ranging from 0-100, where 100 is the theoretical  
601 maximum severity score. This phenotype severity scoring procedure can be expressed as follows.

602 Let us denote:

- 603 •  $p$  : a phenotype in the HPO.
- 604 •  $j$  : the identity of a given annotation metric (i.e. clinical characteristic, such as 'intellectual disability'  
605 or 'congenital onset').
- 606 •  $W_j$ : the assigned weight of metric  $j$ .
- 607 •  $F_j$ : the maximum possible value for metric  $j$ , equal to 3 ("always"). This value is equivalent across all  $j$   
608 annotations.
- 609 •  $F_{pj}$  : the numerically encoded value of annotation metric  $j$  for phenotype  $p$ .
- 610 •  $NSS_p$ : the final composite severity score for phenotype  $p$  after applying normalisation to align values  
611 to a 0-100 scale and ensure equivalent meaning regardless of which other phenotypes are being analysed  
612 in addition to  $p$ . This allows for direct comparability of severity scores across studies with different sets  
613 of phenotypes.

$$614 \quad NSS_p = \frac{\sum_{j=1}^m (F_{pj} \times W_j)}{\sum_{j=1}^m (\max\{F_j\} \times W_j)} \times 100$$

615      Sum of weighted annotation values  
616      across all metrics

617      Numerically encoded annotation value  
618      of metric  $j$  for phenotype  $p$

619      Weight for metric  $j$

620      Theoretical maximum severity score

621      Normalised Severity Score  
622      for each phenotype

619 Using the numerically encoded GPT annotations (0="never", 1="rarely", 2="often", 3="always") we  
620 computed the mean encoded value per cell type within each annotation. One-sided Wilcoxon rank-sum  
621 tests were run using the `rstatix::wilcox_test()` function to test whether each cell type was associated  
622 with more severe phenotypes relative to all other cell types. This procedure was repeated for severity

annotation independently (death, intellectual disability, impaired mobility, etc.) Fig. 6a. Next, we performed a Pearson correlation test between the number of phenotypes that a cell type is significantly associated with (at FDR<0.05) has a relationship with the mean composite GPT severity score of those phenotypes (Fig. 6b). This was performed using the `ggstatsplot::ggscatterstats()` R function.

## 627 Congenital phenotypes are associated with foetal cell types

628 The GPT-4 annotations also enabled us to assess whether foetal cell types were more often significantly  
629 associated with congenital phenotypes in our Human Cell Landscape results as this single-cell reference  
630 contained both adult and foetal versions of cell types (Fig. 7). To do this, we performed a chi-squared  
631 ( $\chi^2$ ) test on the proportion of significantly associated cell types containing any of the substrings or (within  
632 cell types annotations from the original Human Cell Landscape authors<sup>19</sup>) vs. those associated without,  
633 stratified by how often the corresponding phenotype had a congenital onset according to the GPT phenotype  
634 annotations (including ‘never’, ‘rarely’, ‘often’, ‘always’). In addition, a series of  $\chi^2$  tests were performed  
635 within each congenital onset frequency strata, to determine whether the observed proportion of foetal cell  
636 types vs. non-foetal cell types significantly deviated from the proportions expected by chance.

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

641 Next, we aimed to create a continuous metric from -1 to 1 that indicated how biased each phenotype is  
642 towards associations with the foetal or adult form of a cell type. For each phenotype we calculated the  
643 foetal-adult bias score as the difference in the association p-values between the foetal and adult version of the  
644 equivalent cell type (foetal-adult bias :  $p_{adult} - p_{foetal} = \Delta p \in [-1, 1]$ ). A score of 1 indicates the phenotype  
645 is only associated with the foetal version of the cell type and -1 indicates the phenotype is only associated  
646 with the adult version of the cell type.

647 In order to summarise higher-order HPO phenotype categories that were most biased towards foetal or adult  
648 cell types, ontological enrichment tests were run on the phenotypes with the top/bottom 50 greatest/smallest  
649 foetal-adult bias scores. The enrichment tests were performed using the `simona::dag_enrich_on_offsprings`  
650 function, which uses a hypergeometric test to determine whether a list of terms in an ontology are enriched  
651 for offspring terms (descendants) of a given ancestor term within the ontology. Phenotypes categories with  
652 an HPO ontological enrichment a p-value < 0.05 were considered significant.

653 We were similarly interested in which higher-order cell type categories tended to be most commonly associated  
654 with these strongly foetal-/adult-biased phenotype s. Another set of ontological enrichment tests were  
655 run on the cell types associated with the top/bottom 50 phenotypes from the previous analysis. The CL  
656 ontology-aligned IDs for each group cell types were fed into the `simona::dag_enrich_on_offsprings` using

657 the CL ontology. Significantly enriched cell type categories were defined as those with a CL ontological  
658 enrichment p-value < 0.05.

659 **Therapeutic target identification**

660 We developed a systematic and automated strategy for identifying putative cell type-specific gene targets for  
661 each phenotype based on a series of filters at phenotype, cell type, and gene levels.

662 First, we transformed our phenotype-cell type association results and merged them with primary data sources  
663 (e.g. GenCC gene-disease relationships, scRNA-seq atlas datasets) to create a large table of multi-scale  
664 relationships, where each row represented a tetrad of disease-phenotype-cell type-gene relationships. We  
665 then filtered non-significant phenotype-cell type relationships (only associations with  $FDR < 0.05$ ) as well  
666 as phenotype-gene relationships with strong causal evidence (GenCC score > 3). We also removed any  
667 phenotypes that were too broad to be clinically useful, as quantified using the information content (IC)  
668 ( $IC > 8$ ), which measures the how specific each term is within an ontology (i.e. HPO). Gene-cell type  
669 relationships were established by taking genes that had the top 25% expression specificity quantiles within  
670 each cell type. When connecting cell types to diseases via phenotypes, we used a symptom intersection  
671 threshold of >.25. Next, we sorted the remaining results in descending order of phenotype severity using  
672 the GPT4 composite severity scores described earlier. Finally, to limit the size of the resulting multi-scale  
673 networks we took only the top 10 rows, where each row represented a tetrad of disease-phenotype-cell type-gene  
674 relationships. This resulted in number of relatively small, high-confidence disease-phenotype-cell type-gene  
675 networks that could be reasonably interrogated through manual inspection and network visualisation. For  
676 example, if one was interested in the mechanisms causing ‘Recurrent Neisserial infections’, one would need  
677 only select all rows that include this phenotype to find all of its most relevant connection to diseases, cell  
678 types, and genes.

679 The entire target prioritisation procedure can be replicated with a single function: `MSTExplorer::prioritise_targets`.  
680 This function automates all of the reference data gathering (e.g. phenotype metadata, cell type metadata,  
681 cell type signature reference, gene lengths, severity tiers) and takes a variety of arguments at each step for  
682 greater customisability. Each step is described in detail in Table S3. Phenotypes that often or always caused  
683 physical malformations (according to the GPT-4 annotations) were also removed from the final prioritised  
684 targets list, as these were unlikely to be amenable to gene therapy interventions. Finally, phenotypes were  
685 sorted by their composite severity scores such that the most severe phenotypes were ranked the highest.

686 **Therapeutic target validation**

687 To assess whether our prioritised therapeutic targets were likely to be viable, we computed the overlap  
688 between our gene targets and those of existing gene therapies at various stages of clinical development  
689 (Fig. 8). Gene targets were obtained for each therapy from the Therapeutic Target Database (TTD; release

690 2026-02-25) and mapped onto standardised HUGO Gene Nomenclature Committee (HGNC) gene symbols  
691 using the `orthogene` R package. We stratified our overlap metrics according to whether the therapies had  
692 failed (unsuccessful clinical trials or withdrawn), or were non-failed (successful or ongoing clinical trials).  
693 We then conducted hypergeometric tests to determine whether the observed overlap between our prioritised  
694 targets and the non-failed therapy targets was significantly greater than expected by chance (i.e. enrichment).  
695 We also conducted a second hypergeometric test to determine whether the observed overlap between our  
696 prioritised targets and the failed therapy targets was significantly less than expected by chance (i.e. depletion).  
697 Finally, we repeated the analysis against all therapeutic targets, not just those of gene therapies, to determine  
698 whether our prioritised targets had relevance to other therapeutic modalities.

#### 699 Experimental model translatability

700 To improve the likelihood of successful translation between preclinical animal models and human patients, we  
701 created an interspecies translatability prediction tool for each phenotype nominated by our gene therapy  
702 prioritised pipeline (Fig. S8). First, we extracted ontological similarity scores of homologous phenotypes  
703 across species from the MKG<sup>25</sup>. Briefly, the ontological similarity scores ( $SIM_o$ ) are computed for each  
704 homologous pair of phenotypes across two ontologies by calculating the overlap in homologous phenotypes  
705 that are ancestors or descendants of the target phenotype. Next, we generated genotypic similarity scores  
706 ( $SIM_g$ ) for each homologous phenotype pair by computing the proportion of 1:1 orthologous genes using  
707 gene annotation from their respective ontologies. Interspecies orthologs were also obtained from the MKG.  
708 Finally, both scores are multiplied together to yield a unified ontological-genotypic similarity score ( $SIM_{og}$ ).

#### 709 Novel R packages

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

#### 718 Rare Disease Celltyping Portal

719 To further increase the ease of access for stakeholders in the RD community without the need for programmatic  
720 experience, we developed a series of web apps to interactively explore, visualise, and download the results  
721 from our study. Collectively, these web apps are called the Rare Disease Celltyping Portal. The website can  
722 be accessed at <https://neurogenomics-ukdri.dsi.ic.ac.uk/>.

723 The Rare Disease Celltyping Portal integrates diverse datasets, including the HPO, cell types, genes, and  
724 phenotype severity, into a unified platform that allows users to perform flexible, bidirectional queries. Users can  
725 start from any entry point: either phenotype, cell type, genes, or severity, and seamlessly trace relationships  
726 across these dimensions.

727 The portal provides a dynamic and intuitive exploration experience with its real-time interaction capabilities  
728 and responsive interface including network graphs, bar charts, and heat maps. It has the ability to handle large  
729 datasets efficiently and offer fast query response by building with FARM stack (FastAPI, React, MongoDB).  
730 The portal is designed for a broad audience, including researchers, clinicians, and biologists, by offering  
731 user-friendly navigation and interactive visual outputs. By enabling users to intuitively explore complex  
732 biological relationships, the portal aims to accelerate rare disease research, enhance diagnostic accuracy, and  
733 drive therapeutic innovation.

734 All code used to generate the website can be found at <https://github.com/neurogenomics/Rare-Disease-Web-Portal>.  
735

## 736 **Mappings**

737 Mappings from the HPO to other medical ontologies were extracted from the EMBL-EBI Ontology Xref  
738 Service (Oxo; <https://www.ebi.ac.uk/spot/oxo/>) by selecting the National Cancer Institute metathesaurus  
739 (NCIm) as the target ontology and either “SNOMED CT”, “UMLS”, “ICD-9” or “ICD-10CM” as the data  
740 source. HPO terms were then selected as the ID framework with to mediate the cross-ontology mappings.  
741 Mappings between each pair of ontologies were then downloaded, stored in a tabular format. The mappings  
742 files can be accessed with the function `HPOExplorer::get_mappings` or directly via the `HPOExplorer` Releases  
743 page on GitHub (<https://github.com/neurogenomics/HPOExplorer/releases/tag/latest>).

## 744 **Data Availability**

745 All data is publicly available through the following resources:

- 746 • Human Phenotype Ontology (<https://hpo.jax.org>)
- 747 • GenCC (<https://thegencc.org/>)
- 748 • Descartes Human scRNA-seq atlas (<https://cellxgene.cziscience.com/collections/c114c20f-1ef4-49a5-9c2e-d965787fb90c>)
- 749 • Human Cell Landscape scRNA-seq atlas (<https://cellxgene.cziscience.com/collections/38833785-fac5-48fd-944a-0f62a4c23ed1>)
- 750 • Processed Cell Type Datasets (`ctd_DescartesHuman.rds` and `ctd_HumanCellLandscape.rds`; <https://github.com/neurogenomics/MSTEExplorer/releases>)
- 751 • Gene x Phenotype association matrix (`hpo_matrix.rds`; <https://github.com/neurogenomics/MSTEExplorer/releases>)

- 756 • GPT-4 phenotype severity annotations ([https://github.com/neurogenomics/rare\\_disease\\_celltyping/releases/download/latest/gpt\\_check\\_annot.csv.gz](https://github.com/neurogenomics/rare_disease_celltyping/releases/download/latest/gpt_check_annot.csv.gz))
- 757
- 758 • Full phenotype-cell type association test results [https://github.com/neurogenomics/MSTEExplorer/releases/download/v0.1.10/phenomix\\_results.tsv.gz](https://github.com/neurogenomics/MSTEExplorer/releases/download/v0.1.10/phenomix_results.tsv.gz)
- 759
- 760 • Rare Disease Celltyping Portal (<https://neurogenomics-ukdri.dsi.ic.ac.uk/>)
- 761
- 762 • Rare Disease Celltyping Portal data (<https://zenodo.org/records/15147825>)
- 763
- 764 • Complement system gene list (<https://www.genenames.org/data/genegroup/#!/group/492>)
- 765
- 766 • Therapeutic Target Database data (<http://db.idrblab.net/ttd/>)
- 767
- 768 • CellxGene browser: hepatocytes & hepatoblast complement expression (<https://cellxgene.cziscience.com/gene-expression?genes=C5%2CCFI%2CC8B%2CCFB%2CMBL2%2CAPOE&cellTypes=hepatoblast%2Chepatocyte&ver=2>)

## 767 **Code Availability**

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

- 769 • KGExplorer (<https://github.com/neurogenomics/KGExplorer>)
- 770 • HPOExplorer (<https://github.com/neurogenomics/HPOExplorer>)
- 771 • MSTEExplorer (<https://github.com/neurogenomics/MSTEExplorer>)
- 772 • Code to replicate analyses ([https://github.com/neurogenomics/rare\\_disease\\_celltyping](https://github.com/neurogenomics/rare_disease_celltyping))
- 773 • Rare Disease Celltyping Portal code (<https://github.com/neurogenomics/Rare-Disease-Web-Portal>)

## 774 **Acknowledgements**

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

## 778 **Funding**

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

## 782 **References**

- 783 1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A* **179**, 885–892 (2019).
- 784 2. Zhu, Q. *et al.* An integrative knowledge graph for rare diseases, derived from the genetic and rare  
785 diseases information center (GARD). *J. Biomed. Semantics* **11**, 13 (2020).
3. Rare diseases BioResource.

- 786 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).
- 787 5. Molster, C. *et al.* Survey of healthcare experiences of australian adults living with rare diseases.  
*Orphanet J. Rare Dis.* **11**, 30 (2016).
- 788 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).
- 789 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).
- 790 8. Yates, N. & Hinkel, J. The economics of moonshots: Value in rare disease drug development. *Clin.  
Transl. Sci.* **15**, 809–812 (2022).
- 791 9. Nuijten, M. Pricing zolgensma - the world's most expensive drug. *J Mark Access Health Policy* **10**,  
2022353 (2022).
- 792 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).
- 793 11. Gargano, M. A. *et al.* The human phenotype ontology in 2024: Phenotypes around the world. *Nucleic  
Acids Res.* **52**, D1333–D1346 (2024).
- 794 12. Köhler, S. *et al.* Expansion of the human phenotype ontology (HPO) knowledge base and resources.  
*Nucleic Acids Res.* **47**, D1018–D1027 (2019).
- 795 13. 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).
- 796 14. 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).
- 797 15. 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).
- 798 16. 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).
- 799 17. Qi, R. & Zou, Q. Trends and potential of machine learning and deep learning in drug study at  
Single-Cell level. *Research* **6**, 0050 (2023).
- 800 18. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
- 801 19. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).

- 802 20. 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).
- 803 21. 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).
- 804 22. Murphy, K., Schilder, B. M. & Skene, N. G. Harnessing generative AI to annotate the severity of all phenotypic abnormalities within the Human Phenotype Ontology. <https://doi.org/10.1101/2024.06.10.24308475> doi:[10.1101/2024.06.10.24308475](https://doi.org/10.1101/2024.06.10.24308475).
- 805 23. 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).
- 806 24. Diehl, A. D. *et al.* The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
- 807 25. 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).
- 808 26. Heim, C. E. *et al.* Myeloid-derived suppressor cells contribute to staphylococcus aureus orthopedic biofilm infection. *J. Immunol.* **192**, 3778–3792 (2014).
- 809 27. 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).
- 810 28. Stoll, H. *et al.* Staphylococcal enterotoxins Dose-Dependently modulate the generation of Myeloid-Derived suppressor cells. *Front. Cell. Infect. Microbiol.* **8**, 321 (2018).
- 811 29. 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).
- 812 30. Ladhami, S. N. *et al.* Invasive meningococcal disease in patients with complement deficiencies: A case series (2008-2017). *BMC Infect. Dis.* **19**, 522 (2019).
- 813 31. Rosain, J. *et al.* Strains responsible for invasive meningococcal disease in patients with terminal complement pathway deficiencies. *J. Infect. Dis.* **215**, 1331–1338 (2017).
- 814 32. Seal, R. L. *et al.* Genenames.org: The HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).
- 815 33. Zhou, Z., Xu, M.-J. & Gao, B. Hepatocytes: A key cell type for innate immunity. *Cell. Mol. Immunol.* **13**, 301–315 (2016).
- 816 34. Yu, Y. *et al.* Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. *Genome Research* **11**, 1392–1403 (2001).

- 817 35. Wesley, B. T. *et al.* Single-cell atlas of human liver development reveals pathways directing hepatic cell fates. *Nature cell biology* **24**, 1487–1498 (2022).
- 818 36. 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).
- 819 37. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
- 820 38. Sugiyama, K., Tagawa, S. & Toda, M. Methods for visual understanding of hierarchical system structures. *IEEE Trans. Syst. Man Cybern.* **11**, 109–125 (1981).
- 821 39. Ramesh, A. *et al.* Zero-shot text-to-image generation. <https://doi.org/10.48550/arXiv.2102.12092> doi:[10.48550/arXiv.2102.12092](https://doi.org/10.48550/arXiv.2102.12092).
- 822 40. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* **47**, 856–860 (2015).
- 823 41. Ochoa, D. *et al.* Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nature Reviews Drug Discovery* **21**, 551–551 (2022).
- 824 42. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature* 1–6 (2024) doi:[10.1038/s41586-024-07316-0](https://doi.org/10.1038/s41586-024-07316-0).
- 825 43. 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).
- 826 44. Chiu, W. *et al.* An update on gene therapy for inherited retinal dystrophy: Experience in leber congenital amaurosis clinical trials. *International Journal of Molecular Sciences* **22**, 4534 (2021).
- 827 45. 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).
- 828 46. 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).
- 829 47. 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).
- 830 48. 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).
- 831 49. Ferro, A., Sheeler, C., Rosa, J.-G. & Cvetanovic, M. Role of microglia in ataxias. *Journal of molecular biology* **431**, 1792–1804 (2019).
- 832 50. 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).
- 833 51. 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).

- 834 52. Katsanis, N. [The continuum of causality in human genetic disorders](#). *Genome Biology* **17**, (2016).
- 835 53. Freund, M. K. *et al.* [Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits](#). *American Journal of Human Genetics* **103**, 535–552 (2018).
- 836 54. Ziegler, K. C. *et al.* The brain neurovascular epigenome and its association with dementia. *Neuron* <https://doi.org/10.1016/j.neuron.2025.10.001> (2025) doi:[10.1016/j.neuron.2025.10.001](https://doi.org/10.1016/j.neuron.2025.10.001).
- 837 55. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of parkinson’s disease. *Nature Genetics* 10–18 (2020) doi:[10.1038/s41588-020-0610-9](https://doi.org/10.1038/s41588-020-0610-9).
- 838 56. Sargurupremraj, M. *et al.* [Cerebral small vessel disease genomics and its implications across the lifespan](#). *Nature Communications* **11**, 6285 (2020).
- 839 57. Nott, A. & Holtman, I. R. [Genetic insights into immune mechanisms of alzheimer’s and parkinson’s disease](#). *Frontiers in Immunology* **14**, 1168539 (2023).
- 840 58. Novikova, G. Integration of alzheimer’s disease genetics and myeloid cell genomics identifies novel causal variants, regulatory elements, genes and pathways. *Bioarxiv* (2019).
- 841 59. [Heritability enrichment implicates microglia in parkinson’s disease pathogenesis - andersen - 2021 - annals of neurology - wiley online library](#).
- 842 60. Smajić, S. *et al.* [Single-cell sequencing of human midbrain reveals glial activation and a parkinson-specific neuronal state](#). *Brain* **145**, 964–978 (2022).
- 843 61. Menon, M. *et al.* [Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration](#). *Nature Communications* **10**, 4902 (2019).
- 844 62. Lyu, Y. *et al.* [Implication of specific retinal cell-type involvement and gene expression changes in AMD progression using integrative analysis of single-cell and bulk RNA-seq profiling](#). *Scientific Reports* **11**, 15612 (2021).
- 845 63. Lindström, S. *et al.* [Genome-wide analyses characterize shared heritability among cancers and identify novel cancer susceptibility regions](#). *Journal of the National Cancer Institute* **115**, 712–732 (2023).
- 846 64. Strunz, T., Kiel, C., Sauerbeck, B. L. & Weber, B. H. F. [Learning from Fifteen Years of Genome-Wide Association Studies in Age-Related Macular Degeneration](#). *Cells* **9**, 2267 (2020).
- 847 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).
- 848 66. Bulaklak, K. & Gersbach, C. A. The once and future gene therapy. *Nat. Commun.* **11**, 5820 (2020).
- 849 67. Kohn, D. B., Chen, Y. Y. & Spencer, M. J. Successes and challenges in clinical gene therapy. *Gene Ther.* **30**, 738–746 (2023).
- 850 68. Zhao, Z., Shang, P., Mohanraju, P. & Geijzen, N. Prime editing: Advances and therapeutic applications. *Trends Biotechnol.* **41**, 1000–1012 (2023).
- 851 69. Lu, C.-F. FDA takes first step toward international regulation of gene therapies to treat rare diseases. (2024).

- 852 70. 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).
- 853 71. Zanello, G. *et al.* Targeting shared molecular etiologies to accelerate drug development for rare diseases. *EMBO Mol. Med.* **15**, e17159 (2023).
- 854 72. Côté, R. *et al.* The ontology lookup service: Bigger and better. *Nucleic Acids Res.* **38**, W155–60 (2010).
- 855 73. Mathur, R., Rotroff, D., Ma, J., Shojaie, A. & Motsinger-Reif, A. [Gene set analysis methods: A systematic comparison](#). *BioData Mining* **11**, 8 (2018).
- 856 74. Irizarry, R. A., Wang, C., Zhou, Y. & Speed, T. P. [Gene set enrichment analysis made simple](#). *Statistical methods in medical research* **18**, 565–575 (2009).
- 857 75. Maleki, F., Ovens, K., Hogan, D. J. & Kusalik, A. J. [Gene set analysis: Challenges, opportunities, and future research](#). *Frontiers in Genetics* **11**, (2020).
- 858 76. DiStefano, M. T. *et al.* The gene curation coalition: A global effort to harmonize gene-disease evidence resources. *Genet. Med.* **24**, 1732–1742 (2022).
- 859 77. 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).
- 860 78. Skene, N. G. & Grant, S. G. N. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* **10**, 16 (2016).
- 861 79. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).
- 862 80. 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](#). *Frontiers in Neuroscience* **10**, 1–11 (2016).
- 863 81. Lazarin, G. A. *et al.* Systematic classification of disease severity for evaluation of expanded carrier screening panels. *PLoS One* **9**, e114391 (2014).
- 864 82. Solovyeva, V. V. *et al.* [New approaches to tay-sachs disease therapy](#). *Frontiers in Physiology* **9**, (2018).
- 865 83. 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).

866

867

868 **Supplementary Materials**

869 **Supplementary Results**

870 **Congenital phenotypes are associated with foetal cell types**

871 To test whether some cell types tend to show strong differences in their phenotype associations between  
872 their foetal and adult forms. To test this, we performed an analogous enrichment procedure as with the  
873 phenotypes, except using Cell Ontology terms and the Cell Ontology graph. This analysis identified the  
874 cell type category connective tissue cell ( $p = 1.8 \times 10^{-3}$ ,  $\log_2(\text{fold-change})=3.2$ ) as the most foetal-biased cell  
875 type. No cell type categories were significantly enriched for the most adult-biased cell types. This is likely  
876 due to the fact that cell types can be disrupted at different stages of life, resulting in different phenotypes.  
877 Thus there the same cell types may be involved in both the most foetal-biased and adult-biased phenotypes.

878 **Selected example targets**

879 From our prioritised targets, we selected the following four sets of phenotypes or diseases as examples:  
880 ‘GM2-ganglioside accumulation’, ‘Spinocerebellar atrophy’, ‘Neuronal loss in central nervous system’. Only  
881 phenotypes with a GPT severity score greater than 15 were considered to avoid overplotting and to focus on  
882 the more clinically relevant phenotypes Fig. 9a-h. These examples were then selected partly on the basis of  
883 severity rankings, and partly for their relatively smaller, simpler networks than lent themselves to compact  
884 visualisations.

885 Tay-Sachs disease (TSD) is a devastating hereditary condition in which children are born appearing healthy,  
886 which gradually degrades leading to death after 3-5 years. The underlying cause is the toxic accumulation  
887 of gangliosides in the nervous system due to a loss of the enzyme produced by *HEXA*. While this could  
888 in theory be corrected with gene editing technologies, there remain some outstanding challenges. One of  
889 which is identifying which cell types should be targeted to ensure the most effective treatments. Here we  
890 identified alternatively activated macrophages as the cell type most strongly associated with ‘GM2-ganglioside  
891 accumulation’ Fig. 9i. The role of aberrant macrophage activity in the regulation of ganglioside levels is  
892 supported by observation that gangliosides accumulate within macrophages in TSD<sup>45</sup>, as well as experimental  
893 evidence in rodent models<sup>46,47,48</sup>. Our results not only corroborate these findings, but propose macrophages  
894 as the primary causal cell type in TSD, making it the most promising cell type to target in therapies.

895 Another challenge in TSD is early detection and diagnosis, before irreversible damage has occurred. Our  
896 pipeline implicated extravillous trophoblasts of the placenta in ‘GM2-ganglioside accumulation’. While not  
897 necessarily a target for gene therapy (as the child is detached from the placenta after birth), checking these  
898 cells *in utero* for an absence of *HEXA* may serve as a viable biomarker as these cells normally express the gene  
899 at high levels. Early detection of TSD may lengthen the window of opportunity for therapeutic intervention<sup>82</sup>,  
900 especially when genetic sequencing is not available or variants of unknown significance are found within

901 *HEXA*<sup>83</sup>.

902 Spinocerebellar atrophy is a debilitating and lethal phenotype that occurs in diseases such as Spinocerebellar  
903 ataxia and Boucher-Nenhauser syndrome. These diseases are characterised by progressive degeneration of  
904 the cerebellum and spinal cord, leading to severe motor and cognitive impairments. Our pipeline identified  
905 M2 macrophages (labeled as the closest CL term ‘Alternatively activated macrophages’ in Fig. 9j) as the  
906 only causal cell type associated with ‘Spinocerebellar atrophy’. This strongly suggests that degeneration of  
907 cerebellar Purkinje cells are in fact downstream consequences of macrophage dysfunction, rather than being  
908 the primary cause themselves. This is consistent with the known role of macrophages, especially microglia, in  
909 neuroinflammation and other neurodegenerative conditions such as Alzheimer’s and Parkinsons’ disease<sup>49–51</sup>.  
910 While experimental and postmortem observational studies have implicated microglia in spinocerebellar atrophy  
911 previously<sup>49</sup>, our results provide a statistically-supported and unbiased genetic link between known risk  
912 genes and this cell type. Therefore, targeting M2 microglia in the treatment of spinocerebellar atrophy may  
913 therefore represent a promising therapeutic strategy. This is aided by the fact that there are mouse models  
914 that perturb the ortholog of human spinocerebellar atrophy risk genes (e.g. *Atxn1*, *Pnpla6*) and reliably  
915 recapitulate the effects of this diseases at the cellular (e.g. loss of Purkinje cells), morphological (e.g. atrophy  
916 of the cerebellum, spinal cord, and muscles), and functional (e.g. ataxia) levels.

917 Next, we investigated the phenotype ‘Neuronal loss in the central nervous system’. Despite the fact that this  
918 is a fairly broad phenotype, we found that it was only significantly associated with 3 cell types (alternatively  
919 activated macrophage, macrophage, epithelial cell), specifically M2 macrophages and sinusoidal endothelial  
920 cells Fig. 9k.

921 Skeletal dysplasia is a heterogeneous group of over 450 disorders that affect the growth and development of  
922 bone and cartilage. This phenotype can be lethal when deficient bone growth leads to the constriction of vital  
923 organs such as the lungs. Even after surgical interventions, these complications continue to arise as the child  
924 develops. Pharmacological interventions to treat this condition have largely been ineffective. While there  
925 are various cell types involved in skeletal system development, our pipeline nominated chondrocytes as the  
926 causal cell type underlying the lethal form of this condition (Fig. S10). Assuringly, we found that the disease  
927 ‘Achondrogenesis Type 1B’ is caused by the genes *SLC26A2* and *COL2A1* via chondrocytes. We also found  
928 that ‘Platyspondylic lethal skeletal dysplasia, Torrance type’. Thus, in cases where surgical intervention is  
929 insufficient, targeting these genes within chondrocytes may prove a viable long-term solution for children  
930 suffering from lethal skeletal dysplasia.

931 Alzheimer’s disease (AD) is the most common neurodegenerative condition. It is characterised by a set  
932 of variably penetrant phenotypes including memory loss, cognitive decline, and cerebral proteinopathy.  
933 Interestingly, we found that different forms of early onset AD (which are defined by the presence of a specific  
934 disease gene) are each associated with different cell types via different phenotypes (Fig. S10). For example,  
935 AD 3 and AD 4 are primarily associated with cells of the digestive system (‘enterocyte’, ‘gastric goblet

936 cell') and are implied to be responsible for the phenotypes 'Senile plaques', 'Alzheimer disease', 'Parietal  
937 hypometabolism in FDG PET'. Meanwhile, AD 2 is primarily associated with immune cells ('alternatively  
938 activated macrophage') and is implied to be responsible for the phenotypes 'Neurofibrillary tangles', 'Long-  
939 tract signs'. This suggests that different forms of AD may be driven by different cell types and phenotypes,  
940 which may help to explain its variability in onset and clinical presentation.

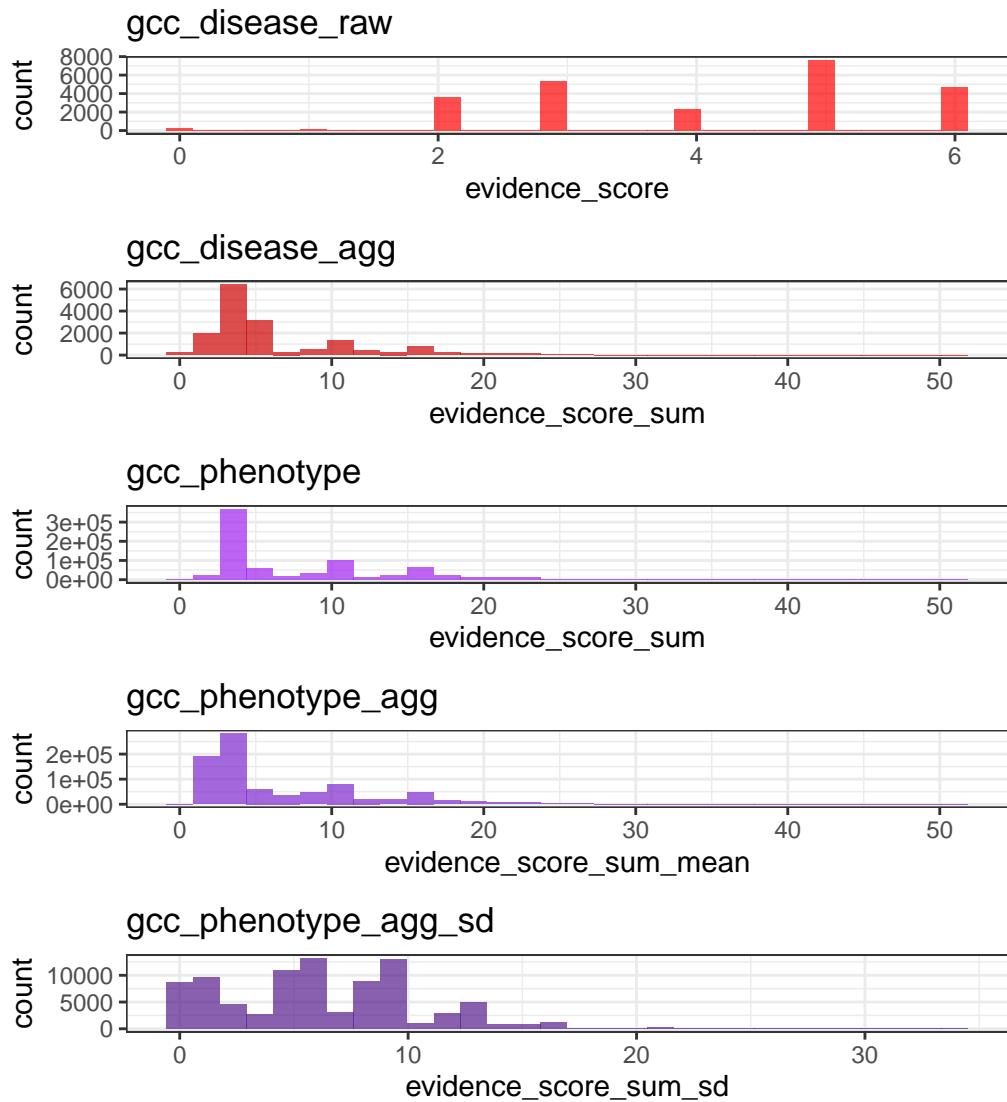
941 Finally, Parkinson's disease (PD) is characterised by motor symptoms such as tremor, rigidity, and bradykinesia.  
942 However there are a number of additional phenotypes associated with the disease that span multiple  
943 physiological systems. PD 19a and PD 8 seemed to align most closely with the canonical understanding of PD  
944 as a disease of the central nervous system in that they implicated oligodendrocytes and neurons (Fig. S10).  
945 Though the reference datasets being used in this study were not annotated at sufficient resolution to distinguish  
946 between different subtypes of neurons, in particular dopaminergic neurons. PD 19a/8 also suggested that risk  
947 variants in *LRRK2* mediate their effects on PD through both myeloid cells and oligodendrocytes by causing  
948 gliosis of the substantia nigra. The remaining clusters of PD mechanisms revolved around chondrocytes (PD  
949 20), amacrine cells of the eye (hereditary late-onset PD), and the respiratory/immune system (PD 14). While  
950 the diversity in cell type-specific mechanisms is somewhat surprising, it may help to explain the wide variety  
951 of cross-system phenotypes frequently observed in PD.

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

## 956 Experimental model translatability

957 We computed interspecies translatability scores using a combination of both ontological ( $SIM_o$ ) and genotypic  
958 ( $SIM_g$ ) similarity relative to each homologous human phenotype and its associated genes Fig. S8. In total, we  
959 mapped 281 non-human phenotypes (in *Caenorhabditis elegans*, *Danio rerio*, *Mus musculus*, *Rattus norvegicus*)  
960 to 375 homologous human phenotypes. Amongst the 5,252 phenotype within our prioritised therapy targets,  
961 154 had viable animal models in at least one non-human species. Per species, the number of homologous  
962 phenotypes was: *Mus musculus* (n=146), *Danio rerio* (n=29), *Rattus norvegicus* (n=5), *Caenorhabditis*  
963 *elegans* (n=1). Amongst our prioritised targets with a GPT-4 severity score of >10, the phenotypes with  
964 the greatest animal model similarity were 'Odontodysplasia' ( $SIM_{og} = 0.97$ ), 'Wide-cupped costochondral  
965 junctions' ( $SIM_{og} = 0.93$ ), 'Short tubular bones of the hand' ( $SIM_{og} = 0.85$ ), 'Megacystis' ( $SIM_{og} = 0.76$ ),  
966 'Cryptophthalmos' ( $SIM_{og} = 0.71$ ).

967 Supplementary Figures



**Distribution of GenCC evidence scores at each processing step.** GenCCC (<https://thegencc.org/>) is a database where semi-quantitative scores for the current strength of evidence attributing disruption of a gene as a causal factor in a given disease. “gcc\_disease\_raw” is the distribution of raw GenCC scores before any aggregation. “gcc\_disease\_agg” is the distribution of GenCC scores after aggregating by disease. “gcc\_phenotype” is the distribution of scores after linking each phenotype to one or more disease. “gcc\_phenotype\_agg” is the distribution of scores after aggregating by phenotype, while “gcc\_phenotype\_agg\_sd” is the standard deviation of those aggregated scores.

Figure S1

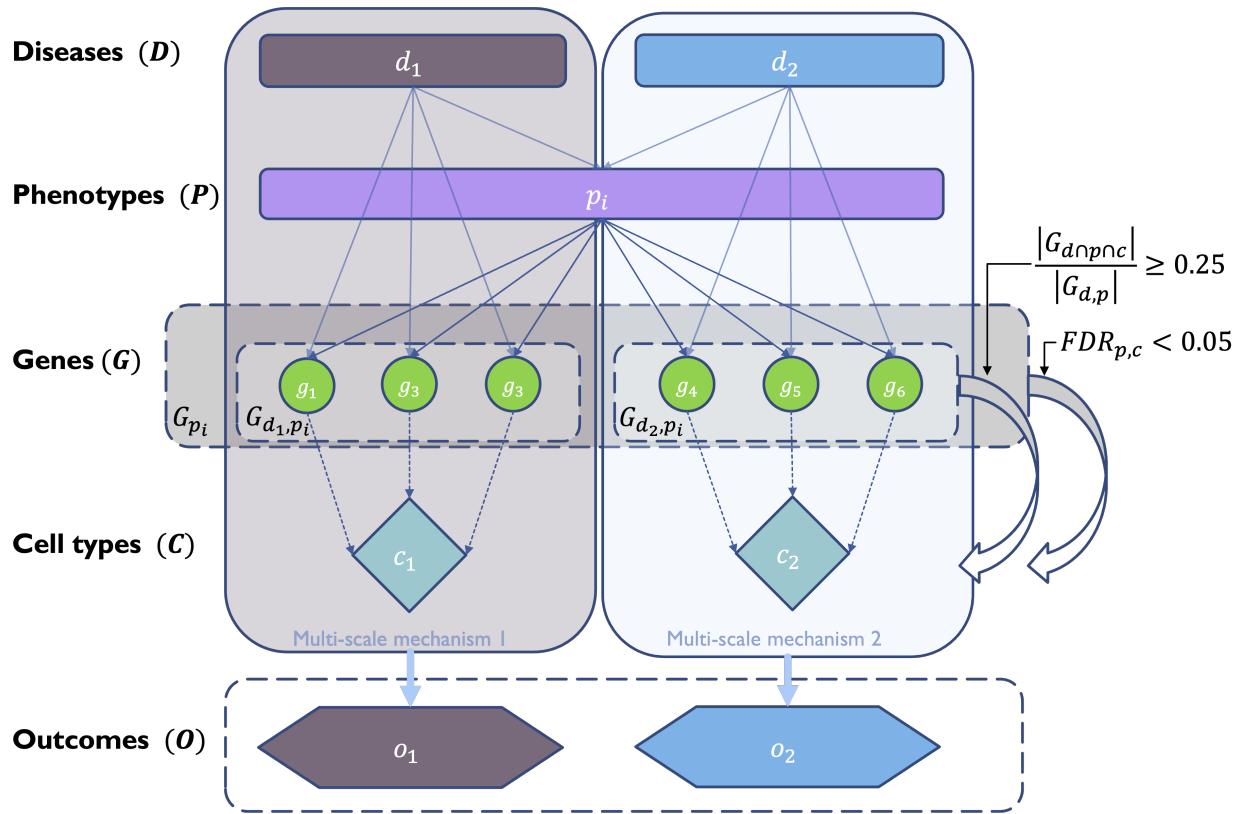
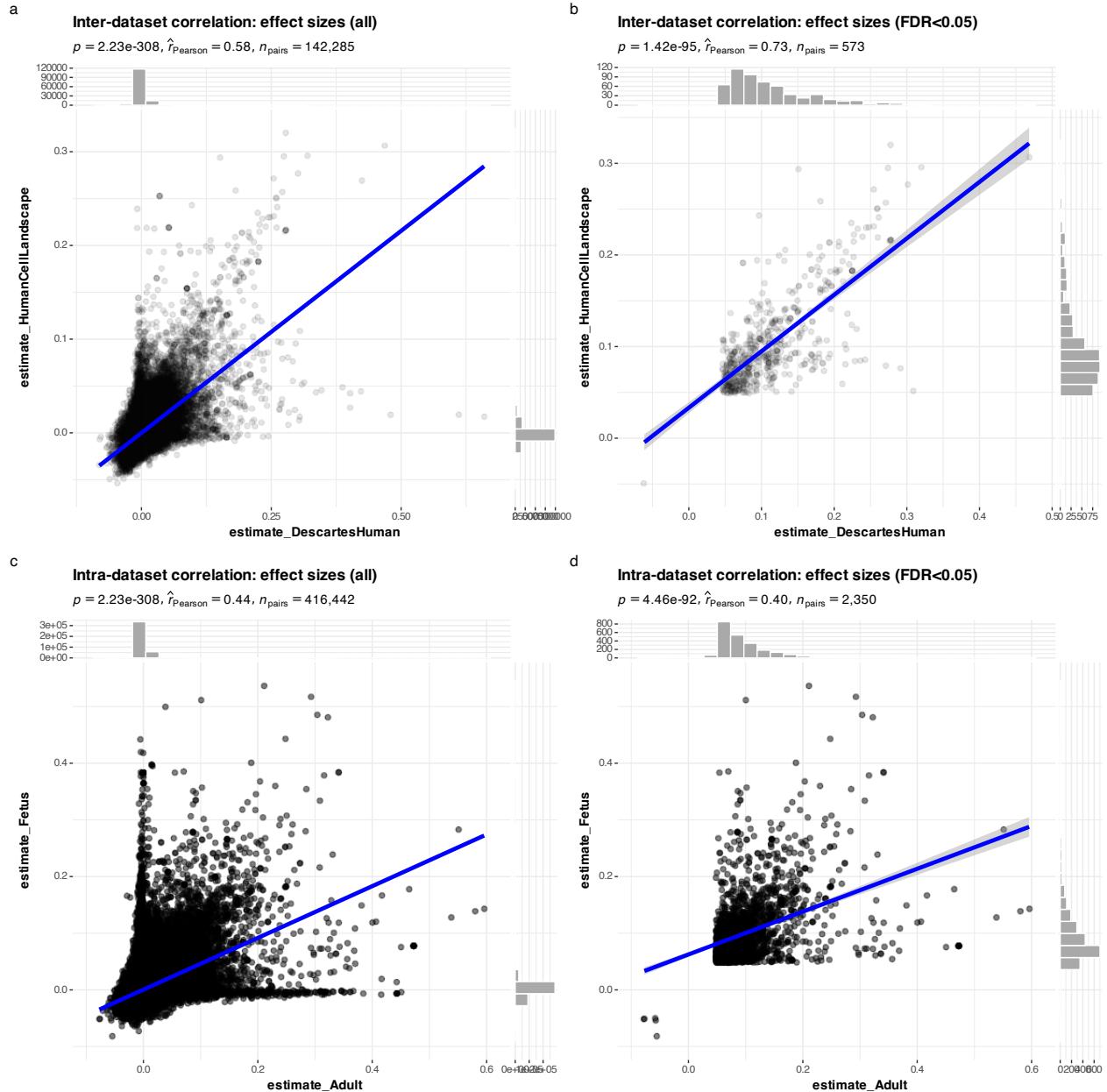
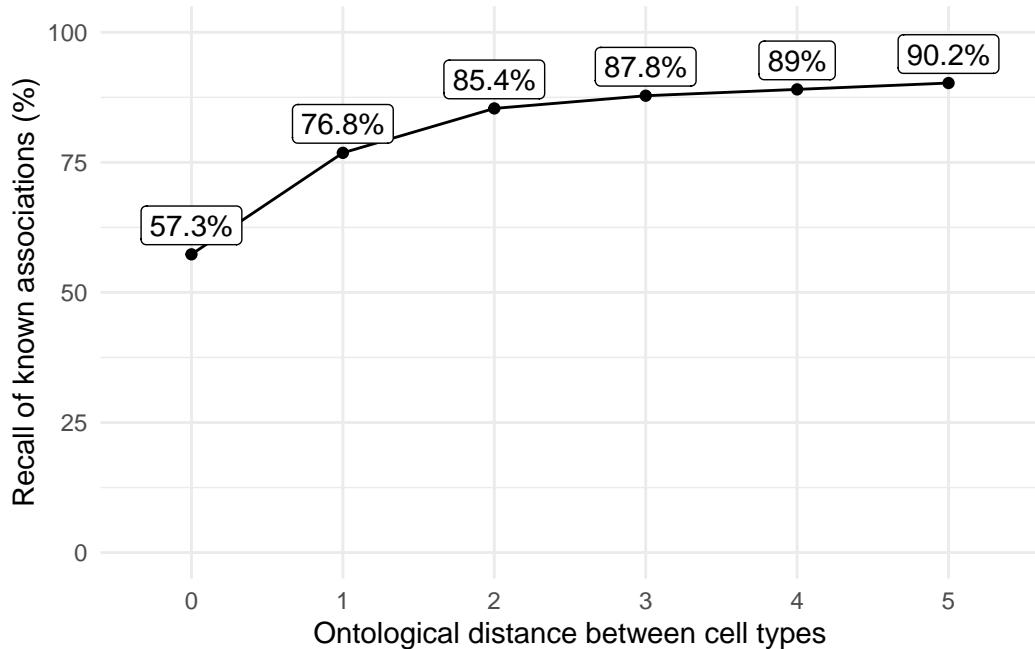


Figure S2: **Diagrammatic overview of multi-scale disease investigation strategy.** Here we provide an abstract example of differential disease etiology across multiple scales: diseases ( $D$ ), phenotypes ( $P$ ), cell types ( $C$ ), genes ( $G$ ), and clinical outcomes ( $O$ ). In the HPO, genes are 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%.



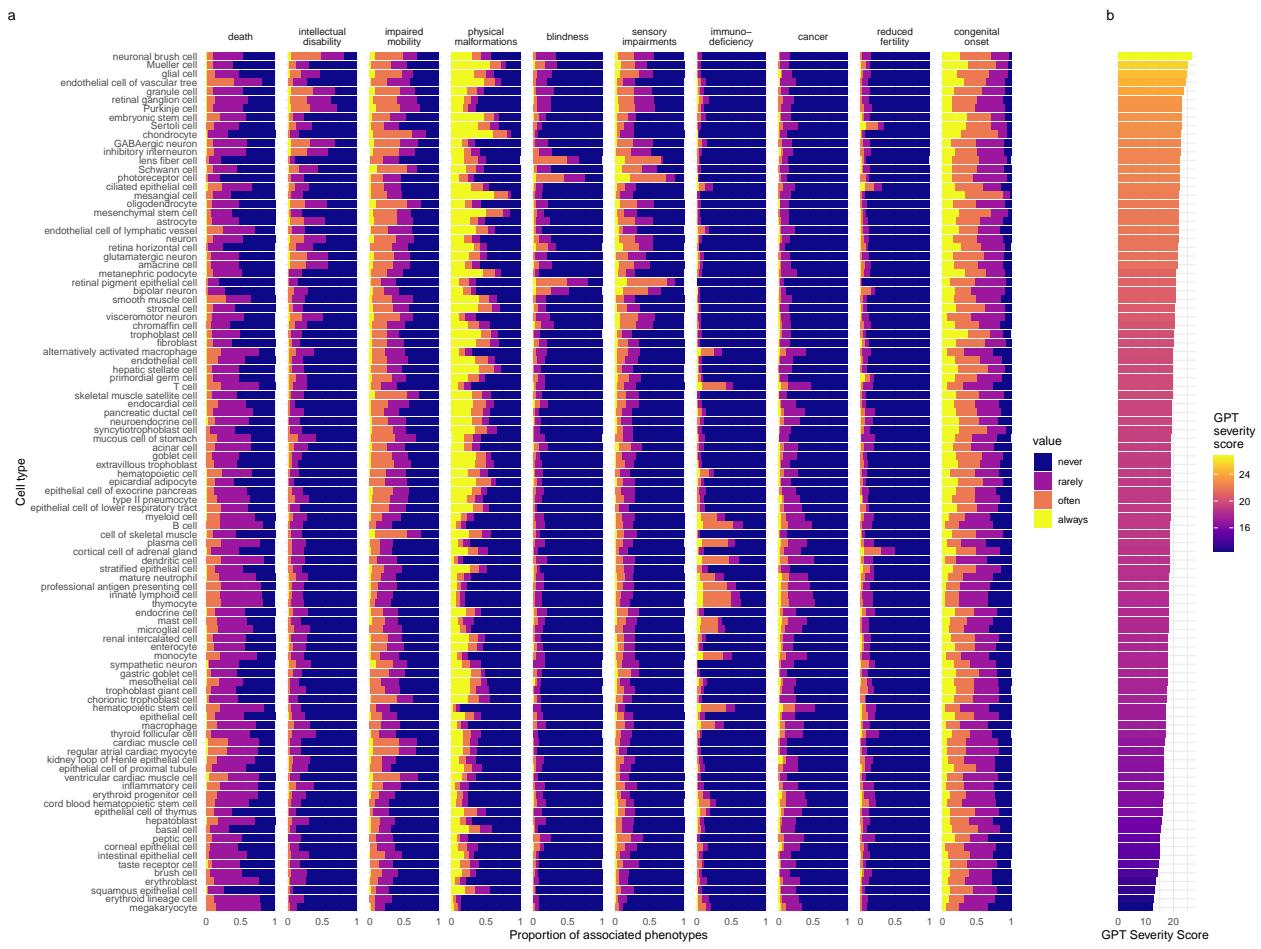
**Inter- and intra-dataset validation across the different CellTypeDataset (CTD) and developmental stages.** Correlations are computed using Pearson correlation coefficient. Point density is plotted using a 2D kernel density estimate. **a** Correlation between the uncorrected p-values from all phenotype-cell type association tests using the Descartes Human vs. Human Cell Landscape CTDs. **b** Correlation between the  $\log_{10}(\text{fold}-\text{change})$  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 foetal samples vs. Human Cell Landscape adult samples. **d** Correlation between the  $\log_{10}(\text{fold}-\text{change})$  from significant phenotype-cell type association tests (FDR<0.05) using the Human Cell Landscape foetal samples vs. Human Cell Landscape adult samples.

Figure S3



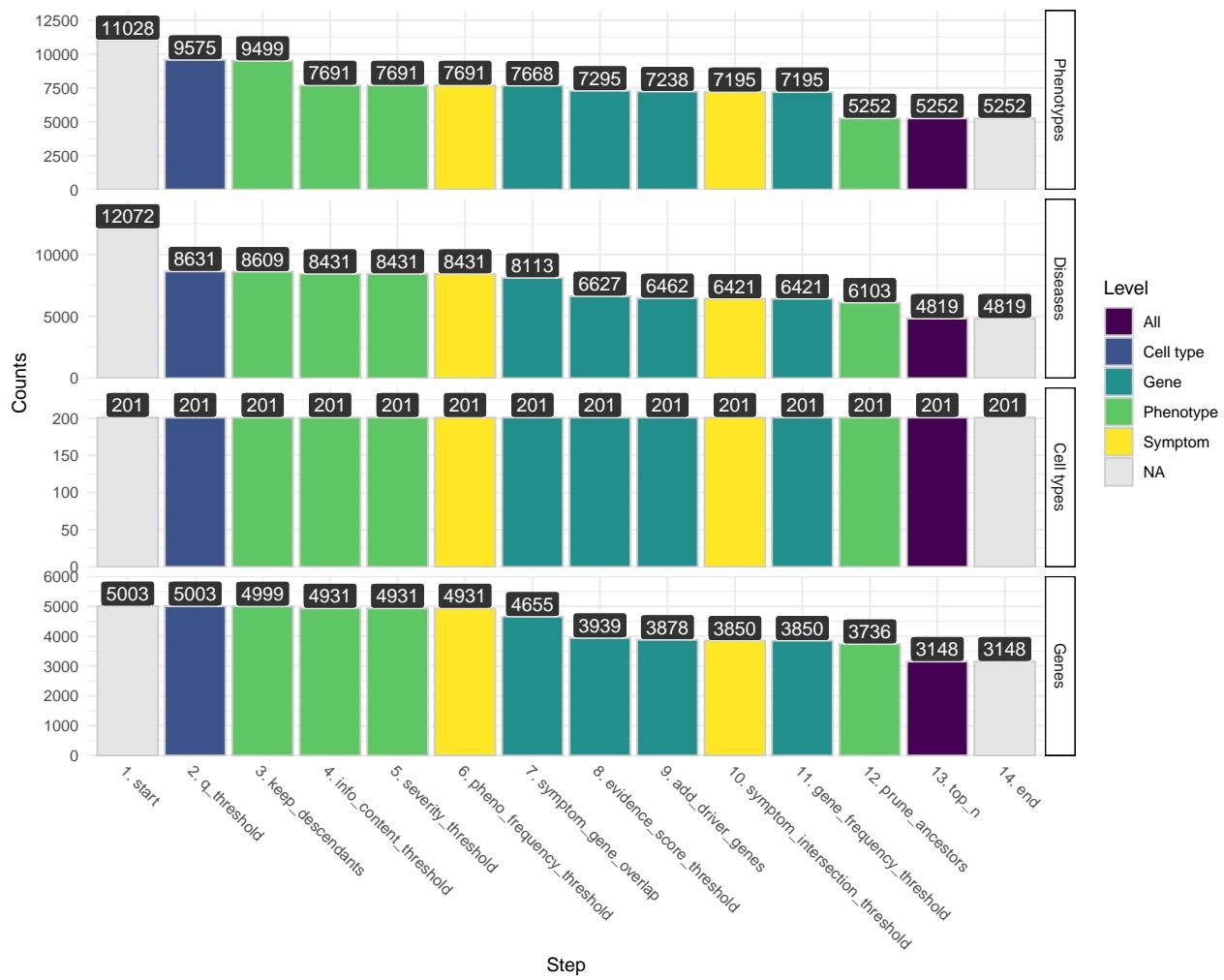
Recall of ground-truth Monarch Knowledge Graph phenotype-cell type relationships at each ontological distance between cell types according to the Cell Ontology.

Figure S4



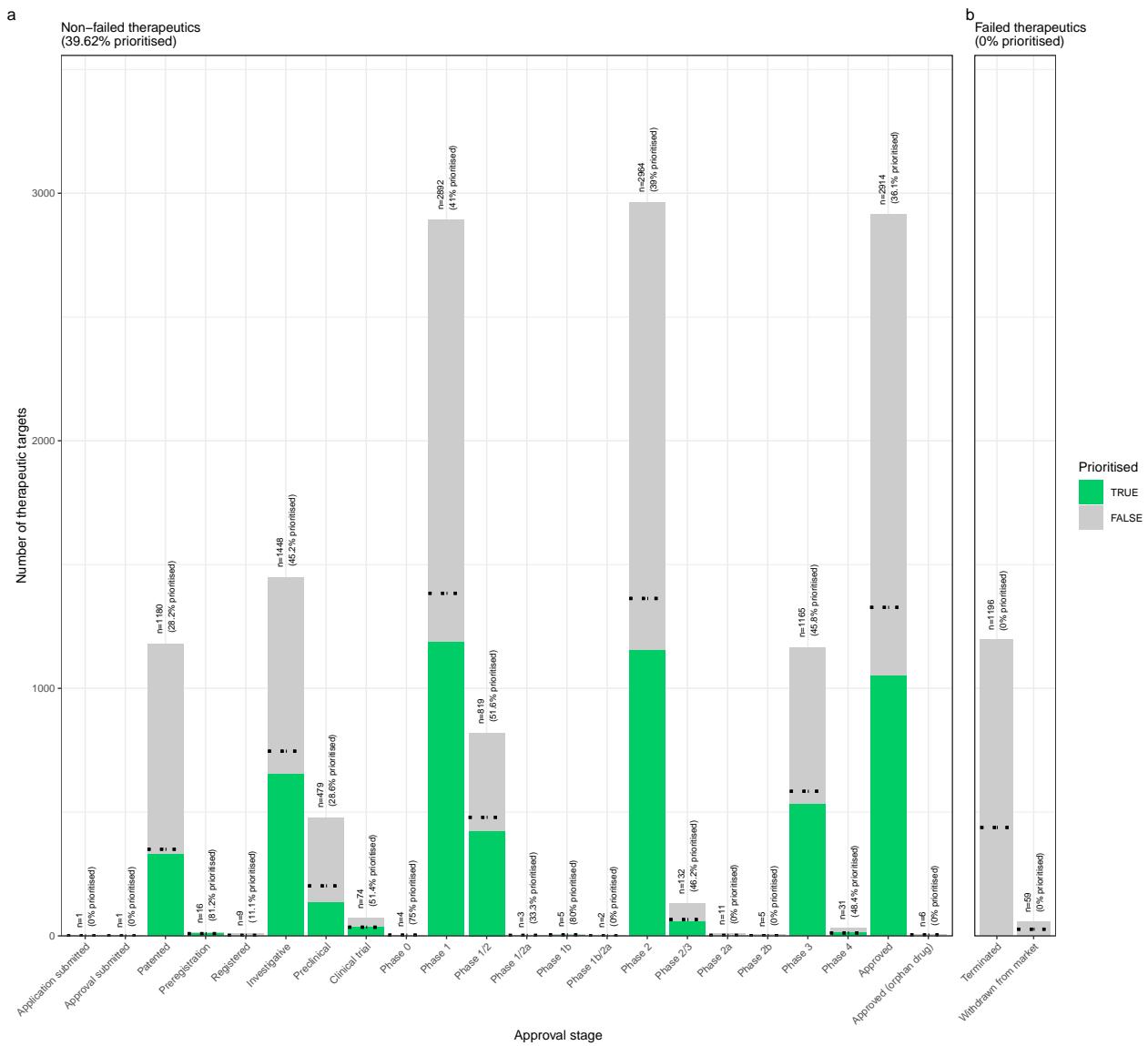
**Cell types ordered by the mean severity of the phenotypes they're associated with.** **a**, The distribution of phenotype severity annotation frequencies aggregated by cell type. **b**, The composite severity score, averaged across all phenotypes associated with each cell type.

Figure S5



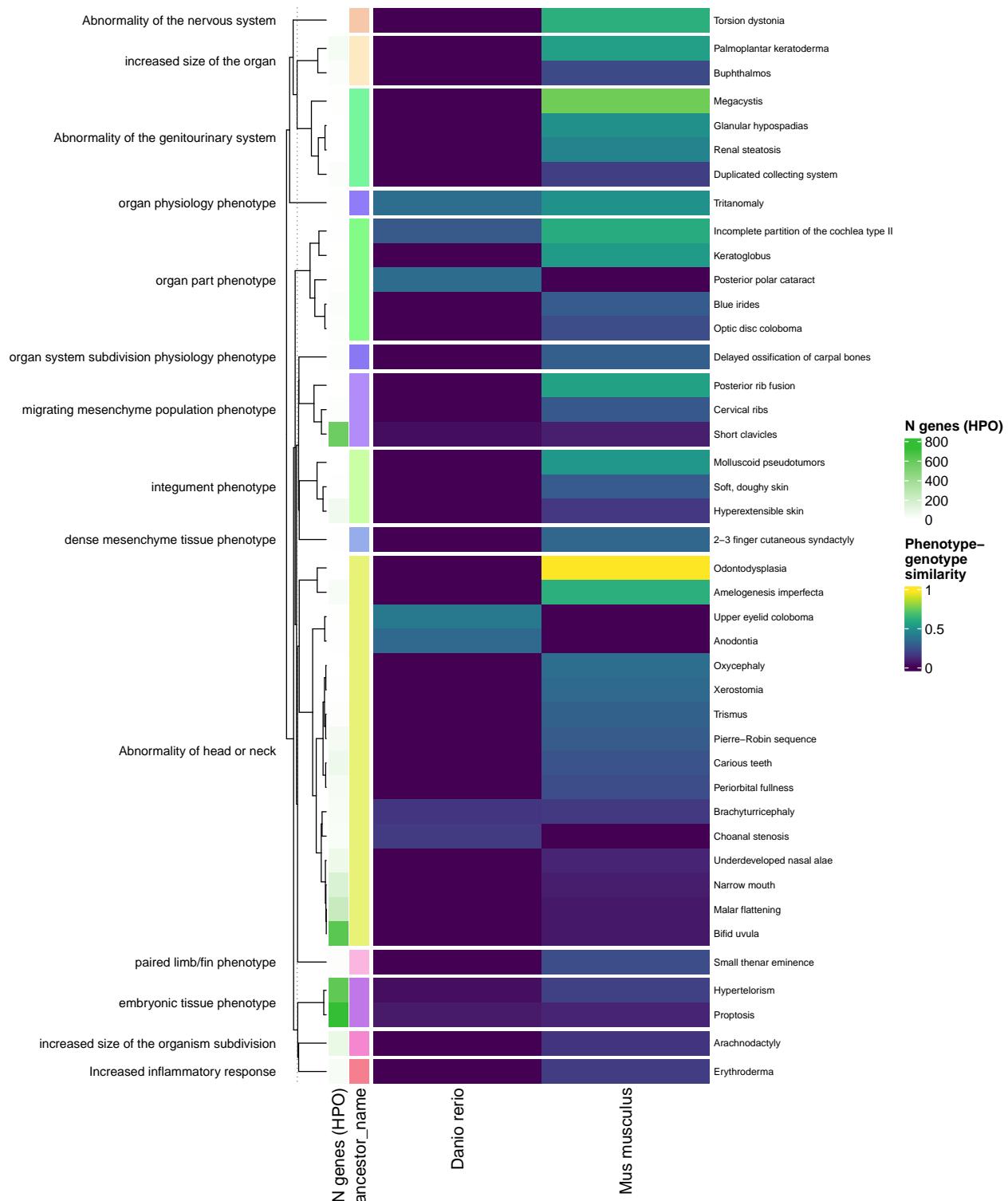
**Prioritised target filtering steps.** This plot visualises the number of unique phenotype-cell type associations, cell types, genes, and phenotypes (*y-axis*) at each filtering step (*x-axis*) within the multi-scale therapeutic target prioritisation pipeline. Each step in the pipeline can be easily adjusted according to user preference and use case. See Table S3 for descriptions and criterion of each filtering step.

Figure S6



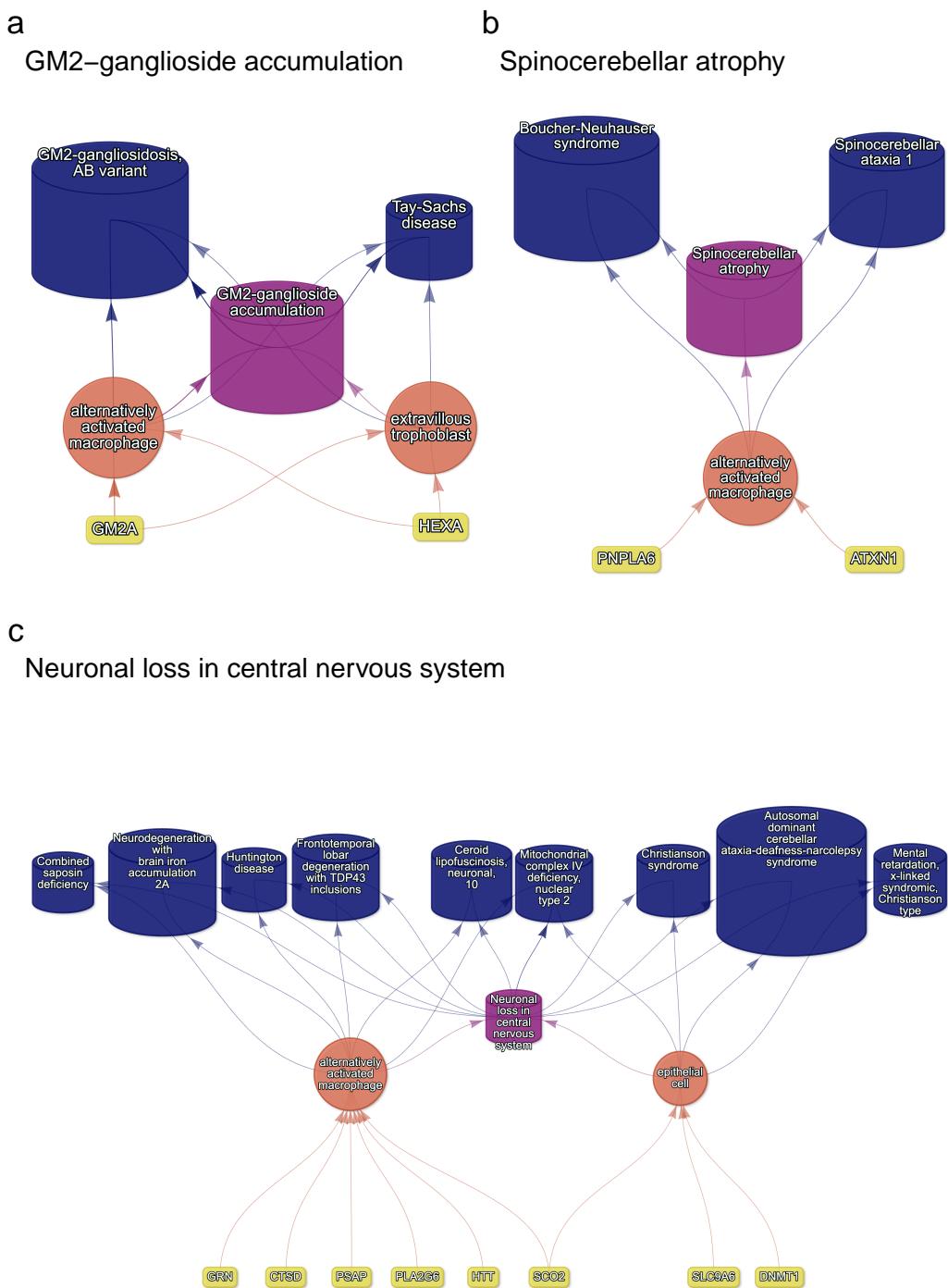
**Validation of prioritised therapeutic targets.** Proportion of existing all therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline.

Figure S7



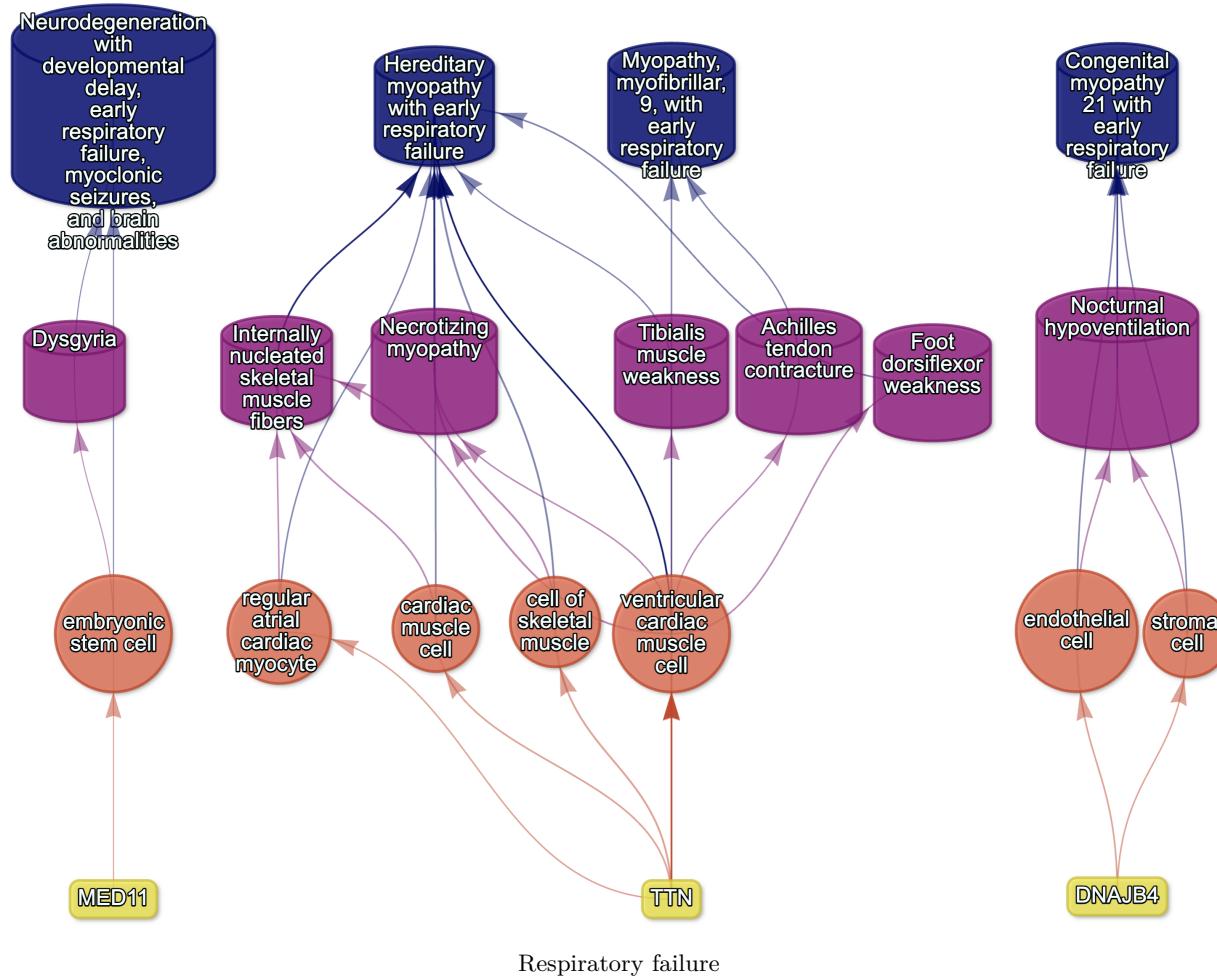
**Identification of translatable experimental models.** Interspecies translatability of the top human phenotypes nominated by the gene therapy prioritised pipeline. Above, the combined ontological-genotypic similarity score ( $SIM_{og}$ ) is displayed as the heatmap fill colour stratified by the model 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*).

Figure S8



**Causal multi-scale networks reveal cell type-specific therapeutic targets.** 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 (orange 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>38</sup>.

Figure S9



**Figure S10: Example cell type-specific gene therapy targets for phenotypes associated with respiratory failure-related 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>38</sup>.

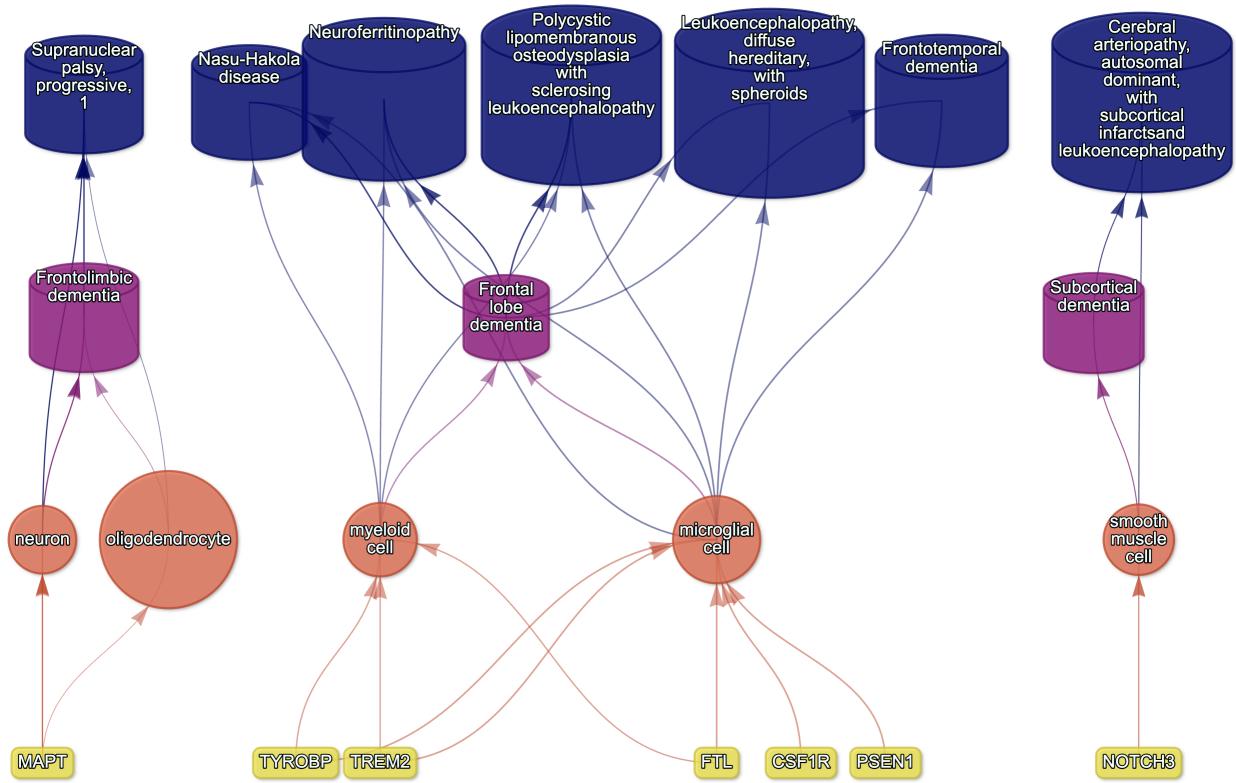


Figure S11: Causal multi-scale network for dementia phenotypes.

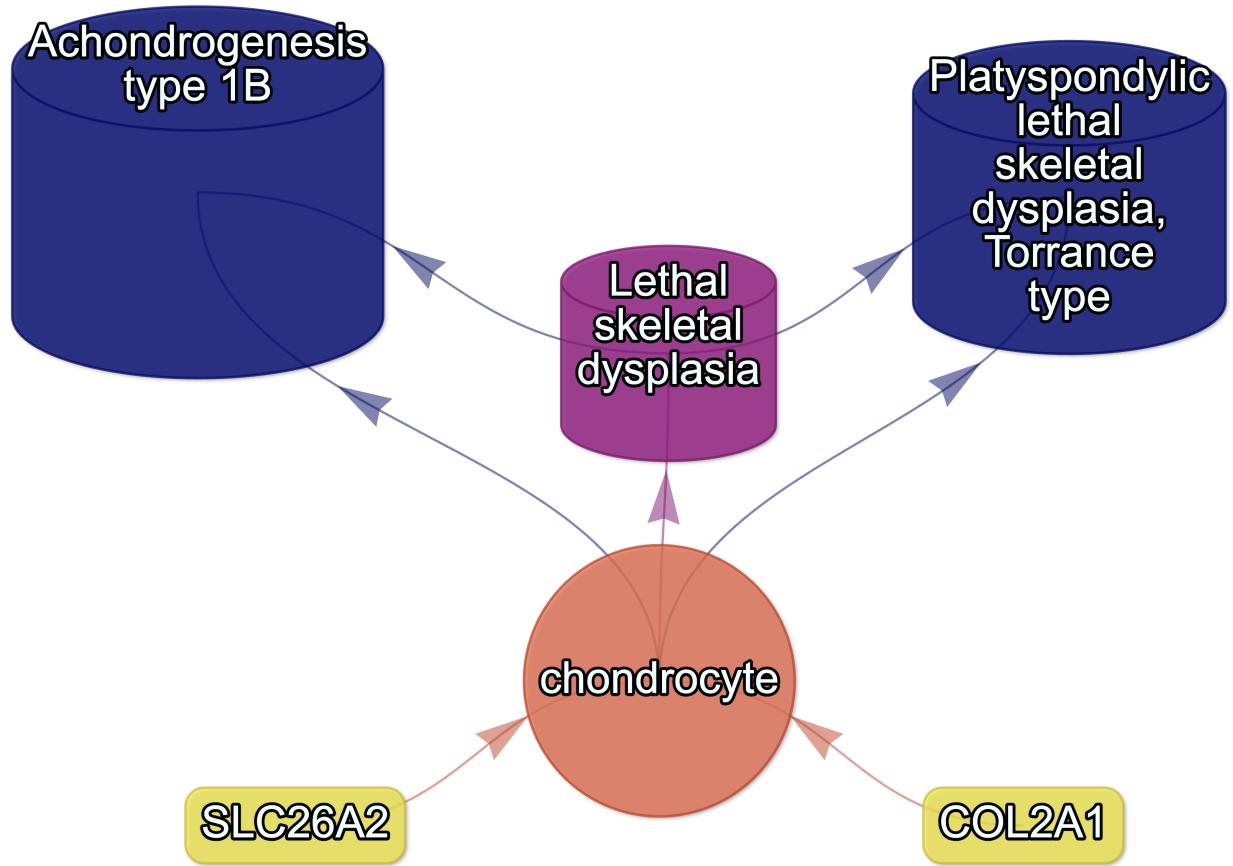


Figure S12: Causal multi-scale network for the phenotype lethal skeletal dysplasia.

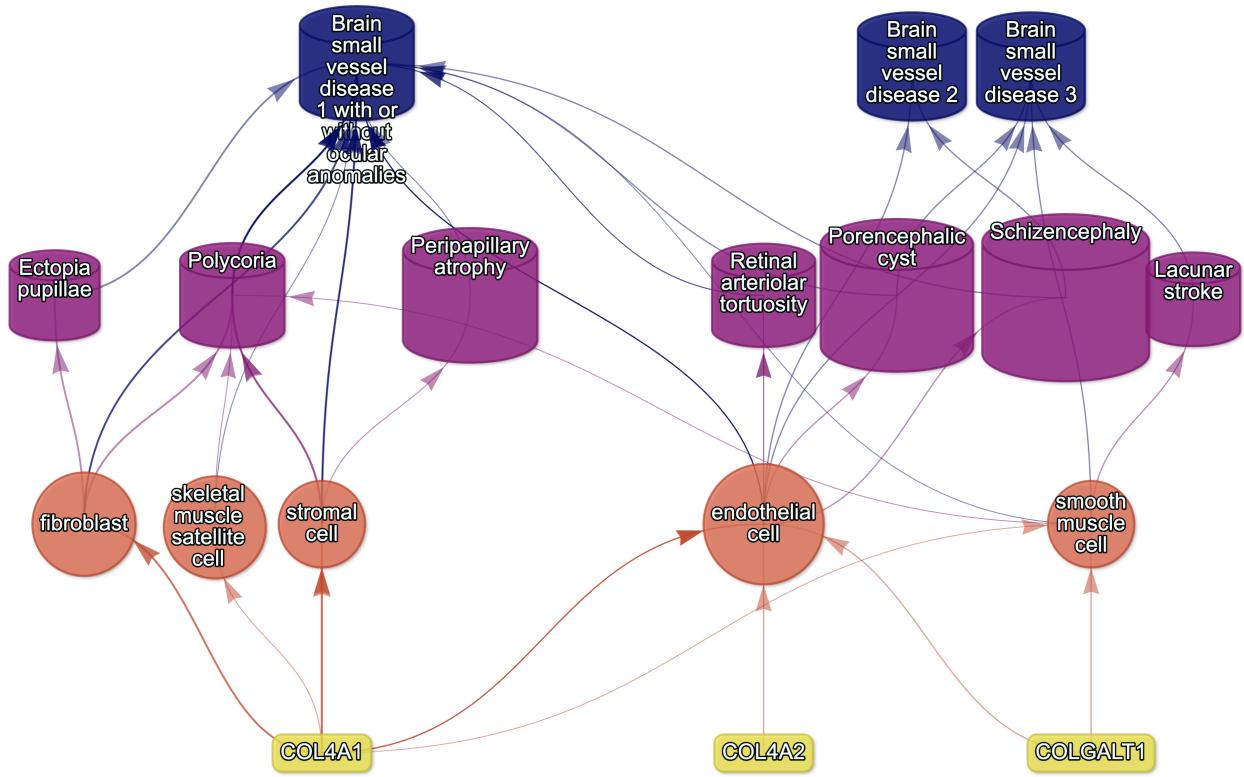


Figure S13: Causal multi-scale network for phenotypes associated with small vessel disease.

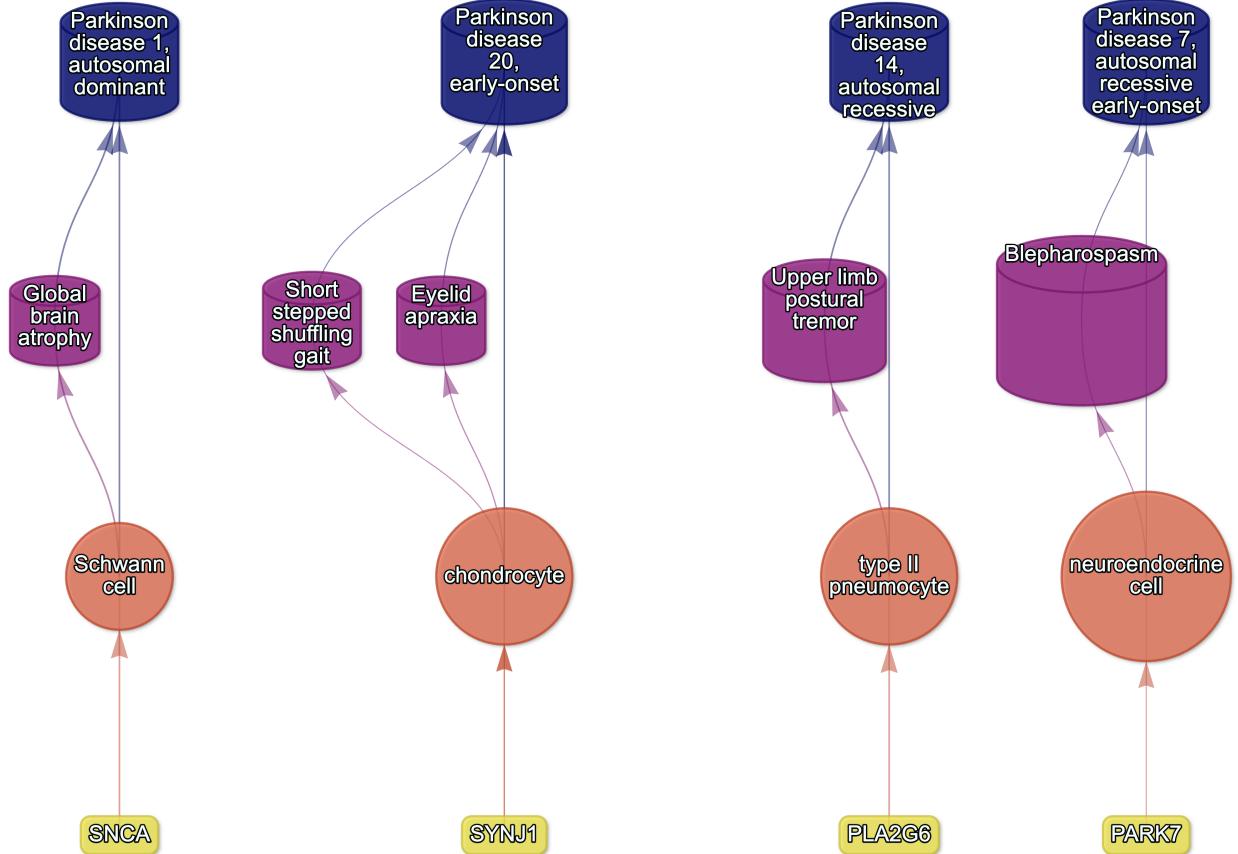


Figure S14: Causal multi-scale network for phenotypes associated with various subtypes of Parkinson's disease.

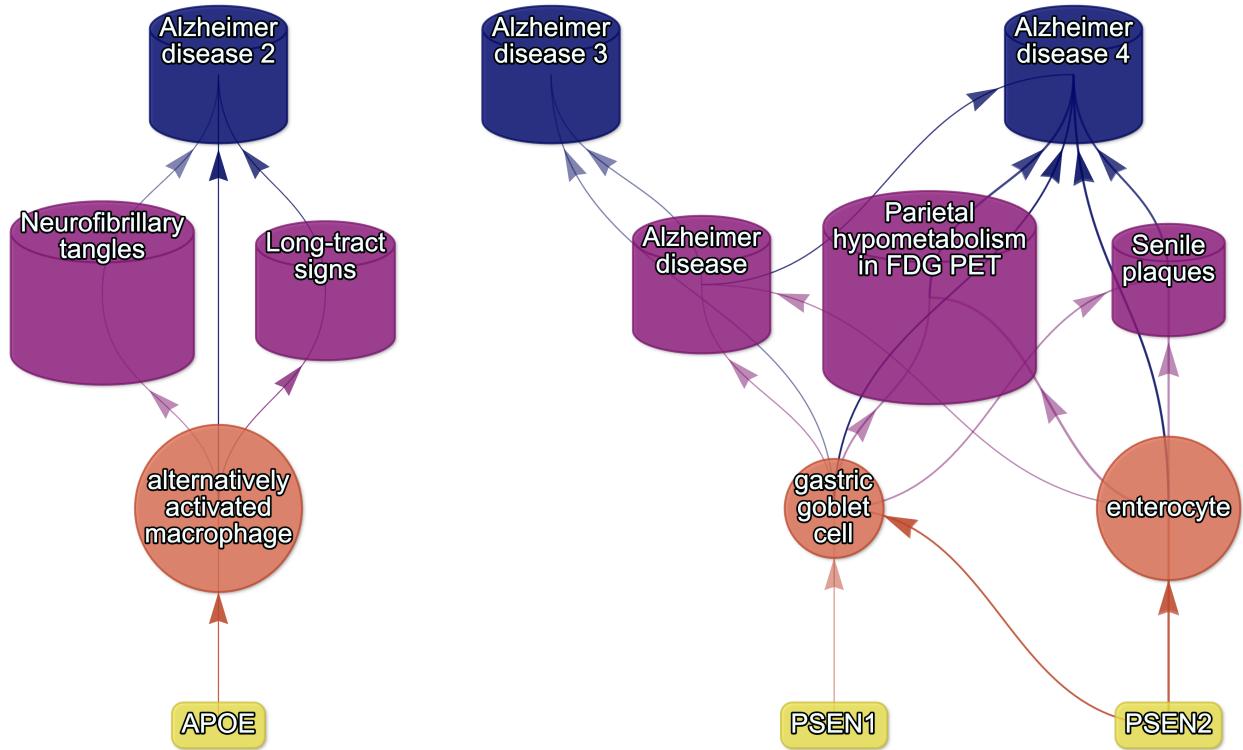


Figure S15: Causal multi-scale network for phenotypes associated with various subtypes of Alzheimer's disease.

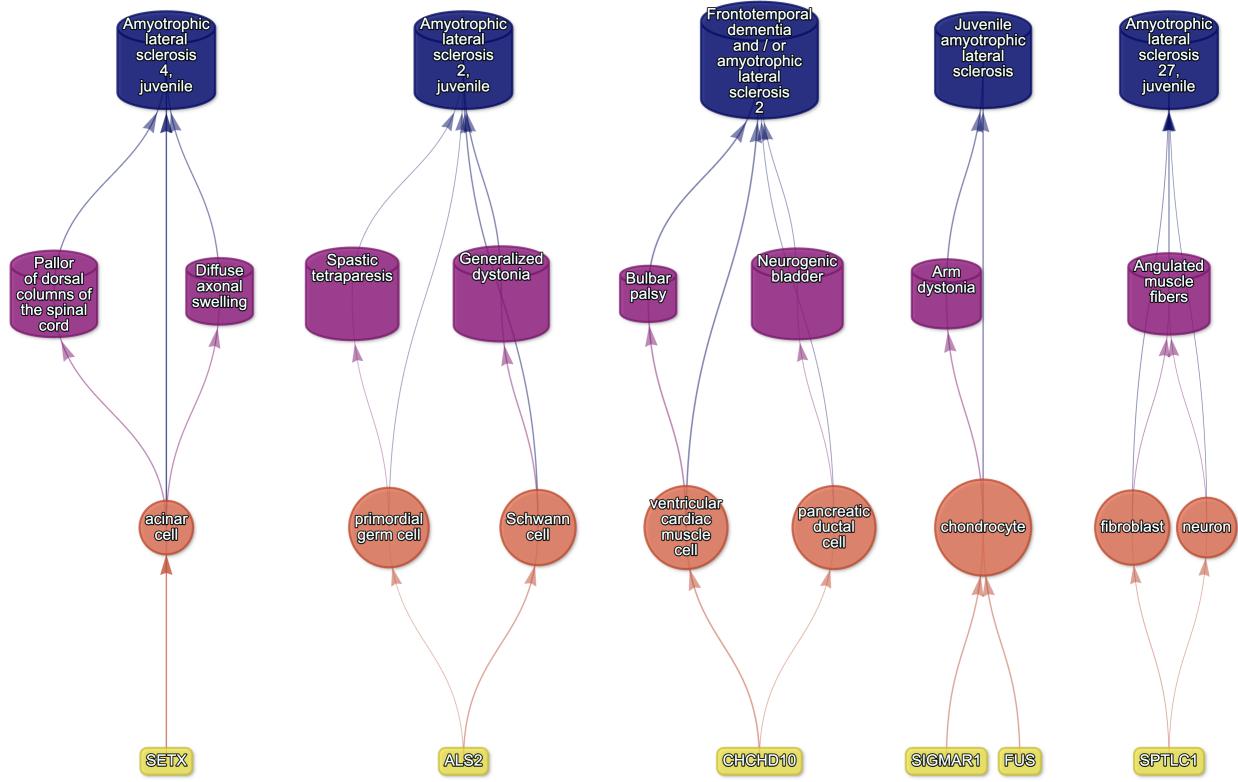


Figure S16: Causal multi-scale network for phenotypes associated with Amyotrophic Lateral Sclerosis (ALS).

968    **Supplementary Tables**

Table S1: **Mappings between HPO phenotypes and other medical ontologies.** “source” indicates the medical ontology and “distance” indicates the cross-ontology distance. “source terms” and “HPO terms” indicates the number of unique IDs mapped from the source ontology and HPO respectively. “mappings” is the total number of cross-ontology mappings within a given distance. Some IDs may have more than one mapping for a given source due to many-to-many relationships.

source	distance	source terms	HPO terms	mappings
ICD10	2	25	23	25
ICD10	3	839	876	1170
ICD9	1	21	21	21
ICD9	2	434	306	462
ICD9	3	1052	920	1816
SNOMED	1	4413	3483	4654
SNOMED	2	75	21	78
SNOMED	3	1796	833	9605
UMLS	1	12898	11601	13049
UMLS	2	140	113	142
UMLS	3	1871	1204	11021

Table S3: **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.

**Table S3: 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
Gene	9. add driver genes	Keep only genes that are driving the association with a given phenotype (inferred by the intersection of phenotype-associated genes and gene with high-specificity quantiles in the target cell type).
Symptom	10. symptom intersection threshold	Minimum proportion of genes overlapping between a symptom gene list (phenotype-associated genes in the context of a particular disease) and the phenotype-cell type association driver genes.
Gene	11. gene frequency threshold	Keep only genes at or above a certain mean frequency threshold (i.e. how frequently a gene is associated with a given phenotype when observed within a disease).
Phenotype	12. prune ancestors	Remove redundant ancestral phenotypes when at least one of their descendants already exist.
All	13. top n	Only return the top N targets per variable group (specified with the “group_vars” argument). For example, setting “group_vars” to “hpo_id” and “top_n” to 1 would only return one target (row) per phenotype ID after sorting.
NA	14. end	NA

Table S2: **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 Landscape).

	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 S4: **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 S6.

HPO branch	Phenotypes		Cell types (overrepresented)	Cell types (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

Table S5: **Encodings for GenCC evidence scores.** Assigned numeric values for the GenCC evidence levels.

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

Table S6: **On-target cell types for each Human Phenotype Ontology (HPO) ancestral branch.** Cell type-phenotype branch pairings were manually curated by comparing high-level HPO terms to terms within the Cell Ontology (CL). Each HPO branch is shown as bolded row dividers. Ancestral CL branch names are shown in the first column, along with the specific CL names and IDs.

CL branch	CL name	CL ID
<b>Abnormality of the cardiovascular system</b>		
cardiocyte	cardiac muscle cell	CL:0000746
cardiocyte	regular atrial cardiac myocyte	CL:0002129
cardiocyte	endocardial cell	CL:0002350
cardiocyte	epicardial adipocyte	CL:1000309
cardiocyte	ventricular cardiac muscle cell	CL:2000046
<b>Abnormality of the endocrine system</b>		
endocrine cell	endocrine cell	CL:0000163
endocrine cell	neuroendocrine cell	CL:0000165
endocrine cell	chromaffin cell	CL:0000166
<b>Abnormality of the eye</b>		
photoreceptor cell / retinal cell	photoreceptor cell	CL:0000210
photoreceptor cell / retinal cell	amacrine cell	CL:0000561
photoreceptor cell / retinal cell	Mueller cell	CL:0000636
photoreceptor cell / retinal cell	retinal pigment epithelial cell	CL:0002586
<b>Abnormality of the immune system</b>		
leukocyte	T cell	CL:0000084
leukocyte	mature neutrophil	CL:0000096
leukocyte	mast cell	CL:0000097
leukocyte	microglial cell	CL:0000129
leukocyte	professional antigen presenting cell	CL:0000145
leukocyte	macrophage	CL:0000235
leukocyte	B cell	CL:0000236
leukocyte	dendritic cell	CL:0000451
leukocyte	monocyte	CL:0000576
leukocyte	plasma cell	CL:0000786
leukocyte	alternatively activated macrophage	CL:0000890
leukocyte	thymocyte	CL:0000893
leukocyte	innate lymphoid cell	CL:0001065
<b>Abnormality of the musculoskeletal system</b>		
cell of skeletal muscle / chondrocyte	chondrocyte	CL:0000138
cell of skeletal muscle / chondrocyte	cell of skeletal muscle	CL:0000188
cell of skeletal muscle / chondrocyte	skeletal muscle satellite cell	CL:0000594
<b>Abnormality of the nervous system</b>		
neural cell	bipolar neuron	CL:0000103
neural cell	granule cell	CL:0000120
neural cell	Purkinje cell	CL:0000121
neural cell	glial cell	CL:0000125
neural cell	astrocyte	CL:0000127
neural cell	oligodendrocyte	CL:0000128
neural cell	microglial cell	CL:0000129
neural cell	neuroendocrine cell	CL:0000165
neural cell	chromaffin cell	CL:0000166
neural cell	photoreceptor cell	CL:0000210
neural cell	inhibitory interneuron	CL:0000498
neural cell	neuron	CL:0000540
neural cell	neuronal brush cell	CL:0000555
neural cell	amacrine cell	CL:0000561
neural cell	GABAergic neuron	CL:0000617
neural cell	Mueller cell	CL:0000636
neural cell	glutamatergic neuron	CL:0000679
neural cell	retinal ganglion cell	CL:0000740
neural cell	retina horizontal cell	CL:0000745
neural cell	Schwann cell	CL:0002573
neural cell	retinal pigment epithelial cell	CL:0002586
neural cell	visceromotor neuron	CL:0005025
neural cell	sympathetic neuron	CL:0011103
<b>Abnormality of the respiratory system</b>		
respiratory epithelial cell / epithelial cell of lung	type II pneumocyte	CL:0002063
respiratory epithelial cell / epithelial cell of lung	epithelial cell of lower respiratory tract	CL:0002632

Table S7: **Results of permutation tests evaluating relationships between phenotype information content and gene number, cell type number, and effect size.** Correlations between phenotypes and cell types may not be independent due to the hierarchical nature of the HPO. To account for this, for each pair of variables we computed the observed correlation (*observed cor*), the mean null correlation (*nullc cor*) from 1000 permutations, and the 95% confidence interval of the null correlations (*null CI [lower, upper]*).

X	y	observed p	observed cor	empirical p	null cor	null CI [lower, upper]
info_content	cell types	6.91e-62	-0.61	2.23e-308	1.99e-03	[-0.07, 0.08]
info_content	estimate	2.23e-308	0.21	2.23e-308	1.18e-04	[-0.01, 0.01]
info_content	genes	2.23e-308	-0.58	2.23e-308	-1.67e-06	[0, 0]

Table S8: **Results of permutation tests evaluating phenotype-cell type association consistency across cell type datasets (Human Cell Landscape vs. Descartes Human; i.e. inter-CTD) and across developmental stages (foetal vs. Adult; i.e. intra-CTD).** To account for the non-independence between phenotypes, we computed empirical p-values for each Pearson correlation comparison over 1,000 permutations. Here, we report observed correlation (*observed cor*), the mean null correlation (*null cor*), 95% confidence interval of the null correlations (*null CI [lower, upper]*). “sig.” indicates that only association tests with FDR<0.05 were included in the correlation.

comparison	observed p	observed cor	empirical p	null cor	null CI [lower, upper]
Inter-CTD: estimate (all)	0e+00	0.58	2e-308	2e-05	[-0.01, 0]
Inter-CTD: estimate (sig.)	1e-95	0.73	2e-308	-7e-04	[-0.08, 0.09]
Intra-CTD: estimate (all)	0e+00	0.44	2e-308	3e-05	[0, 0]
Intra-CTD: estimate (sig.)	4e-92	0.40	2e-308	4e-04	[-0.04, 0.04]

**Table S9: Some HPO phenotype categories or more biased towards foetal- or adult- versions of the same cell type.** We took the top 50 phenotypes with the greatest bias towards foetal-cell type associations (“Foetal-biased”) and the greatest bias towards adult-cell type associations (“Adult-biased”) and fed each list of terms into ontological enrichment tests to get a summary of the representative HPO branches for each group. The phenotypes most biased towards associations with only the foetal versions of cell type and those biased towards the adult versions of cell types. “FDR” is the False Discovery Rate-adjusted p-value from the enrichment test, “log2-fold enrichment” is the log2 fold-change from the enrichment test, and “depth” is the depth of the enriched HPO term in the ontology.

term	name	FDR	log2-fold enrichment	depth
<b>Foetal-biased</b>				
HP:0005105	Abnormal nasal morphology	7.80e-06	4.5	6
HP:0010938	Abnormal external nose morphology	4.19e-05	5.4	7
HP:0000366	Abnormality of the nose	4.28e-05	3.8	5
HP:0000055	Abnormal female external genitalia morphology	6.32e-04	5.2	6
HP:0000271	Abnormality of the face	1.01e-03	1.9	4
HP:0000234	Abnormality of the head	1.59e-03	1.7	3
HP:0000152	Abnormality of head or neck	1.85e-03	1.6	2
HP:0010460	Abnormality of the female genitalia	3.27e-02	2.8	5
HP:0000811	Abnormal external genitalia	3.27e-02	2.8	5
HP:0000078	Abnormality of the genital system	3.27e-02	1.9	3
<b>Adult-biased</b>				
HP:0010647	Abnormal elasticity of skin	1.13e-05	6.0	5
HP:0008067	Abnormally lax or hyperextensible skin	2.04e-04	6.0	6
HP:0011121	Abnormal skin morphology	4.06e-04	2.4	4
HP:0000951	Abnormality of the skin	1.76e-03	2.1	3
HP:0001574	Abnormality of the integument	1.26e-02	1.6	2
HP:0001626	Abnormality of the cardiovascular system	2.39e-02	1.4	2
HP:0030680	Abnormal cardiovascular system morphology	2.39e-02	1.7	3
HP:0025015	Abnormal vascular morphology	3.99e-02	1.9	4
HP:0030962	Abnormal morphology of the great vessels	3.99e-02	2.7	6

Table S10: Examples of specific phenotypes that are most biased towards associations with only the foetal versions of cell types (“Foetal-biased”) and those biased towards the adult versions of cell types (“Adult-biased”). “p-value difference” is the difference in the association p-values between the foetal and adult version of the equivalent cell type (foetal-adult bias :  $p_{adult} - p_{foetal} = \Delta p \in [-1, 1]$ ).

HPO name	HPO ID	CL ID	CL name	p-value difference
<b>Foetal-biased</b>				
Short middle phalanx of the 2nd finger	HP:0009577	CL:0000138	chondrocyte	0.99
Abnormal morphology of the nasal alae	HP:0000429	CL:0000057	fibroblast	0.95
Abnormal labia minora morphology	HP:0012880	CL:0000499	stromal cell	0.94
Acromesomelia	HP:0003086	CL:0000138	chondrocyte	0.93
Left atrial isomerism	HP:0011537	CL:0000163	endocrine cell	0.92
Fixed facial expression	HP:0005329	CL:0000499	stromal cell	0.92
Migraine without aura	HP:0002083	CL:0000163	endocrine cell	0.92
Truncal ataxia	HP:0002078	CL:0000163	endocrine cell	0.92
Anteverted nares	HP:0000463	CL:0000057	fibroblast	0.91
Short 1st metacarpal	HP:0010034	CL:0000138	chondrocyte	0.90
<b>Adult-biased</b>				
Symblepharon	HP:0430007	CL:0000138	chondrocyte	-0.97
Abnormally lax or hyperextensible skin	HP:0008067	CL:0000057	fibroblast	-0.94
Reduced bone mineral density	HP:0004349	CL:0000057	fibroblast	-0.94
Paroxysmal supraventricular tachycardia	HP:0004763	CL:0000138	chondrocyte	-0.93
Lack of skin elasticity	HP:0100679	CL:0000057	fibroblast	-0.92
Excessive wrinkled skin	HP:0007392	CL:0000057	fibroblast	-0.91
Bruising susceptibility	HP:0000978	CL:0000057	fibroblast	-0.91
Corneal opacity	HP:0007957	CL:0000057	fibroblast	-0.90
Broad skull	HP:0002682	CL:0000138	chondrocyte	-0.90
Emphysema	HP:0002097	CL:0000057	fibroblast	-0.89

Table S11: Hypergeometric test results evaluating the enrichment of existing therapy targets documented in the Therapeutic Target Database (TTD) amongst our prioritised targets. Results are shown separately for gene therapies and all therapies. Column keys: universe=gene universe size, p=hypergeometric test p-value, OR=odds ratio, NPV=negative predictive value, PPV=positive predictive value, FDR=false discovery rate, TP=true positives, FP=false positives, FN=false negatives, TN=true negatives.

status	overlap	universe	p	OR	sensitivity	specificity	PPV	NPV	FDR	TP	FP	FN
<b>Gene therapies</b>												
nonfailed	65	5013	5.56e-05	3.00	0.83	0.38	0.02	0.99	0.98	65	3083	13
failed	0	5013	1.00e+00	0.00	NaN	0.37	0.00	1.00	1.00	0	3148	0
<b>All therapies</b>												
nonfailed	591	6432	1.00e+00	0.23	0.26	0.39	0.19	0.49	0.81	591	2557	1664
failed	125	6432	4.44e-23	0.36	0.27	0.49	0.04	0.90	0.96	125	3023	335