

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

³ Brian M. Schilder^{1*} Kitty B. Murphy¹ Hiranyamaya Dash¹ Yichun Zhang¹
⁴ Robert Gordon-Smith¹ Jai Chapman¹ Momoko Otani¹ Nathan G. Skene¹

⁵ 2025-08-08

1. Department of Brain Sciences, Faculty of Medicine, Imperial College London, London, UK

* Corresponding author: brian_schilder@alumni.brown.edu

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
10 by integrating the Human Phenotype Ontology with whole-body single-cell transcriptomic atlases from
11 embryonic, foetal, and adult samples. This revealed significant associations between 201 cell types and
12 9,575/11,028 (86.7%) phenotypes across 8,628 RDs, substantially expanding knowledge of phenotype–cell
13 type links. We then prioritised phenotypes for clinical impact based on severity (e.g. lethality, motor/mental
14 impairment) and gene therapy compatibility (e.g. cell type specificity, postnatal treatability). All results are
15 reproducible and freely available, including via an interactive web portal ([https://neurogenomics-ukdri.dsi.
16 ic.ac.uk/](https://neurogenomics-ukdri.dsi.ic.ac.uk/)), 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)^{1,2}, 75% of RD patients are children, with a 30%
20 rate mortality by age five³. Diagnosis is challenging due to highly variable presentations, averaging five
21 years⁴, with ~46% misdiagnosed and >75% never diagnosed⁵. Prognosis is similarly difficult. Treatments
22 exist for <5% of RDs⁶ and high development costs for small patient populations deter investment^{7,8}, making
23 these therapies among the world’s most expensive^{9,10}. High-throughput therapeutic discovery could lower
24 costs and speed 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^{11–13}, integrated into diagnostics and linked to other ontologies (e.g. SNOMED CT, UMLS, ICD). Over
28 80% of RDs have known genetic causes¹⁴, with HPO gene annotations curated from OMIM, Orphanet, DECI-
29 PHER, and case reports. Yet gene lists alone lack the tissue and cell type context essential for understanding
30 pathogenesis and improving diagnosis, prognosis, and treatment.

31 Single-cell RNA-seq (scRNA-seq) now enables transcriptome-wide profiling at cellular resolution^{15–17}. Comprehensive
32 atlases such as Descartes Human¹⁸ and Human Cell Landscape¹⁹ cover embryonic to adult stages
33 across tissues, providing gene signatures for hundreds of cell subtypes. Integrating RD gene annotations
34 with 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^{20,21}. 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 scores²². 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 variable
45 effects across cell types. Associations were computed between weighted gene lists for each phenotype (from
46 GenCC evidence²³) and cell type gene expression specificity scores derived from normalised scRNA-seq
47 atlases.

48 We analysed two references: Descartes Human (~4M cells from 15 foetal tissues)¹⁸ and Human Cell Landscape
49 (~703k cells from 49 embryonic, foetal, adult tissues)¹⁹. For each phenotype–cell type pair, we ran linear
50 regression and applied FDR correction. In Descartes Human, 19,929 / 848,078 (2.35%) tests were significant
51 across 77 cell types and 7,340 phenotypes. In Human Cell Landscape, the corresponding values were 1.96%
52 significant tests, 124 cell types, and 9,049 phenotypes, with more phenotypes linked to at least one cell type
53 due to greater cell type diversity and life-stage coverage.

54 Across both atlases, the median number of significant cell types per phenotype was 3, indicating specificity
55 of associations. Overall, 8,628/8,631 (~100%) of diseases had significant cell type associations for at least
56 one phenotype. Full stratified results are provided in Table 2.

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

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

62 A cell type was considered *on-target* for an HPO branch if it belonged to a matching CL branch (Table 4).
63 For each HPO branch (Fig. 2b), we tested whether cell types were more often associated with phenotypes in
64 that branch compared to all others, and identified those overrepresented at FDR<0.05. All 7 HPO branches
65 showed disproportionate associations with on-target cell types from their respective organ systems.

66 We hypothesised that more strongly significant phenotype–cell type associations are more likely to be on-
67 target. Grouping $-\log_{10}(\text{p-values})$ into six bins, we calculated the proportion of on-target cell types per HPO–
68 CL branch pairing. Indeed, this proportion consistently increased with association significance ($\rho = 0.63$,
69 $p = 1.1 \times 10^{-6}$). For example, in nervous system abnormalities neural cells constituted only 23% of all
70 tested cell types, yet made up 46% of associations with $-\log_{10}(\text{p-values}) = 6$. This confirms that stronger



Evidence for Gene 1 causing Phenotype A

	Weight	Studies	Score	
No Known	0	x 0	= 0	
Refuted	0	x 0	= 0	
Disputed	1	x 1	= 1	
Limited	2	x 0	= 0	
Supportive	3	x 2	= 6	Sum
Moderate	4	x 1	= 4	
Strong	5	x 3	= 15	
Definitive	6	x 1	= 6	
Total			= 32	

Phenotype x gene evidence score matrix

	Phenotype A	Phenotype B	Phenotype C	...
Gene 1	32	0	1	...
Gene 2	0	16	0	...
Gene 3	2	12	10	...
...

Descartes Human



Human Cell Landscape



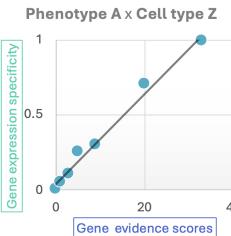
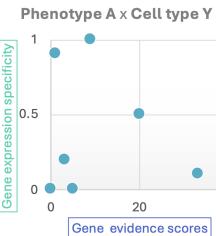
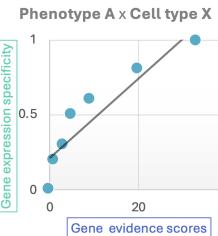
Gene expression specificity in Cell type Z

	Cell Z1	Cell Z2	Cell Z3	Mean	...	Sum (all cell types)	Specificity
Gene 1	0	0	0	0	...	/ 5	= 0
Gene 2	0	1	0	0.33	...	/ 33	= 0.01
Gene 3	9	7	11	9	...	/ 10	= 0.90
...

Cell type x gene expression specificity matrix

	Cell type X	Cell type Y	Cell type Z	...
Gene 1	0.50	0	0	...
Gene 2	0	0.95	0.01	...
Gene 3	0	0.02	0.90	...
...

Generalised Linear Regression Tests

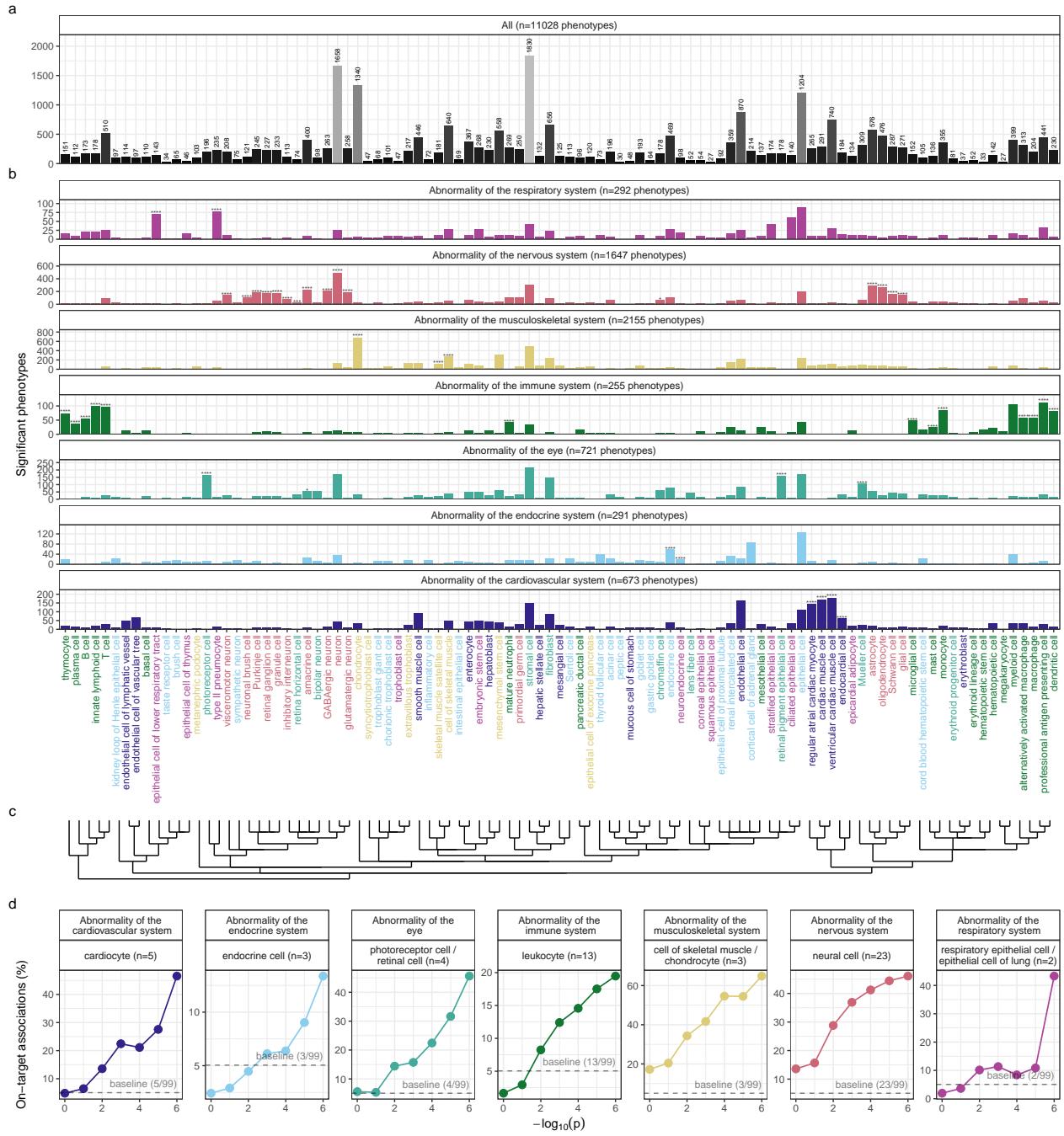


Phenotype-cell type association results

Phenotype	Cell type	P-value	FDR	Z-score
A	X	0.005	0.05	0.25
A	Y	0.98	1	0
A	Z	0.001	0.01	0.90
B	X	1	1	0
B	Y	0.0004	0.004	0.75
B	Z	1	1	0.01
C	X	0.003	0.03	0.20
C	Y	1	1	0
C	Z	0.0007	0.007	0.98
...

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.

⁷¹ associations are more likely to involve on-target cell types, confirming our association strategy captures real
⁷² relationships.



(a) High-throughput analysis reveals cell types underlying thousands of rare disease phenotypes. **a**, Some cell types are much more commonly associated with phenotypes than others. Bar height indicates the total number of significant phenotype enrichments per cell type (FDR<0.05) across all branches of the HPO. **b**, Analyses reveal expected and novel cell type associations within high-level HPO branches. Asterisks above each bar indicate whether that cell type was significantly more often enriched in that branch relative to all other HPO branches, including those not shown here, as a proxy for how specifically that cell type is associated with that branch; FDR<0.0001 (****), FDR<0.001 (**), FDR<0.01 (**), FDR<0.05 (*). **c**, Ontological relatedness of cell types in the Cell Ontology (CL)²⁴. **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

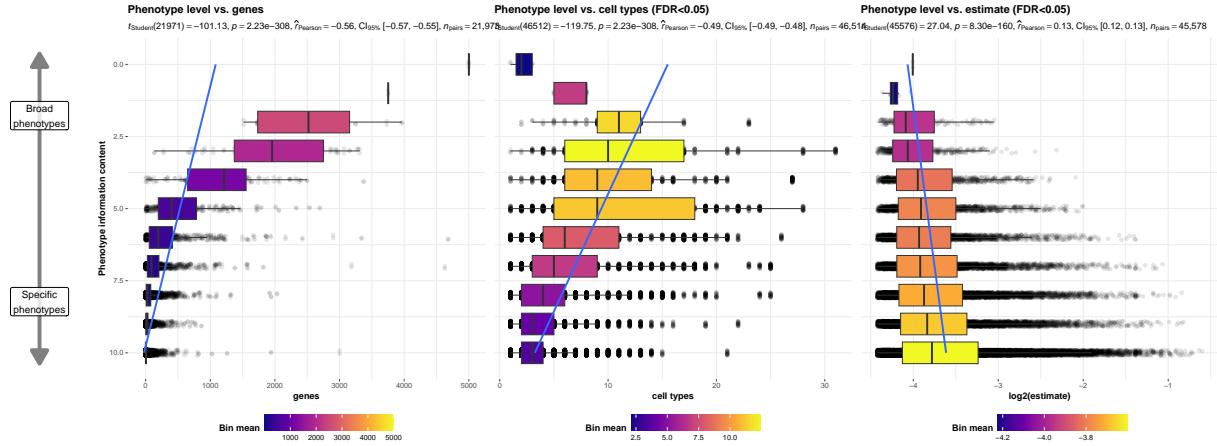
73 **Validation of inter- and intra-dataset consistency**

74 If our methodology works, it should yield consistent phenotype-cell type associations across different datasets.
75 We therefore tested for the consistency of our results across the two single-cell reference datasets (Descartes
76 Human vs. Human Cell Landscape) across the subset of overlapping cell types Fig. 11. In total there were
77 142,285 phenotype-cell type associations to compare across the two datasets (across 10,945 phenotypes and
78 13 cell types annotated to the exact same CL term. We found that the correlation between p-values of
79 the two datasets was high ($\rho=0.91$, $p=5.7 \times 10^{-6}$). Within the subset of results that were significant in
80 both single-cell datasets (FDR<0.05), we found that degree of correlation between the association effect
81 sizes across datasets was even stronger ($\rho=0.82$, $p=5.7 \times 10^{-6}$). We also checked for the intra-dataset
82 consistency between the p-values of the foetal and adult samples in the Human Cell Landscape, showing a
83 very similar degree of correlation as the inter-dataset comparison ($\rho=0.95$, $p=5.0 \times 10^{-15}$). Together,
84 these results suggest that our approach to identifying phenotype-cell type associations is highly replicable
85 and generalisable to new datasets.

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

87 Higher levels of the ontology are broad classes of phenotype (e.g. ‘Abnormality of the nervous system’) while
88 the lower levels can get very detailed (e.g. ‘Spinocerebellar atrophy’). The higher level phenotypes inherit
89 all genes associated with lower level phenotypes, so naturally they have more genes than the lower level
90 phenotypes (Fig. 3a; $\rho=-0.56$, $p=2.2 \times 10^{-308}$).

91 Next, we reasoned that the more detailed and specific a phenotype is, the more likely it is to be driven by
92 one cell type. For example, while ‘Neurodevelopmental abnormality’ could plausibly be driven by any/all
93 cell types in the brain, it is more likely that ‘Impaired visuospatial constructive cognition’ is driven by fewer
94 cell types. This was indeed the case, as we observed a strongly significant negative correlation between the
95 two variables (Fig. 3b; $\rho=-0.49$, $p=2.2 \times 10^{-308}$). We also found that the phenotype-cell type association
96 effect size increased with greater phenotype specificity, reflecting the decreasing overall number of associated
97 cell types at each ontological level (Fig. 3c; $\rho=0.13$, $p=8.3 \times 10^{-160}$).



(a) **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. Boxes are coloured by the mean value within each IC bin (after rounding continuous IC values to the nearest integer).

Figure 3

98 Validation of phenotype-cell type associations using biomedical knowledge graphs

99 To validate phenotype–cell type associations without literature bias, we used the Monarch Knowledge Graph
100 (MKG), a curated database of biomedical concepts and relationships containing 103 known associations²⁵.
101 The MKG served as a benchmark for the field’s current knowledge. For each MKG association, we calculated
102 the proportion of cell types recovered in our results at different ontological distances in the Cell Ontology.
103 Distance 0 indicates the closest possible match (e.g. “monocyte” vs. “monocyte”), with greater distances
104 reflecting progressively broader matches (e.g. distance 1: “monocyte” vs. “classical monocyte”). The theo-
105 retical maximum recall was capped by the percentage of MKG phenotypes for which we identified at least
106 one significant association (FDR_{pc}).

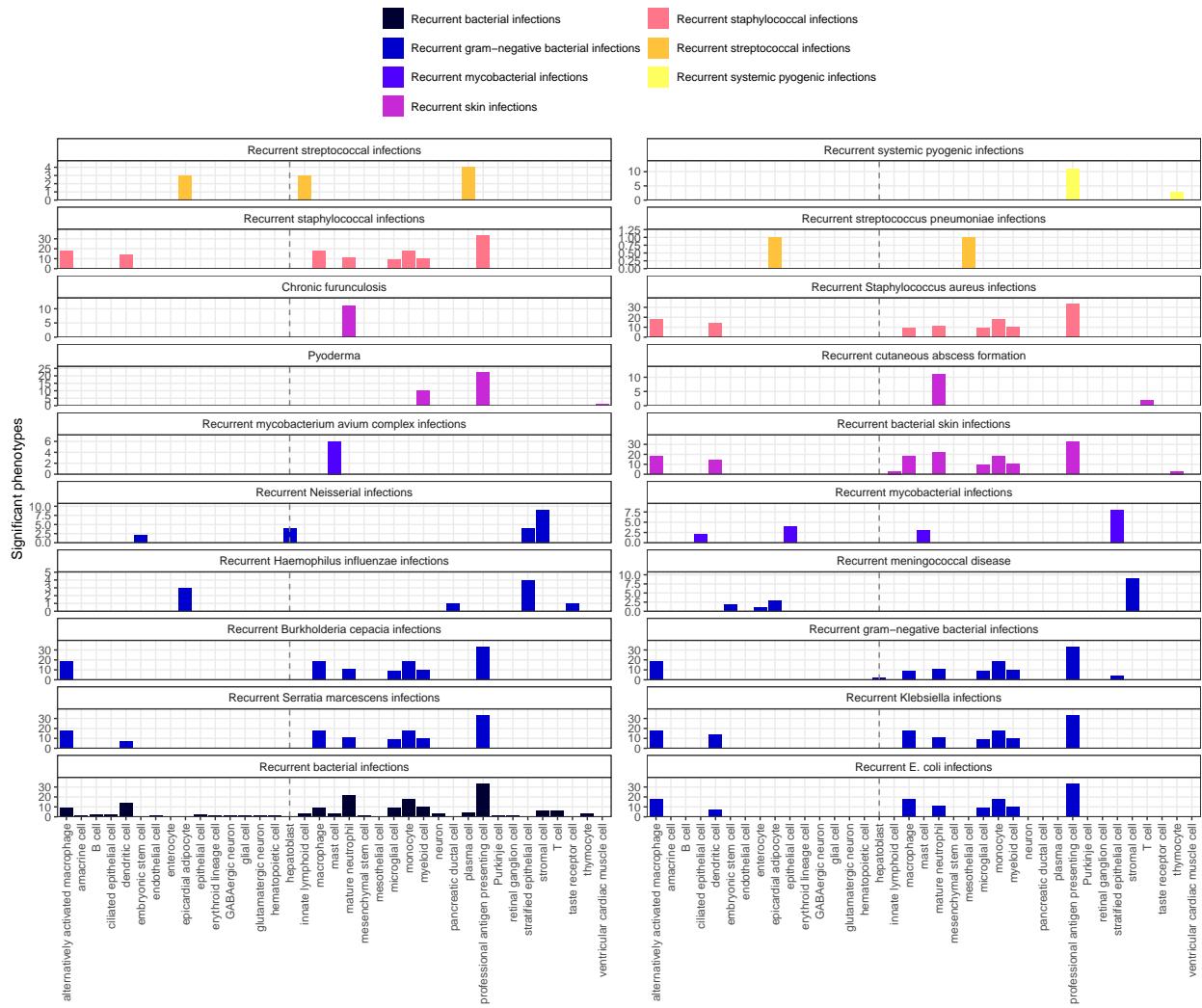
107 Our results included at least one significant cell type for 90% of MKG phenotypes. At distance 0, we recalled
108 57% of associations; at distance 1, recall rose to 77%, reaching a maximum of 90% at the largest allowed
109 distance. Precision could not be computed, as MKG lists only true positives. Overall, these benchmarks
110 show that our approach recovers most known phenotype–cell type associations while generating many novel
111 ones.

112 Phenome-wide analyses discover novel phenotype-cell type associations

113 Having confirmed many phenotype-cell type associations match prior expectations, we explored novel links
114 for undercharacterised phenotypes. ‘Recurrent bacterial infections’ (19 descendants, e.g. staphylococcal,
115 streptococcal, Neisserial) mostly associated with immune cells (e.g. macrophages, dendritic cells, T cells,

monocytes, neutrophils) (Fig. 4). Known links include ‘Recurrent staphylococcal infections’ with myeloid cells^{26–29}, where monocytes were most strongly associated ($FDR=1.0 \times 10^{-30}$, $\beta=0.18$).

A notable novel finding was the association between ‘Recurrent Neisserial infections’ and hepatoblasts (Descartes Human: $FDR=1.1 \times 10^{-6}$, $\beta=8.2 \times 10^{-2}$). This aligns with the role of hepatocyte-derived complement proteins³⁰, complement receptor-expressing Kupffer cells³¹, and complement deficiencies as major Neisserial risk factors^{32,33}. Over 56 complement genes exist³⁴, underscoring the need for targeted, cell type-specific therapeutic strategies. Only hepatoblasts (not mature hepatocytes) showed this association, hinting at early developmental effects.



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

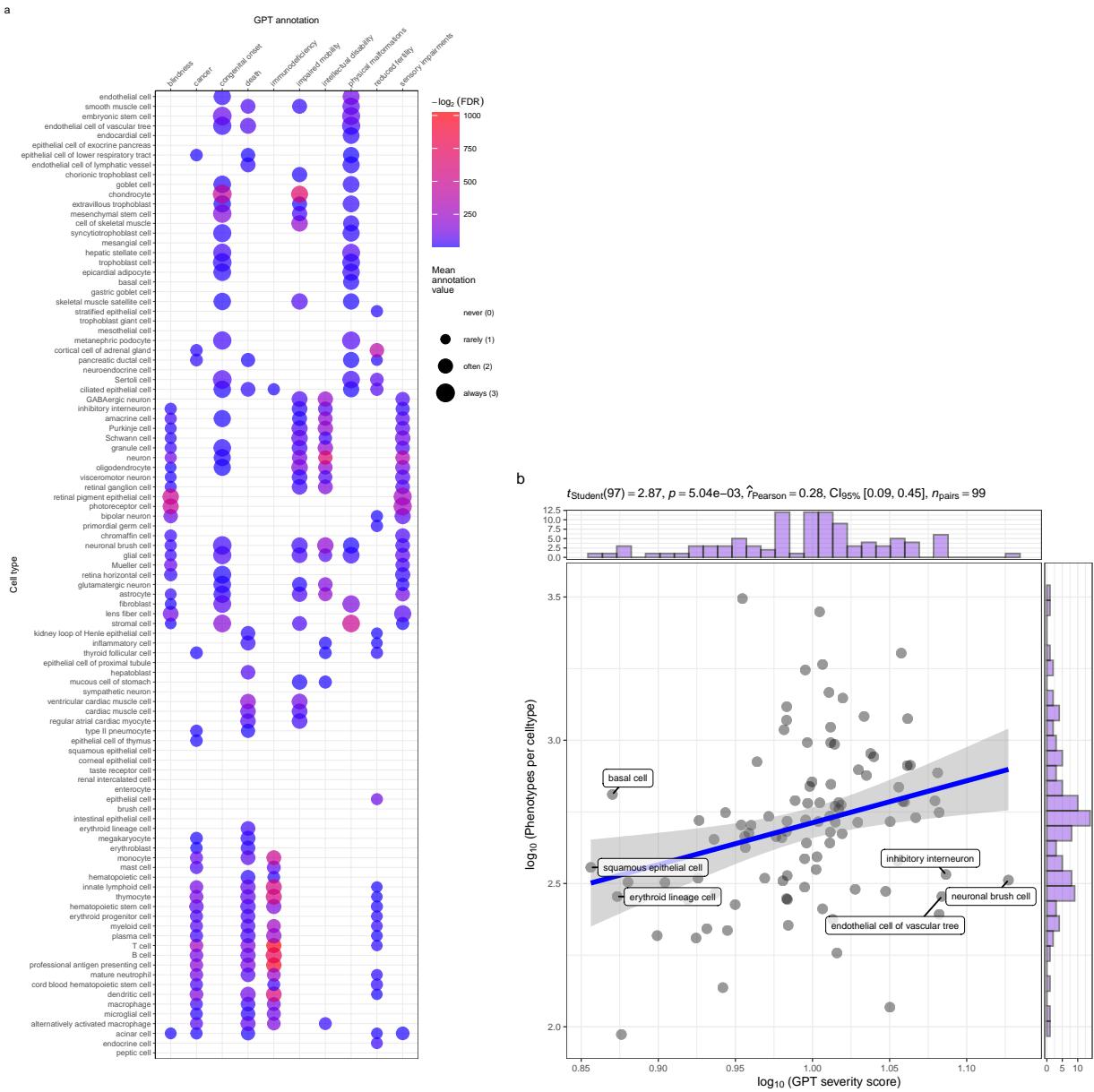
Figure 4

¹²⁴ ‘Recurrent Neisserial infections’ (phenotype of 7 diseases: ‘C5 deficiency’, ‘C6 deficiency’, ‘C7 deficiency’,
¹²⁵ ‘Complement component 8 deficiency, type II’, ‘Complement factor B deficiency’, ‘Complement factor I de-
¹²⁶ ficiency’, ‘Mannose-Binding lectin deficiency’) was also linked to stromal cells ($FDR=4.6 \times 10^{-6}$, $\beta=7.9 \times$
¹²⁷ 10^{-2}), stratified epithelial cells ($FDR=1.7 \times 10^{-23}$, $\beta=0.15$), and embryonic stem cells ($FDR=5.4 \times 10^{-5}$,
¹²⁸ $\beta=7.4 \times 10^{-2}$). Network analysis showed different complement genes (*C5*, *C8*, *C7*) mediated effects via
¹²⁹ hepatoblasts, stratified epithelial cells, and stromal cells, respectively (Supp. Fig. 14). While complement
¹³⁰ genes are widely expressed, subsets act via specific cell types; e.g. only *C6*, *C7*, and *CFI* met driver thresh-
¹³¹ olds in stromal cells. As phenotypes become more granular, we expect convergence toward single-cell-type
¹³² associations. For RNI, the four implicated cell types may represent subtypes with distinct clinical courses
¹³³ or biomarkers, e.g. stromal cell–driven cases differing from stratified epithelial cell–driven ones.

¹³⁴ Prioritising phenotypes based on severity

¹³⁵ Some phenotypes are more severe than others and thus could be prioritised for treatment (e.g. ‘Leukonychia’
¹³⁶ is far less severe than ‘Leukodystrophy’). To systematically rank phenotypes, we used GPT-4 to anno-
¹³⁷ tate severity for 16,982/18,082 (94%) HPO phenotypes²². Benchmarking against ground-truth HPO branch
¹³⁸ annotations showed high accuracy (recall=96%, min=89%, max=100%, SD=4.5%) and strong consistency
¹³⁹ (91%). From these, we computed weighted severity scores (0–100) for all phenotypes. The most severe was
¹⁴⁰ ‘Atrophy/Degeneration affecting the central nervous system’ (*HP:0007367*, score=47), followed by ‘Anen-
¹⁴¹ cephaly’ (*HP:0002323*, score=45). There were 677 phenotypes with score 0 (e.g. ‘Thin toenail’), mean=10
¹⁴² (median=9.4).

¹⁴³ Merging severity scores with significant ($FDR<0.05$) phenotype–cell type associations revealed that neuronal
¹⁴⁴ brush cells had the highest average severity, followed by Mueller cells and glial cells, while megakaryocytes had
¹⁴⁵ the lowest. Numerically encoding GPT annotations (0–3) and applying Wilcoxon tests confirmed expected
¹⁴⁶ links, e.g. retinal pigment epithelial cells with blindness, ventricular cardiac muscle cells with death, and
¹⁴⁷ analogous patterns for reduced fertility, immunodeficiency, impaired mobility, and cancer. Finally, we found
¹⁴⁸ that cell types associated with more phenotypes also tended to have higher mean composite severity ($p=5.0 \times$
¹⁴⁹ 10^{-3} , Pearson=0.28), supporting the idea that broadly involved cell types perform critical physiological
¹⁵⁰ functions whose disruption causes more severe disease.

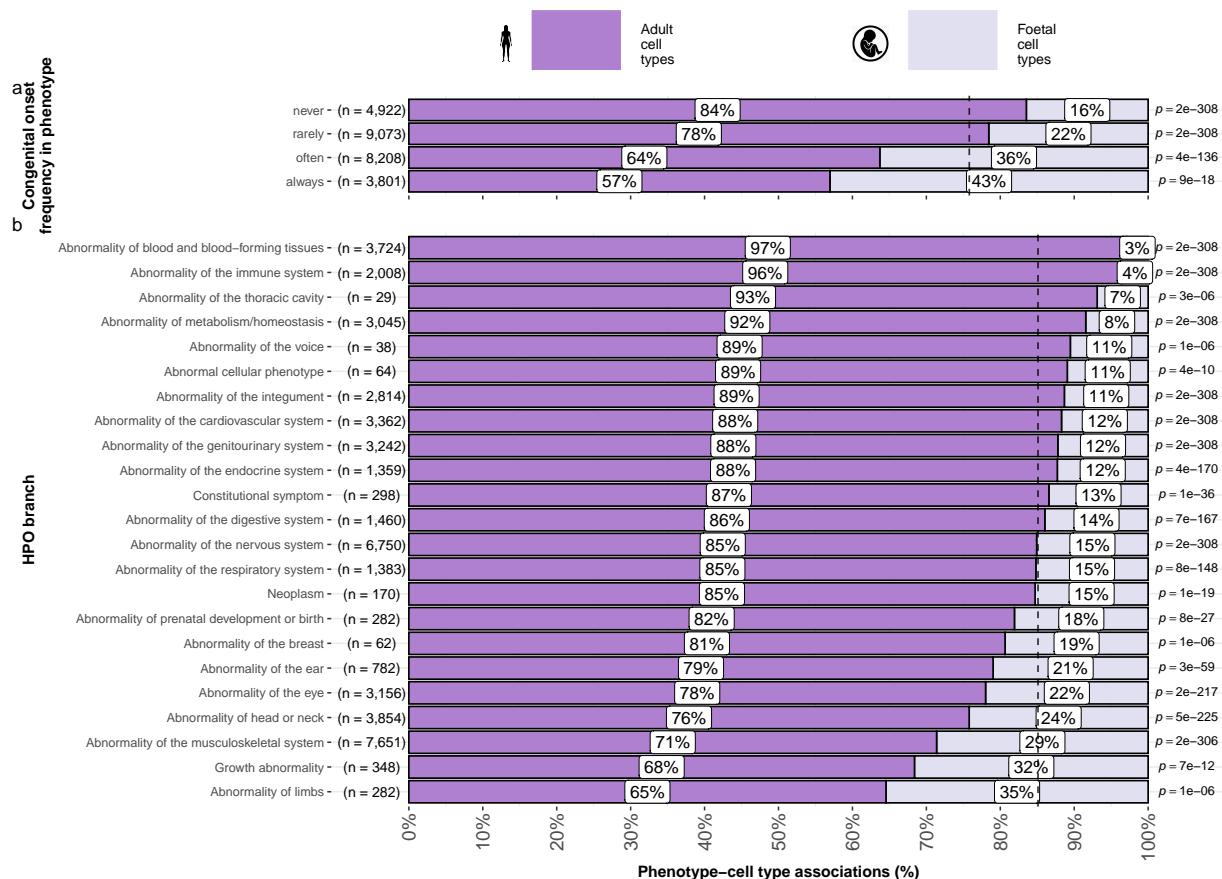


(a) Genetic disruptions to some cell types cause more clinically severe phenotypes than others. **a**, Different cell types are associated with different aspects of phenotypic severity. The dot plot shows the mean encoded frequency value for a given severity annotation (0="never", 1="rarely", 2="often", 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 $-\log_2(\text{FDR})$ when Wilcoxon test FDR values were less than 0.05. All dots with non-significant Wilcoxon tests are instead colored grey. 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 (t_{Student} = Student t-test statistic, p = p-value, $\hat{\rho}_{\text{Pearson}}$ = Pearson correlation coefficient, $CI_{95\%}$ = confidence intervals, n_{pairs} = number of observed data pairs).

Figure 5

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

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



(a) **Foetal vs. adult cell type references provide development context to phenotype aetiology.** **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 χ^2 tests to determine whether the proportion of foetal vs. non-foetal cell types significantly different differs from the proportions expected by chance (the dashed vertical line). The foetal silhouette was generated with DALL-E. The adult silhouette is from phylopic.org and is freely available via CC0 1.0 Universal Public Domain Dedication.

Figure 6

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

¹⁶⁰ ($\hat{V}_{Cramer}=0.22, p<2.2\times10^{-308}$). Branches with the most disproportionate number of foetal cell type associations
¹⁶¹ were ‘Abnormality of limbs’ (35%), ‘Growth abnormality’ (32%), and ‘Abnormality of the musculoskeletal
¹⁶² system’ (29%). The most adult-biased branches were ‘Abnormality of blood and blood-forming tissues’ (97%)
¹⁶³ and ‘Abnormality of the immune system’ (96%).

¹⁶⁴ Some phenotypes involve only foetal or only adult versions of a cell type. We quantified bias by compar-
¹⁶⁵ ing association p-values between foetal and adult versions of the same type (metric range: 1=foetal-only,
¹⁶⁶ -1=adult-only). The top 50 foetal-biased phenotypes revealed were enriched for the HPO branches ‘Abnormal
¹⁶⁷ nasal morphology’ ($p=2.4\times10^{-7}$) and ‘Abnormal external nose morphology’ ($p=2.5\times10^{-6}$), which included
¹⁶⁸ specific phenotypes such as ‘Short middle phalanx of the 2nd finger’. Adult-biased phenotypes were instead
¹⁶⁹ enriched for the branches ‘Abnormal elasticity of skin’ ($p=3.6\times10^{-7}$) and ‘Abnormally lax or hyperexten-
¹⁷⁰ sible skin’ ($p=1.3\times10^{-5}$), with examples like ‘Excessive wrinkled skin’ and ‘Paroxysmal supraventricular
¹⁷¹ tachycardia’. These align with known developmental and age-related processes, supporting our approach for
¹⁷² linking phenotypes to causal cell types.

¹⁷³ Therapeutic target identification

¹⁷⁴ In the above sections, we demonstrated how gene association databases can be used to investigate the cell
¹⁷⁵ types underlying disease phenotypes at scale. While these associations are informative on their own, we
¹⁷⁶ wished to take these results further in order to have a more translational impact. Knowledge of the causal
¹⁷⁷ cell types underlying each phenotype can be incredibly informative for scientists and clinicians in their quest
¹⁷⁸ to study and treat them. Therapeutic targets with supportive genetic evidence have 2.6x higher success rates
¹⁷⁹ in clinical trials^{35–37}. Furthermore, knowing which cell types to target with gene therapy can maximise the
¹⁸⁰ efficacy of highly expensive payloads, and minimise side effects (e.g. immune reaction to viral vectors). Recent
¹⁸¹ biotechnological advances have greatly enhanced our ability to target specific cell types with gene therapy,
¹⁸² making specific and accurate knowledge the correct underlying cell types more pertinent than ever^{20,21}.

¹⁸³ We developed an automated pipeline to identify putative cell type-specific gene targets for each phenotype by
¹⁸⁴ integrating phenotype-cell type association results with primary resources such as GenCC gene-disease rela-
¹⁸⁵ tionships and scRNA-seq atlas datasets, producing a table where each row represented a disease-phenotype-
¹⁸⁶ cell type-gene tetrad. We applied sequential filters to retain only significant phenotype-cell type associations
¹⁸⁷ ($FDR < 0.05$), phenotype-gene pairs with strong causal evidence (GenCC score > 3), phenotypes with high
¹⁸⁸ specificity ($IC > 8$), and gene-cell type links in the top 25% expression specificity quantile, and further
¹⁸⁹ required a symptom intersection > 0.25 when linking cell types to diseases via phenotypes. The filtered re-
¹⁹⁰ sults were ranked by GPT-4 composite severity scores, with only the top 10 tetrads retained per phenotype,
¹⁹¹ yielding compact, high-confidence networks suitable for manual inspection and visualization.

¹⁹² This yielded putative therapeutic targets for 5,252 phenotypes across 4,819 diseases in 201 cell types and
¹⁹³ 3,148 genes (Supp. Fig. 15). While this constitutes a large number of genes in total, each phenotype was

194 assigned a median of 2.0 gene targets (mean=3.3, min=1, max=10). Relative to the number of genes an-
195 notations per phenotype in the HPO overall (median=7.0, mean=62, min=1, max=5,003) this represents
196 a substantial decrease in the number of candidate target genes, even when excluding high-level phenotypes
197 (HPO level>3.0). It is also important to note that the phenotypes in the prioritised targets list are ranked
198 by their severity, allowing us to distinguish between phenotypes with a high medical urgency (e.g. ‘Hydra-
199 nencephaly’) from those with lower medical urgency (e.g. ‘Increased mean corpuscular volume’). This can
200 be useful for clinicians, biomedical scientists, and pharmaceutical manufacturers who wish to focus their
201 research efforts on phenotypes with the greatest need for intervention.

202 Across all phenotypes, epithelial cell were most commonly implicated (838 phenotypes), followed by stromal
203 cell (626 phenotypes), stromal cell (626 phenotypes), neuron (475 phenotypes), chondrocyte (383 pheno-
204 types), and endothelial cell (361 phenotypes). Grouped by higher-order ontology category, ‘Abnormality of
205 the musculoskeletal system’ had the greatest number of enriched phenotypes (959 phenotypes, 857 genes),
206 followed by ‘Abnormality of the nervous system’ (733 phenotypes, 1,138 genes), ‘Abnormality of head or
207 neck’ (543 phenotypes, 986 genes), ‘Abnormality of the genitourinary system’ (443 phenotypes, 695 genes),
208 and ‘Abnormality of the eye’ (377 phenotypes, 545 genes).

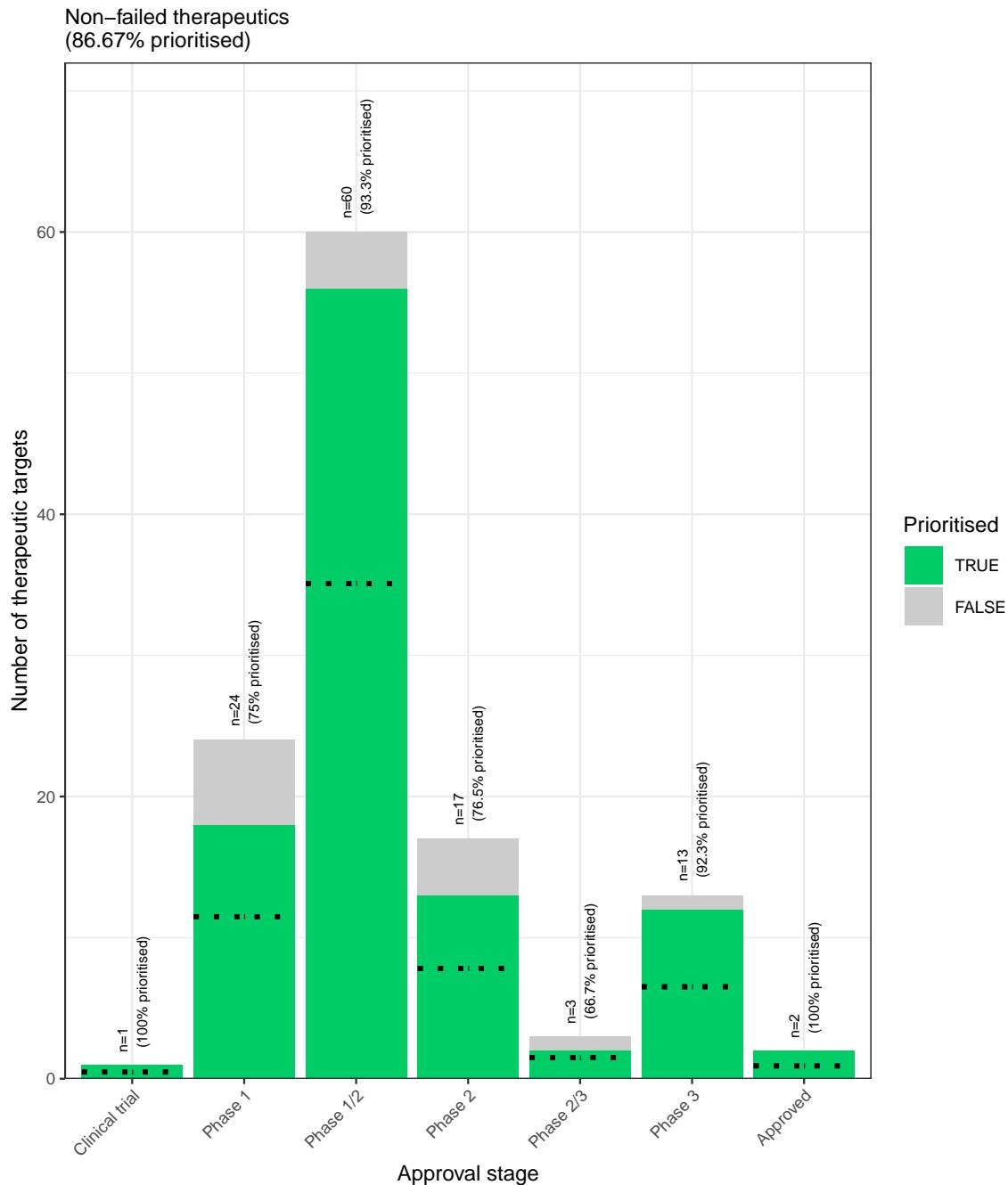
209 Therapeutic target validation

210 To determine whether the genes prioritised by our therapeutic targets pipeline were plausible, we checked
211 what percentage of gene therapy targets we recapitulated. Data on therapeutic approval status was gathered
212 from the Therapeutic Target Database (TTD; release 2025-08-13)³⁸. Overall, we prioritised 87% (120 total)
213 of all non-failed existing gene therapy targets (ie. those which are currently approved, investigative, or
214 undergoing clinical trials). A hypergeometric test confirmed that our prioritised targets were significantly
215 enriched for non-failed gene therapy targets ($p = 1.8 \times 10^{-5}$). For these hypergeometric tests, the background
216 gene set was composed of the union of all phenotype-associated genes in the HPO and all gene therapy
217 targets listed in TTD.

218 Even when considering therapeutics of any kind (Supp. Fig. 16), not just gene therapies, we recapitulated
219 40% of the non-failed therapeutic targets and 0% of the terminated/withdrawn therapeutic targets (n=1,255).
220 Here we found that our prioritised targets were highly significantly depleted for failed therapeutics ($p = 2.2 \times$
221 10^{-142}). This suggests that our multi-scale evidence-based prioritisation pipeline is capable of selectively
222 identifying genes that are likely to be effective therapeutic targets.

223 In addition to aggregate enrichment results, we also provide specific examples of successful gene therapies
224 whose cell type-specific mechanism were recapitulated by our phenotype-cell associations. In particular, our
225 pipeline nominated the gene *RPE65* within ‘retinal pigment epithelial cells’ as the top target for ‘Fundus
226 atrophy’ vision-related phenotypes that are hallmarks of ‘Leber congenital amaurosis, type II’ and ‘Se-
227 vere early-childhood-onset retinal dystrophy’. Indeed, gene therapies targeting *RPE65* within the retina of

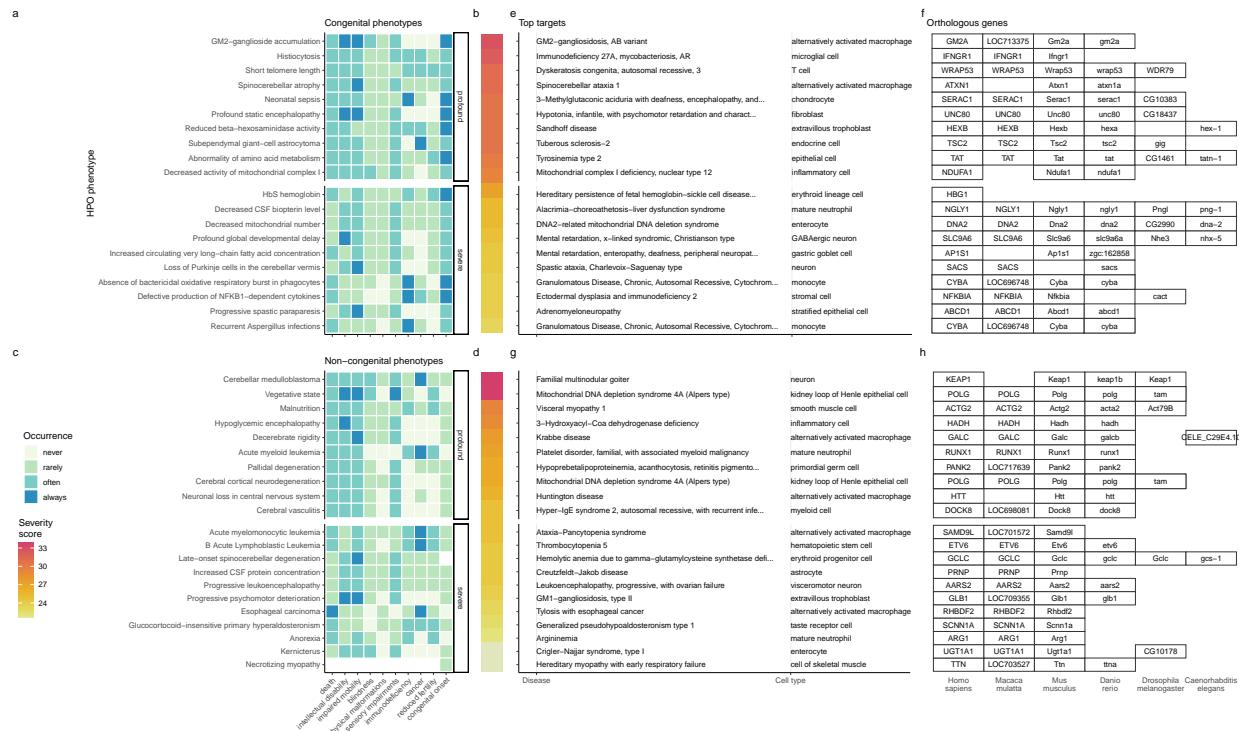
228 patients with these rare genetic conditions are some of the most successful clinical applications of this tech-
229 nology to date, able to restore vision in many cases³⁹. In other cases, a tissue (e.g. liver) may be known to
230 be causally involved in disease genesis, but the precise causal cell types within that tissue remain unknown
231 (e.g. hepatocytes, Kupffer cells, Cholangiocytes, Hepatic stellate cells, Natural killer cells, etc.). Tissue-level
232 investigations (e.g. using bulk transcriptomics or epigenomics) would be dominated by hepatocytes, which
233 comprise 75% of the liver. Our prioritized gene therapy targets can aid in such scenarios by providing the
234 cell type-resolution context most likely to be causal for a given phenotype or set of phenotypes.



(a) **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.

Figure 7

235 **Selected example targets**



(a) **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 Supp. Fig. 18 for network plots of cell type-specific gene therapy targets for several severe phenotypes and their associated diseases.

Figure 8

- 236 From our prioritised targets, we selected four phenotype or disease examples: ‘GM2-ganglioside accumula-
237 tion’, ‘Spinocerebellar atrophy’, ‘Neuronal loss in central nervous system’. To focus on clinically relevant
238 phenotypes and reduce overplotting, we limited selection to those with GPT severity scores above 15 Supp.
239 Fig. 18. Selection was based on severity and network simplicity to allow compact visualisation.
- 240 Tay-Sachs disease (TSD) is a fatal neurodegenerative condition caused by *HEXA* deficiency and ganglioside
241 buildup. We identified alternatively activated macrophages as the cell type most associated with ‘GM2-
242 ganglioside accumulation’ Supp. Fig. 18. This aligns with prior findings of ganglioside accumulation in
243 TSD macrophages^{40,41,42,43}. Our results support macrophages as causal in TSD and the most promising
244 therapeutic target.

245 Spinocerebellar atrophy is a progressive neurodegenerative phenotype in disorders like Spinocerebellar ataxia.
246 Our pipeline implicates M2 macrophages ('Alternatively activated macrophages') as the only causal cell type
247 Supp. Fig. 18. This suggests Purkinje cell loss is downstream of macrophage dysfunction, consistent with
248 microglial roles in neurodegeneration^{44–46}. Our findings provide the first statistically supported link between
249 risk genes and this cell type, which is supported by relevant mouse models (e.g. *Atxn1*, *Pnpla6*) that replicate
250 cellular and behavioural disease phenotypes.

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

255 Additional examples of therapeutics targets include; cardiac muscle and endothelial cells in pheontypes
256 associated with respiratory failure (Supp. Fig. 19a), microglia in frontal lobe dementia (Supp. Fig. 20),
257 chondrocytes in lethal skeletal dyplasia (Supp. Fig. 21), endothelial cells in small vessel disease (Supp.
258 Fig. 22), oligodendrocytes and neurons in Parkinson's disease (Supp. Fig. 23). and multiple gastrointestinal
259 and immune cell types in Alzheimer's disease (Supp. Fig. 24). For further details please refer to the
260 Supplementary Results.

261 **Mappings**

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

268 **Discussion**

269 Investigating rare diseases (RDs) at the phenotype level offers advantages in research and clinical medicine.
270 Most RDs have a single causal gene (7,671/8,631 = 89%). Therefore aggregating genes into phenotype-based
271 sets enables well-powered analyses (mean ~76 genes/phenotype). Phenotypes often converge on shared molec-
272 ular pathways, and a phenotype-centric approach captures interindividual variation disease presentations.
273 This requires mapping the molecular and cellular mechanisms behind each phenotype, which we achieve here
274 at phenome scale.

275 Across 201 cell types and 11,047 phenotypes, we found >46,514 significant phenotype–cell type relationships,
276 enabling multi-scale mechanistic tracing. Results replicate known links, add cellular context, and uncover
277 novel associations. Extensive benchmarking confirmed expected associations, aided by comprehensive phe-

278 notype and cell type ontologies. Key findings include enrichment of anatomically matched associations,
279 correlation of phenotype specificity with association strength, precise subtypes for recurrent infections, and
280 links between congenital onset frequency and developmental cell types.

281 Less than 5% of RDs have treatments⁶. However, advances in CRISPR, prime editing, antisense oligonu-
282 cleotides, viral/lipid delivery^{47–50} are accelerating. The FDA’s new program⁵¹ aims to expand gene/cell
283 therapy approvals in years rather than decades⁵², but success depends on understanding the causal mech-
284 anisms of each RD. Here, we built a reproducible pipeline for nominating cell type–resolved therapeutic
285 targets (Fig. 8), factoring in association strength, gene specificity, severity, therapy delivery suitability, and
286 model translatability. We recovered 87% of active gene therapies, confirming strong enrichment. Highlighted
287 cases include macrophage-driven phenotypes in Tay-Sachs, spinocerebellar ataxia, and Alzheimer’s disease,
288 pinpointing specific phenotypes (e.g. neurofibrillary tangles) to causal cell types.

289 Current limitations of our study include missing certain rare cell subtypes and states (e.g. immune cell
290 responses, diseased states, aging) and incomplete knowledge of gene–phenotype associations. With the ex-
291 pectation that data will continue to improve over time, our pipeline is fully containerised and documented
292 for end-to-end reproducibility. Comprehensive, ontology-driven frameworks like ours enable discovery, di-
293 agnosis, and basket trial design for shared molecular etiologies across many diseases⁵³. Furthermore, we
294 invite collaborations to validate and translate these predictions, and have publicly released all results via
295 R packages and the Rare Disease Celltyping Portal (<https://neurogenomics-ukdri.dsi.ic.ac.uk/>) to support
296 broad access for researchers, clinicians, and patients.

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

300 Methods

301 Human Phenotype Ontology

302 The latest version of the HPO (release 2024-02-08) was downloaded from the EMBL-EBI Ontology Lookup
303 Service⁵⁴ and imported into R using the `HPOExplorer` package. This R object was used to extract ontolog-
304 ical relationships between phenotypes as well as to assign absolute and relative ontological levels to each
305 phenotype. The latest version of the HPO phenotype-to-gene mappings and phenotype annotations were
306 downloaded from the official HPO GitHub repository and imported into R using `HPOExplorer`. This contains
307 lists of genes associated with phenotypes via particular diseases, formatted as three columns in a table (gene,
308 phenotype, disease).

309 However, not all genes have equally strong evidence of causality with a disease or phenotype, especially when
310 considering that the variety of resources used to generate these annotations (OMIM, Orphanet, DECIPHER)

311 use variable methodologies (e.g. expert-curated review of the medical literature vs. automated text mining
 312 of the literature). Therefore we imported data from the Gene Curation Coalition (GenCC)^{55,56}, which (as
 313 of 2025-08-02) 24,112 evidence scores across 7,566 diseases and 5,533 genes. Evidence scores are defined
 314 by GenCC using a standardised ordinal rubric which we then encoded as a semi-quantitative score ranging
 315 from 0 (no evidence of disease-gene relationship) to 6 (strongest evidence of disease-gene relationship) (see
 316 Table 5). As each Disease-Gene pair can have multiple entries (from different studies) with different levels
 317 of evidence, we then summed evidence scores per Disease-Gene pair to generate aggregated Disease-by-Gene
 318 evidence scores. This procedure can be described as follows.

319 Let us denote:

- 320 • D as diseases.
- 321 • P as phenotypes in the HPO.
- 322 • G as genes
- 323 • S as the evidence scores describing the strength of the relationship between each Disease-Gene pair.
- 324 • M_{ij} as the aggregated Disease-by-Gene evidence score matrix.

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

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

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

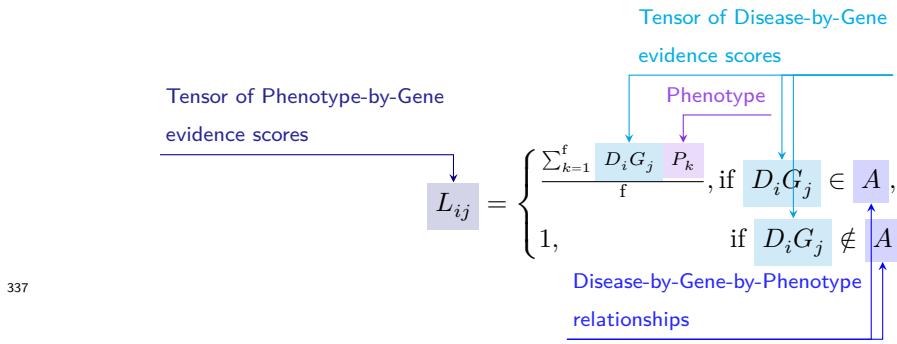
$$A_{ijk} = D_i G_j P_k$$

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

334

335

336



337

338

339

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

341

342

343 Histograms of evidence score distributions at each step in processing can be found in Fig. 9.

344 Single-cell transcriptomic atlases

345 In this study, the gene by cell type specificity matrix was constructed using the Descartes Human transcriptome atlas of foetal gene expression, which contains a mixture of single-nucleus and single-cell RNA-seq data (collected with sci-RNA-seq3)¹⁸. This dataset contains 377,456 cells representing 77 distinct cell types across 15 tissues. All 121 human foetal samples ranged from 72 to 129 days in estimated postconceptual age. **346** To independently replicate our findings, we also used the Human Cell Landscape which contains single-cell transcriptomic data (collected with microwell-seq) from embryonic, foetal, and adult human samples across **347** 49 tissues¹⁹.

352 Specificity matrices were generated separately for each transcriptomic atlas using the R package EWCE (**353** v1.11.3)⁵⁷. Within each atlas, cell types were defined using the authors' original freeform annotations **354** in order to preserve the granularity of cell subtypes as well as incorporate expert-identified rare cell types. **355** Cell types were only aligned and aggregated to the level of corresponding Cell Ontology (CL)²⁴ annotations **356** afterwards when generating summary figures and performing cross-atlas analyses. Using the original **357** gene-by-cell count matrices from each single-cell atlas, we computed gene-by-cell type expression specificity **358** matrices as follows. Genes with very no expression across any cell types were considered to be uninformative **359** and were therefore removed from the input gene-by-cell matrix $F(g, i, c)$.

360 Next, we calculated the mean expression per cell type and normalised the resulting matrix to transform it **361** into a gene-by-cell type expression specificity matrix ($S_{g,c}$). In other words, each gene in each cell type had **362** a 0-1 score where 1 indicated the gene was mostly specifically expressed in that particular cell type relative **363** to all other cell types. This procedure was repeated separately for each of the single-cell atlases and can be **364** summarised as:

365

366

Compute mean expression of each gene per cell type

Gene-by-cell type specificity matrix

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

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

367

368

369

370 Phenotype-cell type associations

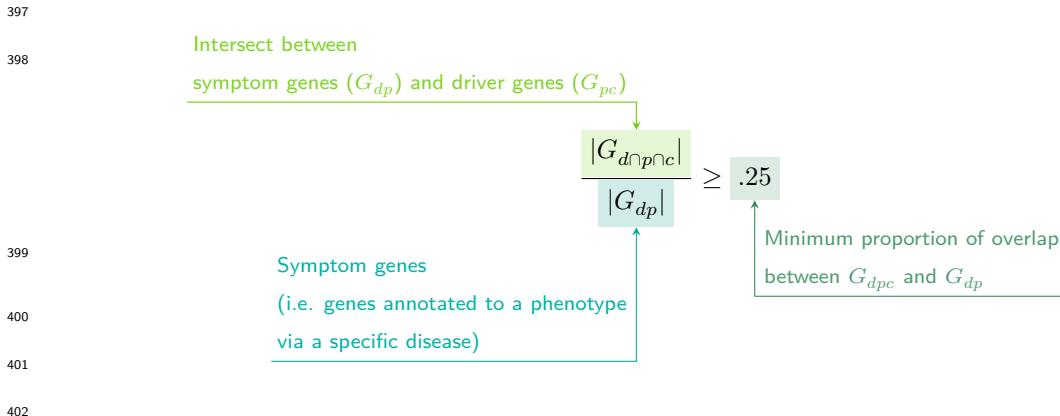
371 To test for relationships between each pairwise combination of phenotype ($n=11,047$) and cell type ($n=201$)
 372 we ran a series of univariate generalised linear models implemented via the `stats::glm` function in R. First,
 373 we filtered the gene-by-phenotype evidence score matrix (L_{ij}) and the gene-by-cell type expression specificity
 374 matrix (S_{gc}) to only include genes present in both matrices ($n=4,949$ genes in the Descartes Human analyses;
 375 $n=4,653$ genes in the Human Cell Landscape analyses). Then, within each matrix any rows or columns with a
 376 sum of 0 were removed as these were uninformative data points that did not vary. To improve interpretability
 377 of the results β coefficient estimates across models (i.e. effect size), we performed a scaling prestep on all
 378 dependent and independent variables. Initial tests showed that this had virtually no impact on the total
 379 number of significant results or any of the benchmarking metrics based on p-value thresholds Fig. 2. This
 380 scaling prestep improved our ability to rank cell types by the strength of their association with a given
 381 phenotype as determined by separate linear models.

382 We repeated the aforementioned procedure separately for each of the single-cell references. Once all results
 383 were generated using both cell type references (2,206,994 association tests total), we applied Benjamini-
 384 Hochberg false discovery rate⁵⁸ (denoted as FDR_{pc}) to account for multiple testing. Of note, we applied
 385 this correction across all results at once (as opposed to each single-cell reference separately) to ensure the
 386 FDR_{pc} was stringently controlled for across all tests performed in this study.

387 Symptom-cell type associations

388 Here we define a symptom as a phenotype as it presents within the context of the specific disease. The features
 389 of a given symptom can be described as the subset of genes annotated to phenotype p via a particular disease
 390 d , denoted as G_{dp} (see Fig. 10). To attribute our phenotype-level cell type enrichment signatures to specific
 391 diseases, we first identified the gene subset that was most strongly driving the phenotype-cell type association
 392 by computing the intersect of genes that were both in the phenotype annotation and within the top 25%
 393 specificity percentile for the associated cell type. We then computed the intersect between symptom genes
 394 (G_{dp}) and driver genes (G_{pc}), resulting in the gene subset $G_{d \cap p \cap c}$. Only $G_{d \cap p \cap c}$ gene sets with 25% or greater

395 overlap with the symptom gene subset (G_{dp}) were kept. This procedure was repeated for all phenotype-cell
 396 type-disease triads, which can be summarised as follows:



403 Validation of expected phenotype-cell type relationships

404 We first sought to confirm that our tests (across both single-cell references) were able to recover expected
 405 phenotype-cell type relationships across seven high-level branches within the HPO (Fig. 2), including ab-
 406 normalities of the cardiovascular system, endocrine system, eye, immune system, musculoskeletal system,
 407 nervous system, and respiratory system. Within each branch the number of significant tests in a given
 408 cell type were plotted (Fig. 2b). Mappings between freeform annotations (the level at which we performed
 409 our phenotype-cell type association tests) provided by the original atlas authors and their closest CL term
 410 equivalents were provided by CellxGene⁵⁹. CL terms along the *x-axis* of Fig. 2b were assigned colours corre-
 411 sponding to which HPO branch showed the greatest number of enrichments (after normalising within each
 412 branch to account for differences in scale). The normalised colouring allows readers to quickly assess which
 413 HPO branch was most often associated with each cell type, while accounting for differences in the number
 414 of phenotypes across branches. We then ran a series of Analysis of Variance (ANOVA) tests to determine
 415 whether (within a given branch) a given cell type was more often enriched ($FDR < 0.05$) within that branch
 416 relative to all of the other HPO branches of an equivalent level in the ontology (including all branches not
 417 shown in Fig. 2b). After applying Benjamini-Hochberg multiple testing correction⁵⁸ (denoted as $FDR_{b,c}$),
 418 we annotated each respective branch-by-cell type bar according to the significance (**** : $FDR_{b,c} < 1e-04$,
 419 *** : $FDR_{b,c} < 0.001$, ** : $FDR_{b,c} < 0.01$, * : $FDR_{b,c} < 0.05$). Cell types in Fig. 2a-b were ordered along
 420 the *x-axis* according to a dendrogram derived from the CL ontology (Fig. 2c), which provides ground-truth
 421 semantic relationships between all cell types (e.g. different neuronal subtypes are grouped together).

422 As an additional measure of the accuracy of our phenotype-cell types test results we identified conceptually
 423 matched branches across the HPO and the CL (Fig. 2d and Table 6). For example, ‘Abnormality of the
 424 cardiovascular system’ in the HPO was matched with ‘cardiocytes’ in the CL which includes all cell types
 425 specific to the heart. Analogously, ‘Abnormality of the nervous system’ in the HPO was matched with ‘neural

426 cell' in the CL which includes all descendant subtypes of neurons and glia. This cross-ontology matching
427 was repeated for each HPO branch and can be referred to as on-target cell types. Within each branch, the
428 $-\log_{10}(FDR_{pc})$ values of on-target cell types were binned by rounding to the nearest integer (*x-axis*) and
429 the percentage of tests for on-target cell types relative to all cell types were computed at each bin (*y-axis*)
430 (Fig. 2d). The baseline level (dotted horizontal line) illustrates the percentage of on-target cell types relative
431 to the total number of observed cell types. Any percentages above this baseline level represent greater than
432 chance representation of the on-target cell types in the significant tests.

433 Validation of inter- and intra-dataset consistency

434 We tested for inter-dataset consistency of our phenotype-cell type association results across different single-
435 cell reference datasets (Descartes Human and Human Cell Landscape). For all tests reported here, the
436 relevant association metrics (p-values or effect size) were first averaged to the level of ancestral HPO terms
437 (5 levels down the hierarchy) to reduce figure size. For association tests with exactly matching Cell Ontology
438 ID across the two references, we tested for a relationship between the p-values generated with each of the
439 references by fitting linear regression model (`stats::lm` via the R function `ggstatsplot::ggscatterstats`).
440 Next, we performed an additional linear regression between the effect sizes (each GLM model's R^2 estimates
441 after applying a \log_2 fold-change transformation) of all significant phenotype-cell type associations ($FDR <$
442 0.05) with exactly matching cell types across the two references.

443 We also tested for intra-dataset consistency within the Human Cell Landscape by running additional linear
444 regressions between the phenotype-cell type association test statistics of the foetal and the adult samples (us-
445 ing both p-values and model R^2 estimates). While we would not expect the same exact cell type associations
446 across different developmental stages, we would nevertheless expect there to be some degree of correlation
447 between the developing and mature versions of the same cell types.

448 More specific phenotypes are associated with fewer genes and cell types

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

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

455 Next, IC scores were quantised into 10 bins using the `ceiling` R function to improve visualisation. We

456 then performed a series of linear regressions between phenotype binned IC scores and: 1) number of genes
457 annotated per HPO phenotype, 2) the number of significantly associated cell types per HPO phenotype, and
458 3) the model estimate of each significant phenotype-cell type associations (at FDR < 0.05) after taking the
459 log of the absolute value ($\log_2(|estimate|)$).

460 **Monarch Knowledge Graph recall**

461 Finally, we gathered known phenotype-cell type relationships from the Monarch Knowledge Graph (MKG),
462 a comprehensive database of links between many aspects of disease biology²⁵. This currently includes 103
463 links between HPO phenotypes (n=103) and CL cell types (n=79). Of these, we only considered the 82
464 phenotypes that we were able to test given that our ability to generate associations was dependent on
465 the existence of gene annotations within the HPO. We considered instances where we found a significant
466 relationship between exactly matching pairs of HPO-CL terms as a hit.

467 However, as the cell types in MKG were not necessarily annotated at the same level as our single-cell refer-
468 ences, we considered instances where the MKG cell type was an ancestor term of our cell type (e.g. ‘myeloid
469 cell’ vs. ‘monocyte’), or *vice versa*, as hits. We also adjusted ontological distance by computing the ratio
470 between the observed ontological distance and the smallest possible ontological distance for that cell type
471 given the cell type that were available in our references ($dist_{adjusted} = (\frac{dist_{observed}+1}{dist_{minimum}+1}) - 1$). This provides
472 a way of accurately measuring how dissimilar our identified cell types were for each phenotype-cell type
473 association (Fig. 12).

474 **Prioritising phenotypes based on severity**

475 Only a small fraction of the the phenotypes in HPO (<1%) have metadata annotations containing informa-
476 tion on their time course, consequences, and severity. This is due to the time-consuming nature of manually
477 annotating thousands of phenotypes. To generate such annotations at scale, we previously used Generative
478 Pre-trained Transformer 4 (GPT-4), a large language model (LLM) as implemented within OpenAI’s Appli-
479 cation Programming Interface (API)²². After extensive prompt engineering and ground-truth benchmarking,
480 we were able to acquire annotations on how often each phenotype directly causes intellectual disability, death,
481 impaired mobility, physical malformations, blindness, sensory impairments, immunodeficiency, cancer, re-
482 duced fertility, or is associated with a congenital onset. These criteria were previously defined in surveys
483 of medical experts as a means of systematically assessing phenotype severity⁶⁰. Responses for each metric
484 were provided in a consistent one-word format which could be one of: ‘never’, ‘rarely’, ‘often’, ‘always’. This
485 procedure was repeated in batches (to avoid exceeding token limits) until annotations were gathered for
486 16,982/18,082 HPO phenotypes.

487 We then encoded these responses into a semi-quantitative scoring system (‘never’=0, ‘rarely’=1, ‘often’=2,
488 ‘always’=3), which were then weighted by multiplying a semi-subjective scoring of the relevance of each

489 metric to the concept of severity on a scale from 1.0-6.0, with 6.0 being the most severe ('death'=6,
 490 'intellectual_disability'=5, 'impaired_mobility'=4, 'physical_malformations'=3, 'blindness'=4, 'sen-
 491 sory_impairments'=3, 'immunodeficiency'=3, 'cancer'=3, 'reduced_fertility'=1, 'congenital_onset'=1).
 492 Finally, the product of the score was normalised to a quantitative severity score ranging from 0-100, where
 493 100 is the theoretical maximum severity score. This phenotype severity scoring procedure can be expressed
 494 as follows.

495 Let us denote:

- 496 • p : a phenotype in the HPO.
- 497 • j : the identity of a given annotation metric (i.e. clinical characteristic, such as 'intellectual disability'
498 or 'congenital onset').
- 499 • W_j : the assigned weight of metric j .
- 500 • F_j : the maximum possible value for metric j , equal to 3 ("always"). This value is equivalent across all
501 j annotations.
- 502 • F_{pj} : the numerically encoded value of annotation metric j for phenotype p .
- 503 • NSS_p : the final composite severity score for phenotype p after applying normalisation to align values
504 to a 0-100 scale and ensure equivalent meaning regardless of which other phenotypes are being analysed
505 in addition to p . This allows for direct comparability of severity scores across studies with different
506 sets of phenotypes.

507 Sum of weighted annotation values
 508 across all metrics
 509 Numerically encoded annotation value
 510 of metric j for phenotype p
 511 Weight for metric j
 512

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

510 Theoretical maximum severity score

513 Using the numerically encoded GPT annotations (0="never", 1="rarely", 2="often", 3="always") we com-
 514 puted the mean encoded value per cell type within each annotation. One-sided Wilcoxon rank-sum tests
 515 were run using the `rstatix::wilcox_test()` function to test whether each cell type was associated with
 516 more severe phenotypes relative to all other cell types. This procedure was repeated for severity annotation
 517 independently (death, intellectual disability, impaired mobility, etc.) Fig. 5a. Next, we performed a Pear-

518 son correlation test between the number of phenotypes that a cell type is significantly associated with (at
519 FDR<0.05) has a relationship with the mean composite GPT severity score of those phenotypes (Fig. 5b).
520 This was performed using the `ggstatsplot::ggscatterstats()` R function.

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

522 The GPT-4 annotations also enabled us to assess whether foetal cell types were more often significantly
523 associated with congenital phenotypes in our Human Cell Landscape results as this single-cell reference
524 contained both adult and foetal versions of cell types (Fig. 6). To do this, we performed a chi-squared (χ^2)
525 test on the proportion of significantly associated cell types containing any of the substrings ‘fetal’, ‘fetus’,
526 ‘primordial’, ‘hESC’ or ‘embryonic’ (within cell types annotations from the original Human Cell Landscape
527 authors¹⁹) vs. those associated without, stratified by how often the corresponding phenotype had a congenital
528 onset according to the GPT phenotype annotations (including ‘never’, ‘rarely’, ‘often’, ‘always’). In addition,
529 a series of χ^2 tests were performed within each congenital onset frequency strata, to determine whether the
530 observed proportion of foetal cell types vs. non-foetal cell types significantly deviated from the proportions
531 expected by chance.

532 We next tested whether the proportion of tests with significant associations with foetal cell types varied
533 across the major HPO branches using a χ^2 test. We also performed separate χ^2 test within each branch to
534 determine whether the proportion of significant associations with foetal cell types was significantly different
535 from chance.

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

542 In order to summarise higher-order HPO phenotype categories that were most biased towards foetal
543 or adult cell types, ontological enrichment tests were run on the phenotypes with the top/bottom
544 50 greatest/smallest foetal-adult bias scores. The enrichment tests were performed using the
545 `simona::dag_enrich_on_offsprings` function, which uses a hypergeometric test to determine whether a
546 list of terms in an ontology are enriched for offspring terms (descendants) of a given ancestor term within
547 the ontology. Phenotypes categories with an HPO ontological enrichment a p-value < 0.05 were considered
548 significant.

549 We were similarly interested in which higher-order cell type categories tended to be most commonly associated
550 with these strongly foetal-/adult-biased phenotype s. Another set of ontological enrichment tests were run on
551 the cell types associated with the top/bottom 50 phenotypes from the previous analysis. The CL ontology-

552 aligned IDs for each group cell types were fed into the `simona::dag_enrich_on_offsprings` using the CL
553 ontology. Significantly enriched cell type categories were defined as those with a CL ontological enrichment
554 p-value < 0.05.

555 Therapeutic target identification

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

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

575 The entire target prioritisation procedure can be replicated with a single function: `MSTExplorer::prioritise_targets`.
576 This function automates all of the reference data gathering (e.g. phenotype metadata, cell type metadata,
577 cell type signature reference, gene lengths, severity tiers) and takes a variety of arguments at each step for
578 greater customisability. Each step is described in detail in Table 3. Phenotypes that often or always caused
579 physical malformations (according to the GPT-4 annotations) were also removed from the final prioritised
580 targets list, as these were unlikely to be amenable to gene therapy interventions. Finally, phenotypes were
581 sorted by their composite severity scores such that the most severe phenotypes were ranked the highest.

582 Therapeutic target validation

583 To assess whether our prioritised therapeutic targets were likely to be viable, we computed the overlap
584 between our gene targets and those of existing gene therapies at various stages of clinical development

585 (Fig. 7). Gene targets were obtained for each therapy from the Therapeutic Target Database (TTD; release
586 2025-08-13) and mapped onto standardised HUGO Gene Nomenclature Committee (HGNC) gene symbols
587 using the `orthogene` R package. We stratified our overlap metrics according to whether the therapies had
588 failed (unsuccessful clinical trials or withdrawn), or were non-failed (successful or ongoing clinical trials).
589 We then conducted hypergeometric tests to determine whether the observed overlap between our prioritised
590 targets and the non-failed therapy targets was significantly greater than expected by chance (i.e. enrichment).
591 We also conducted a second hypergeometric test to determine whether the observed overlap between our
592 prioritised targets and the failed therapy targets was significantly less than expected by chance (i.e. depletion).
593 Finally, we repeated the analysis against all therapeutic targets, not just those of gene therapies, to determine
594 whether our prioritised targets had relevance to other therapeutic modalities.

595 Experimental model translatability

596 To improve the likelihood of successful translation between preclinical animal models and human patients,
597 we created an interspecies translatability prediction tool for each phenotype nominated by our gene therapy
598 prioritised pipeline (Supp. Fig. 17). First, we extracted ontological similarity scores of homologous pheno-
599 types across species from the MKG²⁵. Briefly, the ontological similarity scores (SIM_o) are computed for each
600 homologous pair of phenotypes across two ontologies by calculating the overlap in homologous phenotypes
601 that are ancestors or descendants of the target phenotype. Next, we generated genotypic similarity scores
602 (SIM_g) for each homologous phenotype pair by computing the proportion of 1:1 orthologous genes using
603 gene annotation from their respective ontologies. Interspecies orthologs were also obtained from the MKG.
604 Finally, both scores are multiplied together to yield a unified ontological-genotypic similarity score (SIM_{og}).

605 Novel R packages

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

614 Rare Disease Celltyping Portal

615 To further increase the ease of access for stakeholders in the RD community without the need for program-
616 matic experience, we developed a series of web apps to interactively explore, visualise, and download the
617 results from our study. Collectively, these web apps are called the Rare Disease Celltyping Portal. The

618 website can be accessed at <https://neurogenomics-ukdri.dsi.ic.ac.uk/>.

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

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

630 All code used to generate the website can be found at [https://github.com/neurogenomics/Rare-Disease-
Web-Portal](https://github.com/neurogenomics/Rare-Disease-
631 Web-Portal).

632 **Mappings**

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

640 **Data Availability**

641 All data is publicly available through the following resources:

- 642 • Human Phenotype Ontology (<https://hpo.jax.org>)
- 643 • GenCC (<https://thegencc.org/>)
- 644 • Descartes Human scRNA-seq atlas ([https://cellxgene.cziscience.com/collections/c114c20f-1ef4-49a5-
9c2e-d965787fb90c](https://cellxgene.cziscience.com/collections/c114c20f-1ef4-49a5-
645 9c2e-d965787fb90c))
- 646 • Human Cell Landscape scRNA-seq atlas ([https://cellxgene.cziscience.com/collections/38833785-fac5-
48fd-944a-0f62a4c23ed1](https://cellxgene.cziscience.com/collections/38833785-fac5-
647 48fd-944a-0f62a4c23ed1))
- 648 • Processed Cell Type Datasets (`ctd_DescartesHuman.rds` and `ctd_HumanCellLandscape.rds`; [https://github.com/neurogenomics/MSTExplorer/releases](https://
649 github.com/neurogenomics/MSTExplorer/releases))

- 650 • Gene x Phenotype association matrix (*hpo_matrix.rds*; <https://github.com/neurogenomics/MSTEExplorer/releases>)
651
652 • GPT-4 phenotype severity annotations (https://github.com/neurogenomics/rare_disease_celltyping/releases/download/latest/gpt_check_annot.csv.gz)
653
654 • Full phenotype-cell type association test results https://github.com/neurogenomics/MSTEExplorer/releases/download/v0.1.10/phenomix_results.tsv.gz
655
656 • Rare Disease Celltyping Portal (<https://neurogenomics-ukdri.dsi.ic.ac.uk/>)

657 **Code Availability**

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

- 659 • KGExplorer (<https://github.com/neurogenomics/KGExplorer>)
660 • HPOExplorer (<https://github.com/neurogenomics/HPOExplorer>)
661 • MSTEExplorer (<https://github.com/neurogenomics/MSTEExplorer>)
662 • Code to replicate analyses (https://github.com/neurogenomics/rare_disease_celltyping)
663 • Cell type-specific gene target prioritisation (https://neurogenomics.github.io/RareDiseasePrioritisation/reports/prioritise_targets)
664
665 • Complement system gene list (<https://www.genenames.org/data/genegroup/#!/group/492>)

666 **Acknowledgements**

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

670 **Funding**

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

674 **References**

- 675 1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A* **179**, 885–892 (2019).
- 676 2. Zhu, Q. *et al.* An integrative knowledge graph for rare diseases, derived from the genetic and rare
677 diseases information center (GARD). *J. Biomed. Semantics* **11**, 13 (2020).
- 678 3. Rare diseases BioResource.
- 679 4. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed
680 disease: Beyond the exome. *Genome Med.* **14**, 23 (2022).

- 679 5. Molster, C. *et al.* Survey of healthcare experiences of australian adults living with rare diseases. *Orphanet J. Rare Dis.* **11**, 30 (2016).
- 680 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).
- 681 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).
- 682 8. Yates, N. & Hinkel, J. The economics of moonshots: Value in rare disease drug development. *Clin. Transl. Sci.* **15**, 809–812 (2022).
- 683 9. Nuijten, M. Pricing zolgensma - the world's most expensive drug. *J Mark Access Health Policy* **10**, 2022353 (2022).
- 684 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).
- 685 11. Gargano, M. A. *et al.* The human phenotype ontology in 2024: Phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
- 686 12. Köhler, S. *et al.* Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
- 687 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).
- 688 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).
- 689 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).
- 690 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).
- 691 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).
- 692 18. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
- 693 19. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
- 694 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).

- 695 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).
- 696 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. doi:[10.1101/2024.06.10.24308475](https://doi.org/10.1101/2024.06.10.24308475).
- 697 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).
- 698 24. Diehl, A. D. *et al.* The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
- 699 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).
- 700 26. Heim, C. E. *et al.* Myeloid-derived suppressor cells contribute to staphylococcus aureus orthopedic biofilm infection. *J. Immunol.* **192**, 3778–3792 (2014).
- 701 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).
- 702 28. Stoll, H. *et al.* Staphylococcal enterotoxins Dose-Dependently modulate the generation of Myeloid-Derived suppressor cells. *Front. Cell. Infect. Microbiol.* **8**, 321 (2018).
- 703 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).
- 704 30. Zhou, Z., Xu, M.-J. & Gao, B. Hepatocytes: A key cell type for innate immunity. *Cell. Mol. Immunol.* **13**, 301–315 (2016).
- 705 31. Dixon, L. J., Barnes, M., Tang, H., Pritchard, M. T. & Nagy, L. E. Kupffer cells in the liver. *Compr. Physiol.* **3**, 785–797 (2013).
- 706 32. Ladhami, S. N. *et al.* Invasive meningococcal disease in patients with complement deficiencies: A case series (2008-2017). *BMC Infect. Dis.* **19**, 522 (2019).
- 707 33. Rosain, J. *et al.* Strains responsible for invasive meningococcal disease in patients with terminal complement pathway deficiencies. *J. Infect. Dis.* **215**, 1331–1338 (2017).
- 708 34. Seal, R. L. *et al.* Genenames.org: The HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).
- 709 35. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* **47**, 856–860 (2015).
- 710 36. Ochoa, D. *et al.* Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nature Reviews Drug Discovery* **21**, 551–551 (2022).

- 711 37. 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).
- 712 38. 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).
- 713 39. 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).
- 714 40. 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).
- 715 41. 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).
- 716 42. 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).
- 717 43. 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).
- 718 44. Ferro, A., Sheeler, C., Rosa, J.-G. & Cvitanovic, M. Role of microglia in ataxias. *Journal of molecular
biology* **431**, 1792–1804 (2019).
- 719 45. 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).
- 720 46. 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).
- 721 47. 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).
- 722 48. Bulaklak, K. & Gersbach, C. A. The once and future gene therapy. *Nat. Commun.* **11**, 5820 (2020).
- 723 49. Kohn, D. B., Chen, Y. Y. & Spencer, M. J. Successes and challenges in clinical gene therapy. *Gene
Ther.* **30**, 738–746 (2023).
- 724 50. Zhao, Z., Shang, P., Mohanraju, P. & Geijsen, N. Prime editing: Advances and therapeutic applica-
tions. *Trends Biotechnol.* **41**, 1000–1012 (2023).
- 725 51. Lu, C.-F. FDA takes first step toward international regulation of gene therapies to treat rare diseases.
(2024).
- 726 52. 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).
- 727 53. Zanello, G. *et al.* Targeting shared molecular etiologies to accelerate drug development for rare
diseases. *EMBO Mol. Med.* **15**, e17159 (2023).

- 728 54. Côté, R. *et al.* The ontology lookup service: Bigger and better. *Nucleic Acids Res.* **38**, W155–60
(2010).
- 729 55. DiStefano, M. T. *et al.* The gene curation coalition: A global effort to harmonize gene-disease evidence resources. *Genet. Med.* **24**, 1732–1742 (2022).
- 730 56. 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).
- 731 57. 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).
- 732 58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).
- 733 59. 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).
- 734 60. Lazarin, G. A. *et al.* Systematic classification of disease severity for evaluation of expanded carrier screening panels. *PLoS One* **9**, e114391 (2014).
- 735 61. Solovyeva, V. V. *et al.* New approaches to tay-sachs disease therapy. *Frontiers in Physiology* **9**, (2018).
- 736 62. 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).
- 737 63. Sugiyama, K., Tagawa, S. & Toda, M. Methods for visual understanding of hierarchical system structures. *IEEE Trans. Syst. Man Cybern.* **11**, 109–125 (1981).

738

739

740 **Supplementary Materials**

741 **Supplementary Results**

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

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

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

751 We visualised the putative causal relationships between genes, cell types and diseases associated with RNI as
752 a network (Fig. 14). The phenotype ‘Recurrent Neisserial infections’ was connected to cell types through the
753 aforementioned association test results ($\text{FDR} < 0.05$). Genes that were primarily driving these associations
754 (i.e. genes that were both strongly linked with ‘Recurrent Neisserial infections’ and were highly specifically
755 expressed in the given cell type) were designated as “driver genes” and retained for plotting. Across all
756 phenotypes in the HPO, more specific phenotypes (terms in the HPO with greater IC) are not only more
757 specific to certain cell types (Fig. 3b), but are also associated with genes that have greater cell type-specific
758 expression within those cell types. Even so, we should note that the choice of which specificity quantiles to
759 include is arbitrary. It should also be noted that simply because a gene is not specific to a cell type does not
760 mean it is not important for the function of the cell type. Indeed, there are many genes that are ubiquitously
761 expressed throughout many tissues in the body and are essential for cell function. Gene expression specificity
762 is nevertheless a useful metric to help distinguish many hundreds of cell (sub)types with overlapping gene
763 signatures.

764 **Selected example targets**

765 From our prioritised targets, we selected the following four sets of phenotypes or diseases as examples:
766 ‘GM2-ganglioside accumulation’, ‘Spinocerebellar atrophy’, ‘Neuronal loss in central nervous system’. Only
767 phenotypes with a GPT severity score greater than 15 were considered to avoid overplotting and to focus on
768 the more clinically relevant phenotypes Fig. 8a-h. These examples were then selected partly on the basis of
769 severity rankings, and partly for their relatively smaller, simpler networks than lent themselves to compact
770 visualisations.

771 Tay-Sachs disease (TSD) is a devastating hereditary condition in which children are born appearing healthy,
772 which gradually degrades leading to death after 3-5 years. The underlying cause is the toxic accumulation

773 of gangliosides in the nervous system due to a loss of the enzyme produced by *HEXA*. While this could
774 in theory be corrected with gene editing technologies, there remain some outstanding challenges. One of
775 which is identifying which cell types should be targeted to ensure the most effective treatments. Here
776 we identified alternatively activated macrophages as the cell type most strongly associated with ‘GM2-
777 ganglioside accumulation’ Fig. 8i. The role of aberrant macrophage activity in the regulation of ganglioside
778 levels is supported by observation that gangliosides accumulate within macrophages in TSD⁴⁰, as well as
779 experimental evidence in rodent models^{41,42,43}. Our results not only corroborate these findings, but propose
780 macrophages as the primary causal cell type in TSD, making it the most promising cell type to target in
781 therapies.

782 Another challenge in TSD is early detection and diagnosis, before irreversible damage has occurred. Our
783 pipeline implicated extravillous trophoblasts of the placenta in ‘GM2-ganglioside accumulation’. While not
784 necessarily a target for gene therapy (as the child is detached from the placenta after birth), checking these
785 cells *in utero* for an absence of *HEXA* may serve as a viable biomarker as these cells normally express
786 the gene at high levels. Early detection of TSD may lengthen the window of opportunity for therapeutic
787 intervention⁶¹, especially when genetic sequencing is not available or variants of unknown significance are
788 found within *HEXA*⁶².

789 Spinocerebellar atrophy is a debilitating and lethal phenotype that occurs in diseases such as Spinocerebellar
790 ataxia and Boucher-Nenhauser syndrome. These diseases are characterised by progressive degeneration of
791 the cerebellum and spinal cord, leading to severe motor and cognitive impairments. Our pipeline identified
792 M2 macrophages (labeled as the closest CL term ‘Alternatively activated macrophages’ in Fig. 8j) as the
793 only causal cell type associated with ‘Spinocerebellar atrophy’. This strongly suggests that degeneration of
794 cerebellar Purkinje cells are in fact downstream consequences of macrophage dysfunction, rather than being
795 the primary cause themselves. This is consistent with the known role of macrophages, especially microglia, in
796 neuroinflammation and other neurodegenerative conditions such as Alzheimer’s and Parkinsons’ disease^{44–46}.
797 While experimental and postmortem observational studies have implicated microglia in spinocerebellar atro-
798 phy previously⁴⁴, our results provide a statistically-supported and unbiased genetic link between known risk
799 genes and this cell type. Therefore, targeting M2 microglia in the treatment of spinocerebellar atrophy may
800 therefore represent a promising therapeutic strategy. This is aided by the fact that there are mouse models
801 that perturb the ortholog of human spinocerebellar atrophy risk genes (e.g. *Atxn1*, *Pnpla6*) and reliably
802 recapitulate the effects of this diseases at the cellular (e.g. loss of Purkinje cells), morphological (e.g. atrophy
803 of the cerebellum, spinal cord, and muscles), and functional (e.g. ataxia) levels.

804 Next, we investigated the phenotype ‘Neuronal loss in the central nervous system’. Despite the fact that this
805 is a fairly broad phenotype, we found that it was only significantly associated with 3 cell types (alternatively
806 activated macrophage, macrophage, epithelial cell), specifically M2 macrophages and sinusoidal endothelial
807 cells Fig. 8k.

808 Skeletal dysplasia is a heterogeneous group of over 450 disorders that affect the growth and development of
809 bone and cartilage. This phenotype can be lethal when deficient bone growth leads to the constriction of
810 vital organs such as the lungs. Even after surgical interventions, these complications continue to arise as the
811 child develops. Pharmacological interventions to treat this condition have largely been ineffective. While
812 there are various cell types involved in skeletal system development, our pipeline nominated chondrocytes
813 as the causal cell type underlying the lethal form of this condition (Fig. 19). Assuringly, we found that
814 the disease ‘Achondrogenesis Type 1B’ is caused by the genes *SLC26A2* and *COL2A1* via chondrocytes.
815 We also found that ‘Platyspondylic lethal skeletal dysplasia, Torrance type’. Thus, in cases where surgical
816 intervention is insufficient, targeting these genes within chondrocytes may prove a viable long-term solution
817 for children suffering from lethal skeletal dysplasia.

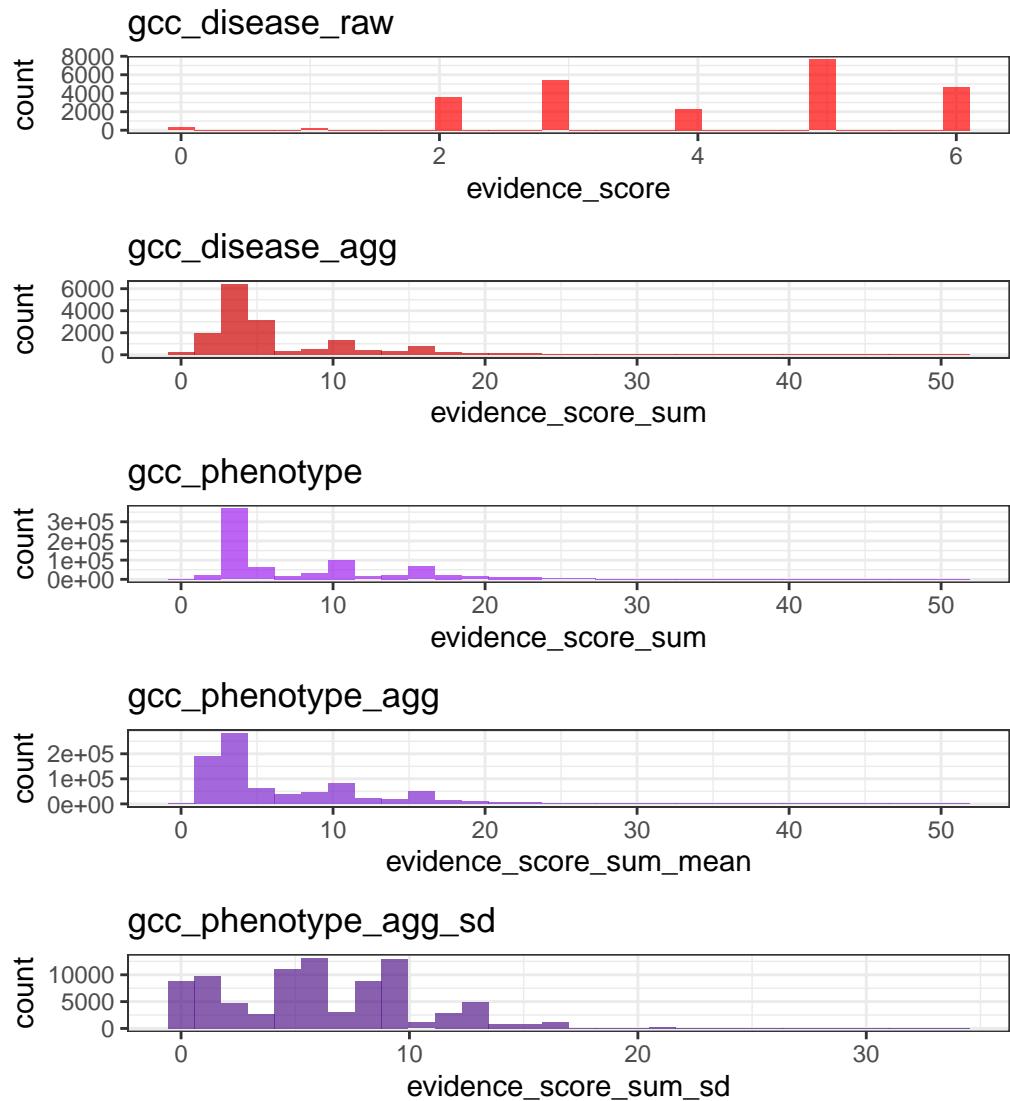
818 Alzheimer’s disease (AD) is the most common neurodegenerative condition. It is characterised by a set of
819 variably penetrant phenotypes including memory loss, cognitive decline, and cerebral proteinopathy. Inter-
820 estingly, we found that different forms of early onset AD (which are defined by the presence of a specific
821 disease gene) are each associated with different cell types via different phenotypes (Fig. 19). For example,
822 AD 3 and AD 4 are primarily associated with cells of the digestive system (‘enterocyte’, ‘gastric goblet
823 cell’) and are implied to be responsible for the phenotypes ‘Senile plaques’, ‘Alzheimer disease’, ‘Parietal
824 hypometabolism in FDG PET’. Meanwhile, AD 2 is primarily associated with immune cells (‘alternatively
825 activated macrophage’) and is implied to be responsible for the phenotypes ‘Neurofibrillary tangles’, ‘Long-
826 tract signs’. This suggests that different forms of AD may be driven by different cell types and phenotypes,
827 which may help to explain its variability in onset and clinical presentation.

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

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

843 **Experimental model translatability**

844 We computed interspecies translatability scores using a combination of both ontological (SIM_o) and geno-
845 typic (SIM_g) similarity relative to each homologous human phenotype and its associated genes Supp. Fig. 17.
846 In total, we mapped 1,221 non-human phenotypes (in *Caenorhabditis elegans*, *Danio rerio*, *Mus musculus*,
847 *Rattus norvegicus*) to 3,319 homologous human phenotypes. Amongst the 5,252 phenotype within our pri-
848 oritised therapy targets, 1,788 had viable animal models in at least one non-human species. Per species, the
849 number of homologous phenotypes was: *Mus musculus* (n=1705), *Danio rerio* (n=244), *Rattus norvegicus*
850 (n=85), *Caenorhabditis elegans* (n=23). Amongst our prioritised targets with a GPT-4 severity score of >10,
851 the phenotypes with the greatest animal model similarity were “Rudimentary to absent tibiae” ($SIM_{og} = 1$),
852 “Hypoglutaminemia” ($SIM_{og} = 1$), “Bilateral ulnar hypoplasia” ($SIM_{og} = 0.99$), “Disproportionate short-
853 ening of the tibia” ($SIM_{og} = 0.99$), “Acrobrachycephaly” ($SIM_{og} = 0.98$).



(a) **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 9

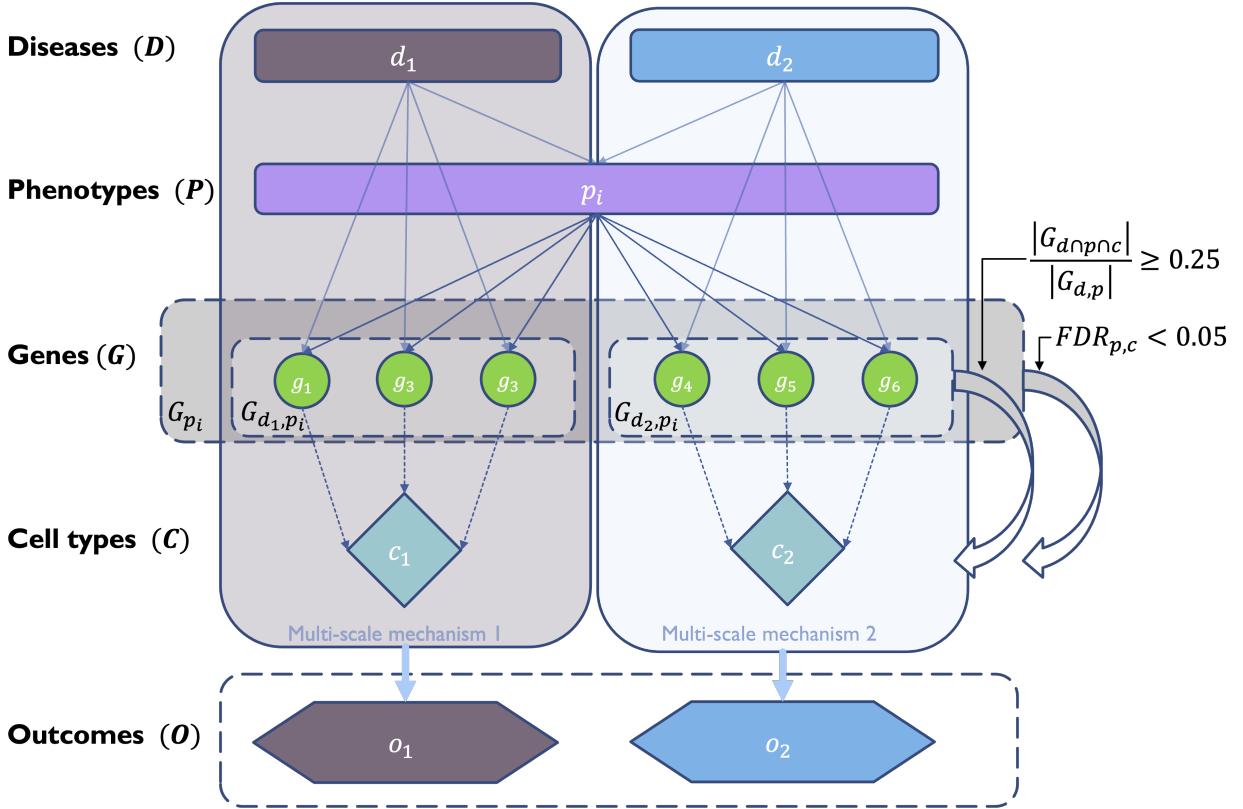
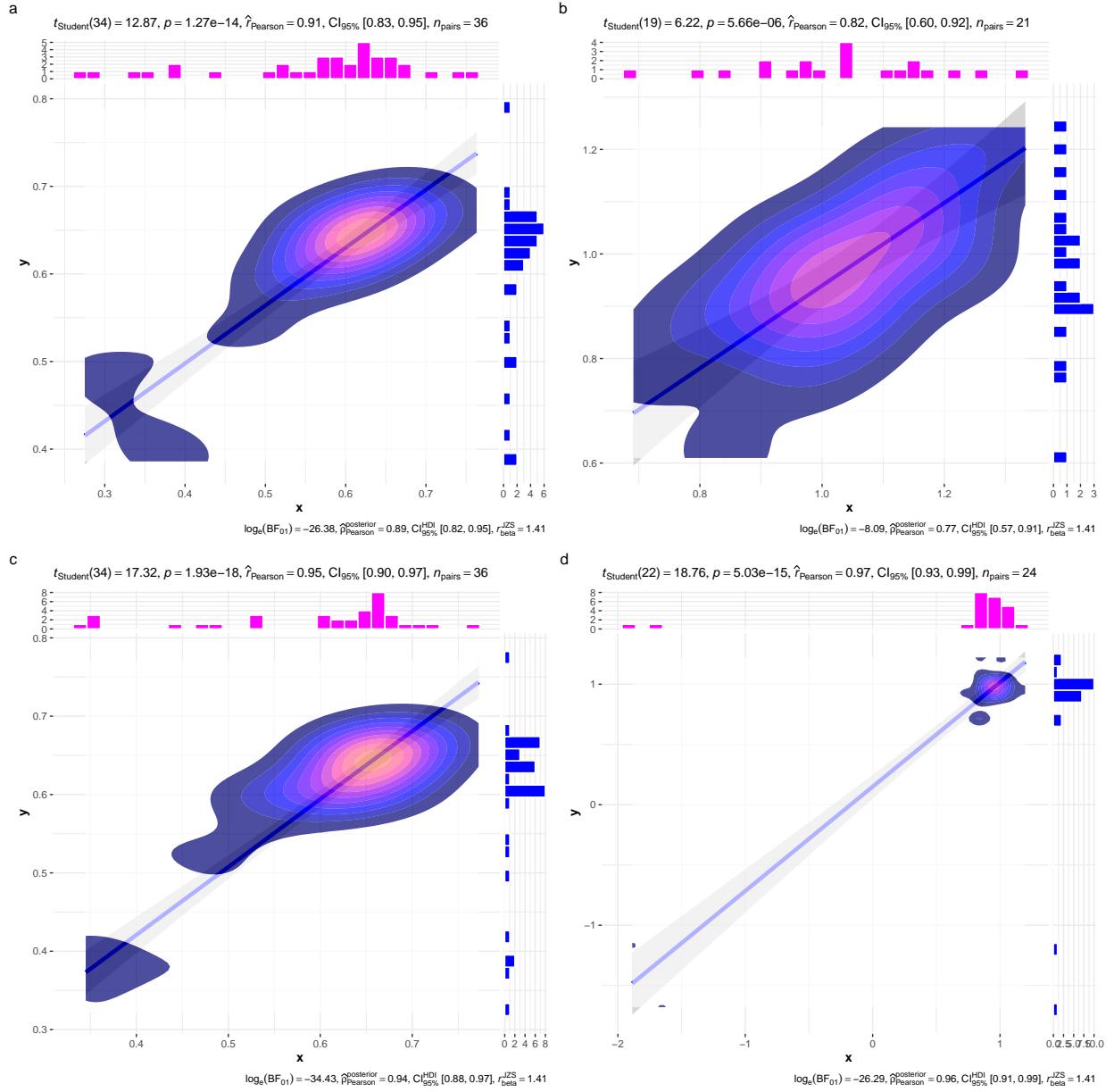
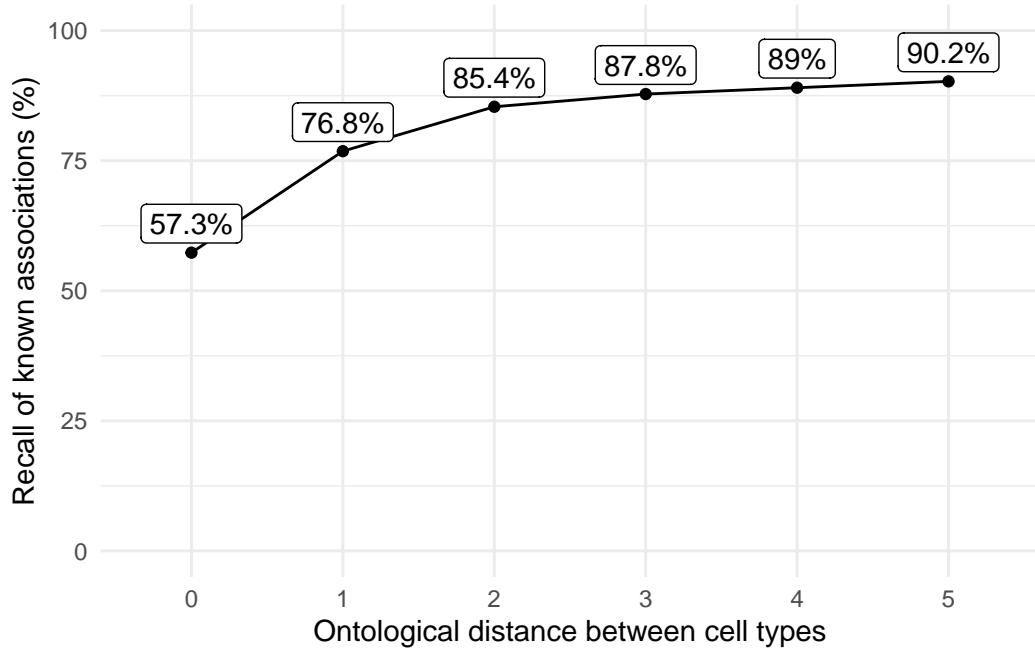


Figure 10: **Diagrammatic overview of multi-scale disease investigation strategy.** Here we provide an abstract example of differential disease aetiology across multiple scales: diseases (D), phenotypes (P), cell types (C), genes (G), and clinical outcomes (O). In the HPO, genes are assigned to phenotypes via particular diseases (G_{dp}). Therefore, the final gene list for each phenotype is aggregated from across multiple diseases (G_p). We performed association tests for all pairwise combinations of cell types and phenotypes and filtered results after multiple testing corrections ($FDR < 0.05$). Each phenotype in the context of a given disease is referred to here as a symptom. Links were established between symptoms and cell types through proportional gene set overlap at a minimum threshold of 25%.



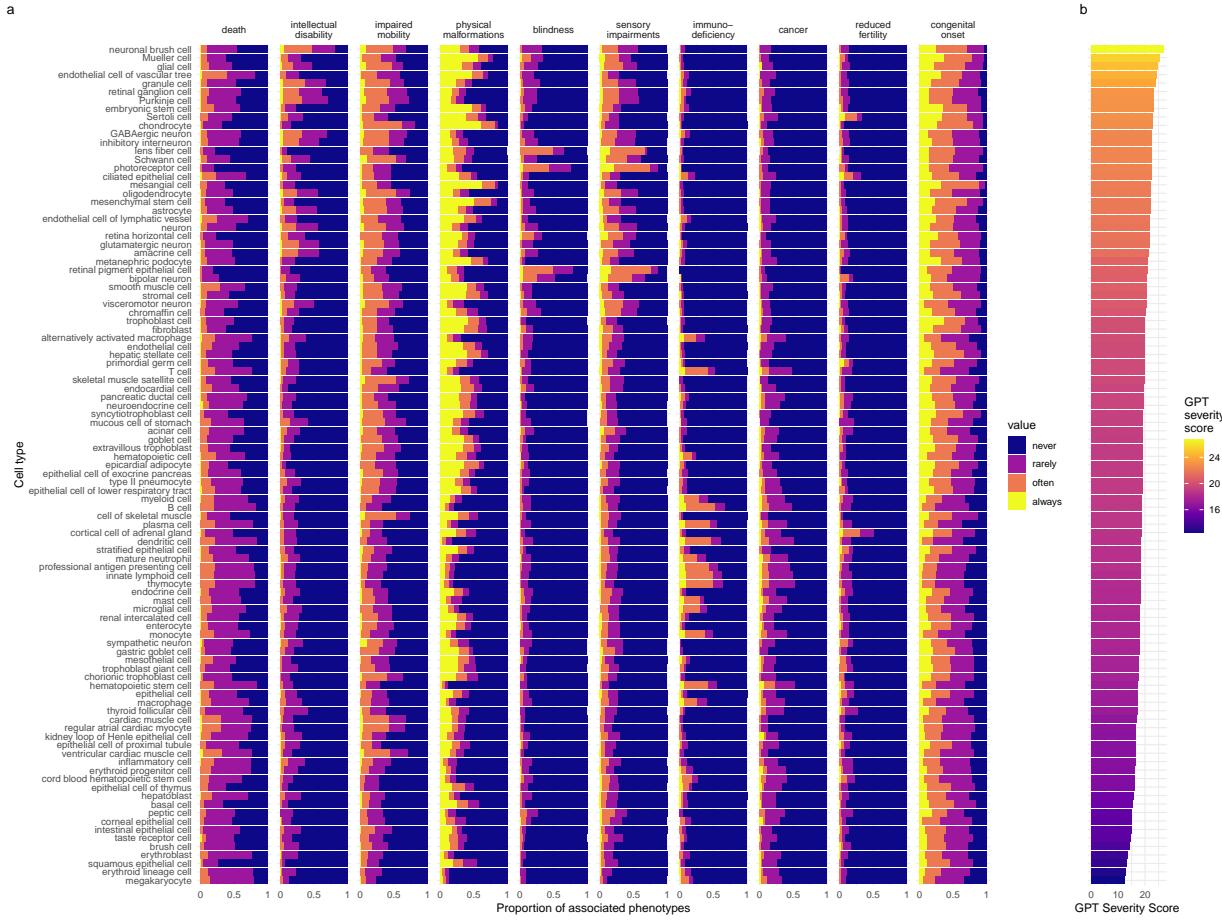
(a) Inter- and intra-dataset validation across the different CellTypeDataset (CTD) and developmental stages. Correlations are computed using Pearson correlation coefficient. Point density is plotted using a 2D kernel density estimate. **a** Correlation between the uncorrected p-values from all phenotype-cell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

Figure 11



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

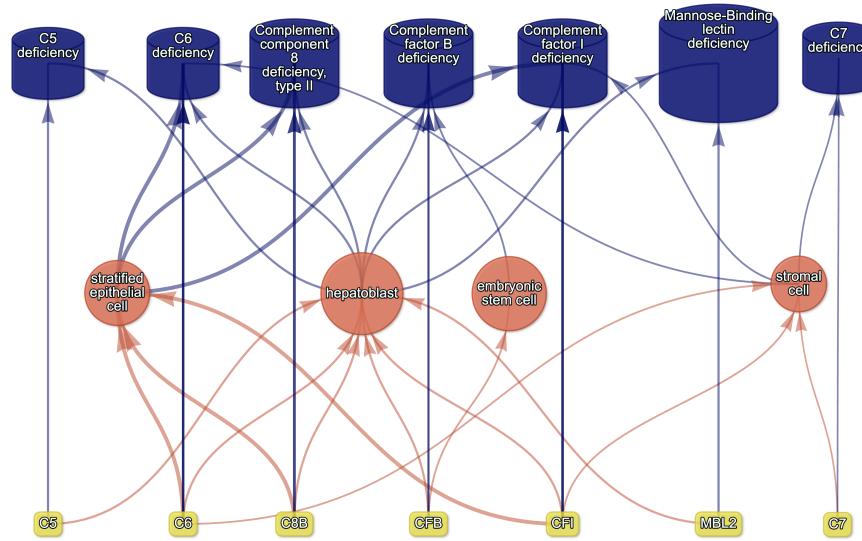
Figure 12



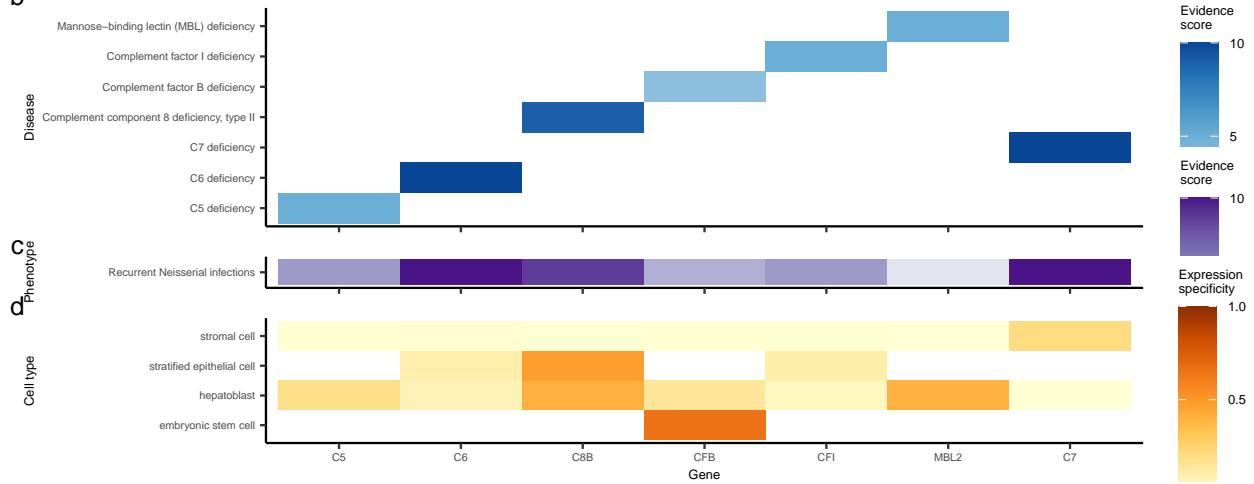
(a) **Cell types ordered by the mean severity of the phenotypes they're associated with.** **a**, The distribution of phenotype severity annotation frequencies aggregated by cell type. **b**, The composite severity score, averaged across all phenotypes associated with each cell type.

Figure 13

a

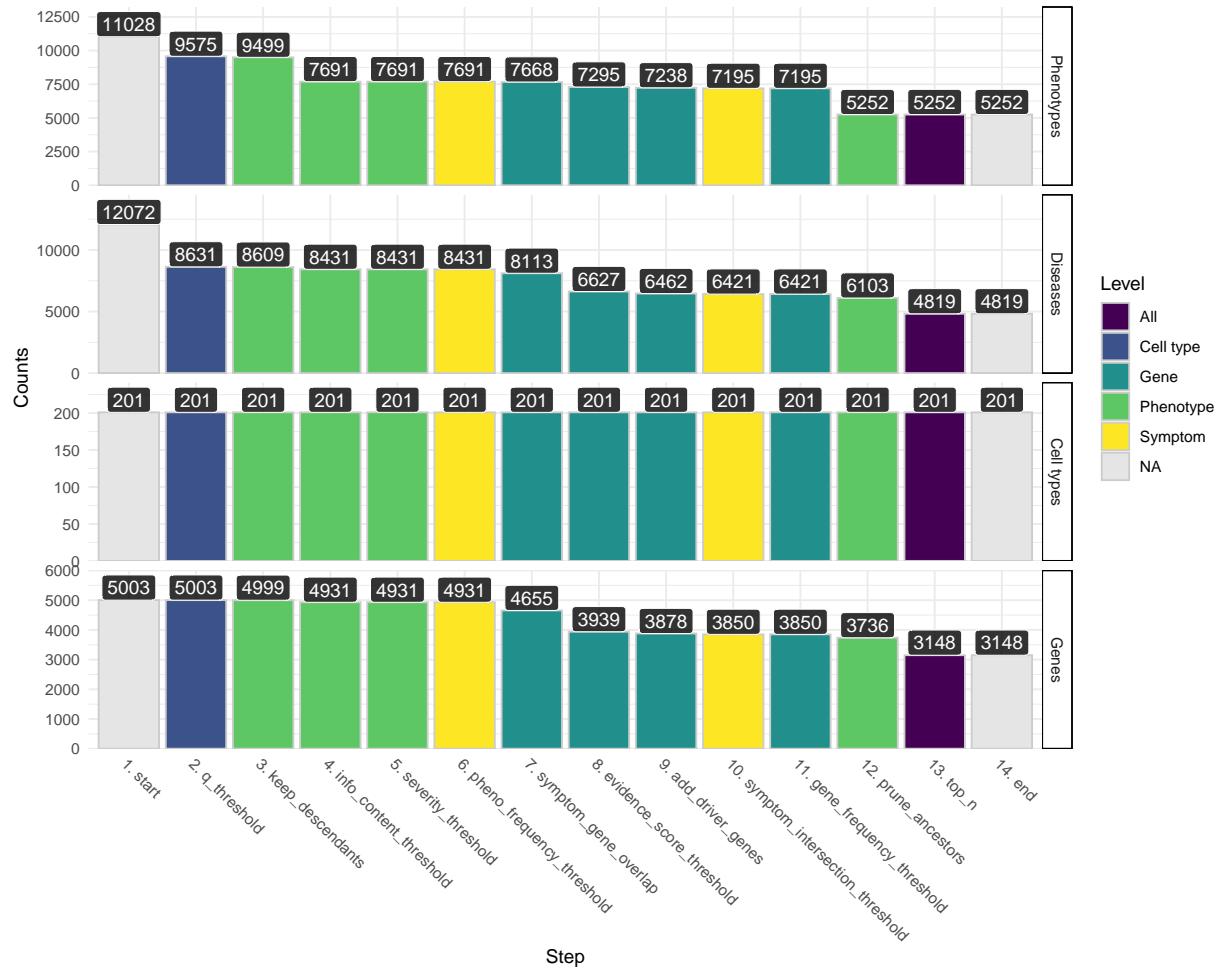


b



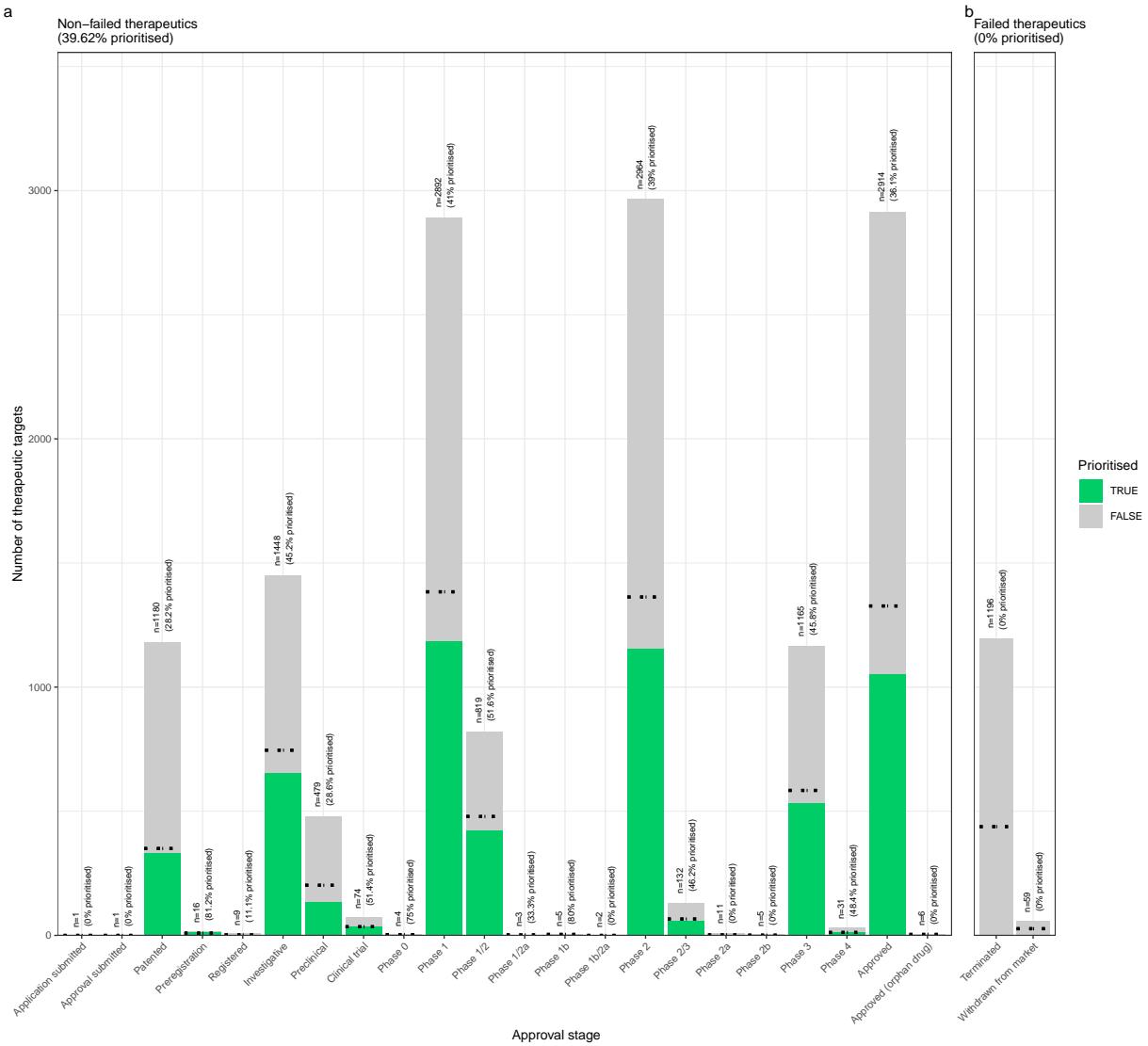
(a) **Causal network of recurrent Neisserial infections (RNI) reveals multi-scale disease aetiology.** RNI is a phenotype in seven different monogenic diseases caused by disruptions to specific complement system genes. Four cell types were significantly associated with RNI. **a**, One can trace how genes 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 algorithm⁶³. **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 14



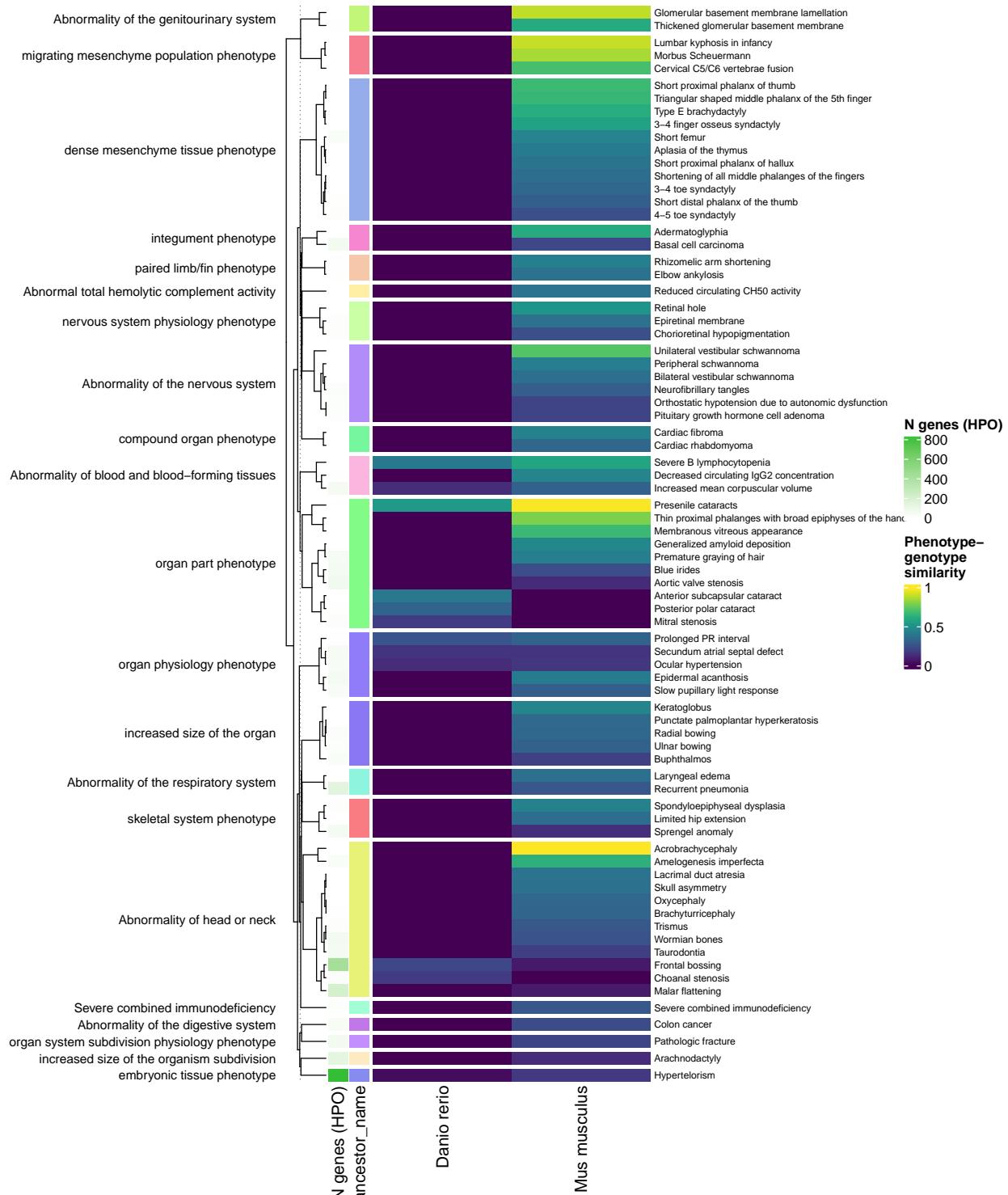
(a) **Prioritised target filtering steps.** This plot visualises the number of unique phenotype-cell type associations, cell types, genes, and phenotypes (*y-axis*) at each filtering step (*x-axis*) within the multi-scale therapeutic target prioritisation pipeline. Each step in the pipeline can be easily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

Figure 15



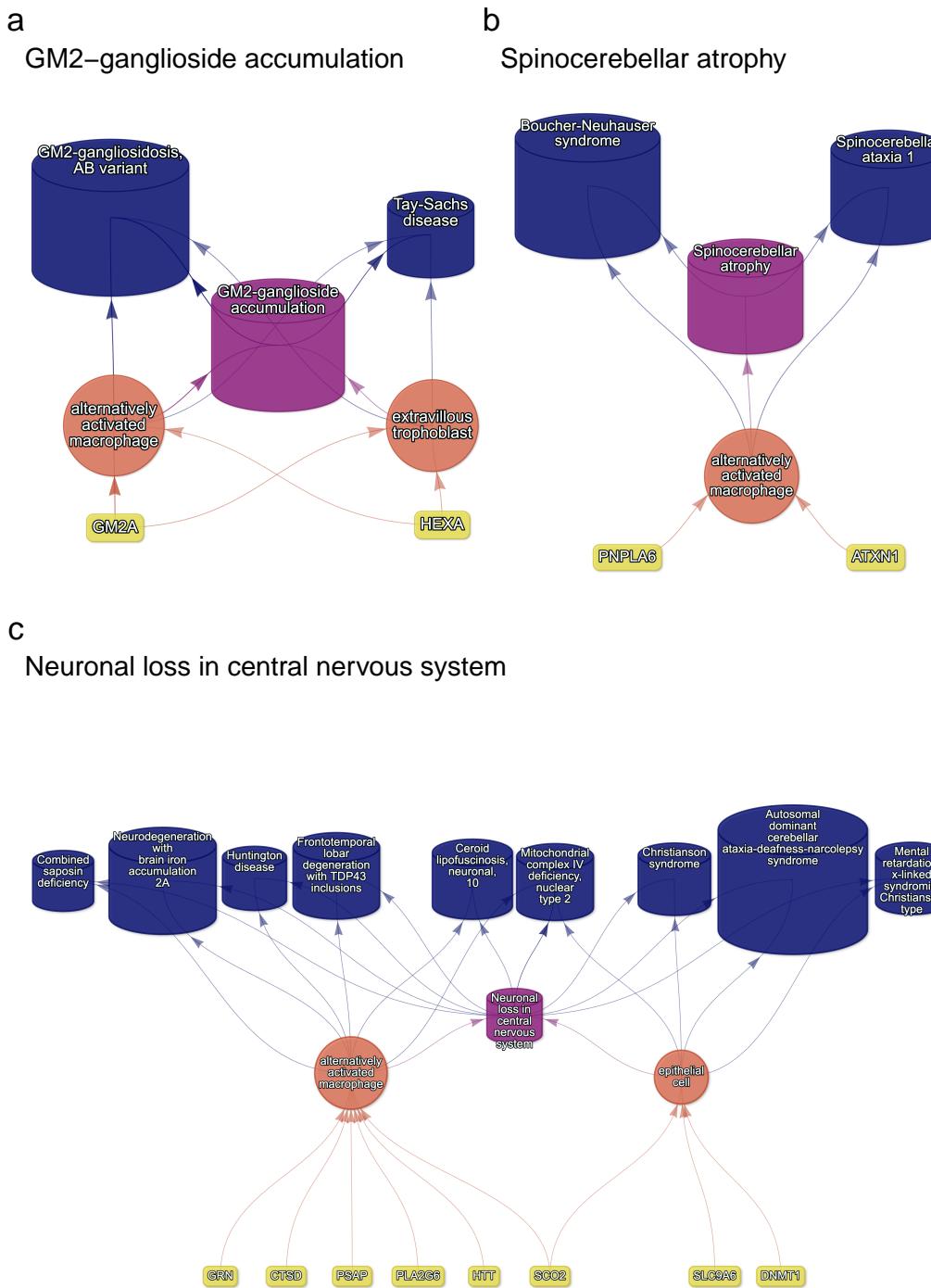
(a) **Validation of prioritised therapeutic targets.** Proportion of existing all therapy targets (documented in the Therapeutic Target Database) recapitulated by our prioritisation pipeline.

Figure 16



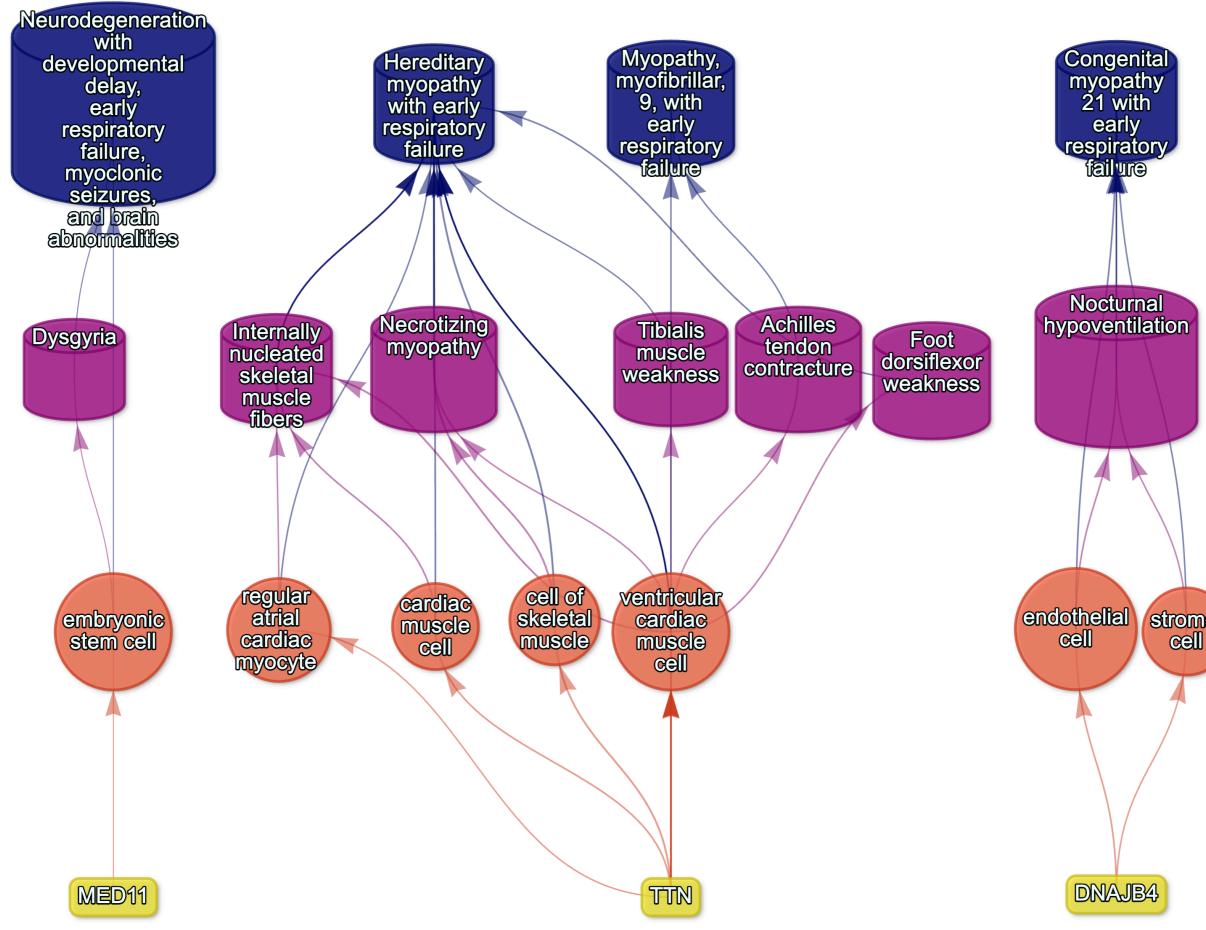
(a) **Identification of translatable experimental models.** Interspecies translatability of the top 200 human phenotypes nominated by the gene therapy prioritised pipeline. Above, the combined ontological-genotypic similarity score (SIM_{og}) is displayed as the heatmap fill colour stratified by the 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 17



(a) **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 HPO¹¹. 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 algorithm⁶³.

Figure 18



(a) Respiratory failure

Figure 19: **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 HPO¹¹. 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 algorithm⁶³.

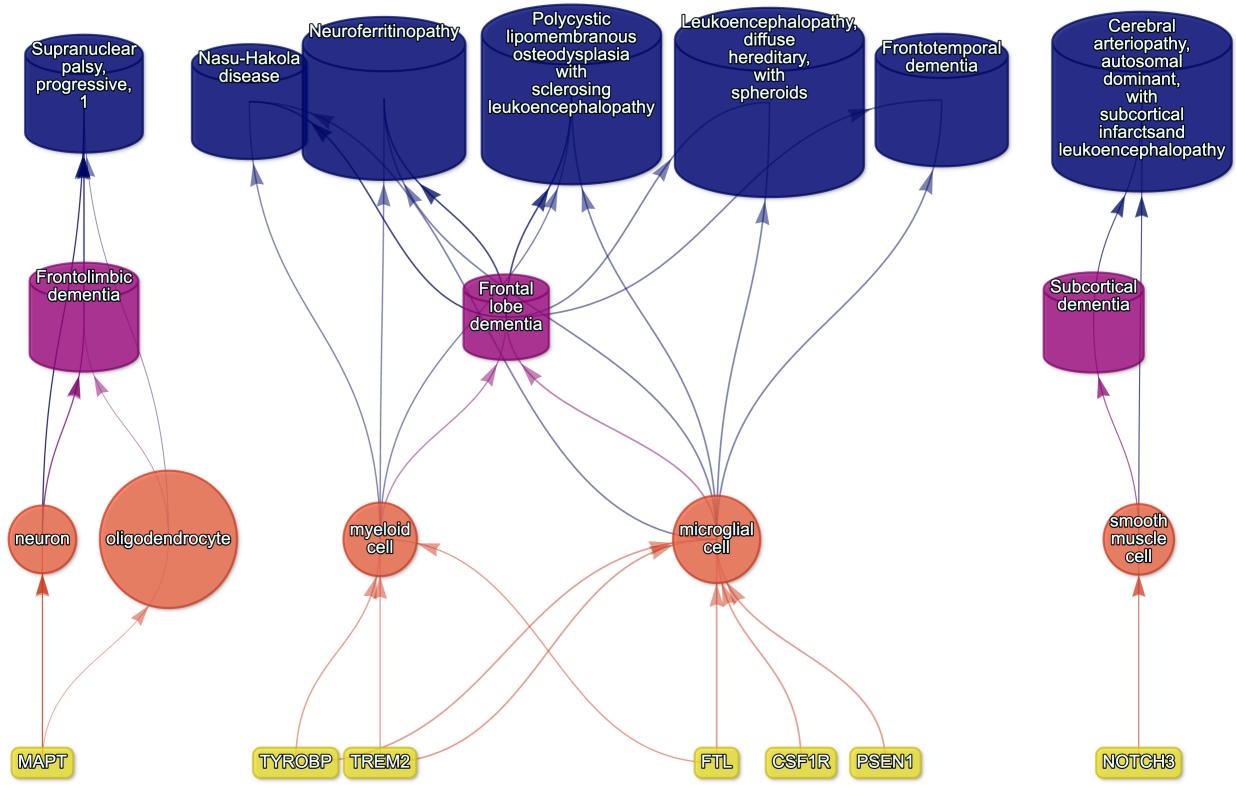


Figure 20: Causal multi-scale network for dementia phenotypes.

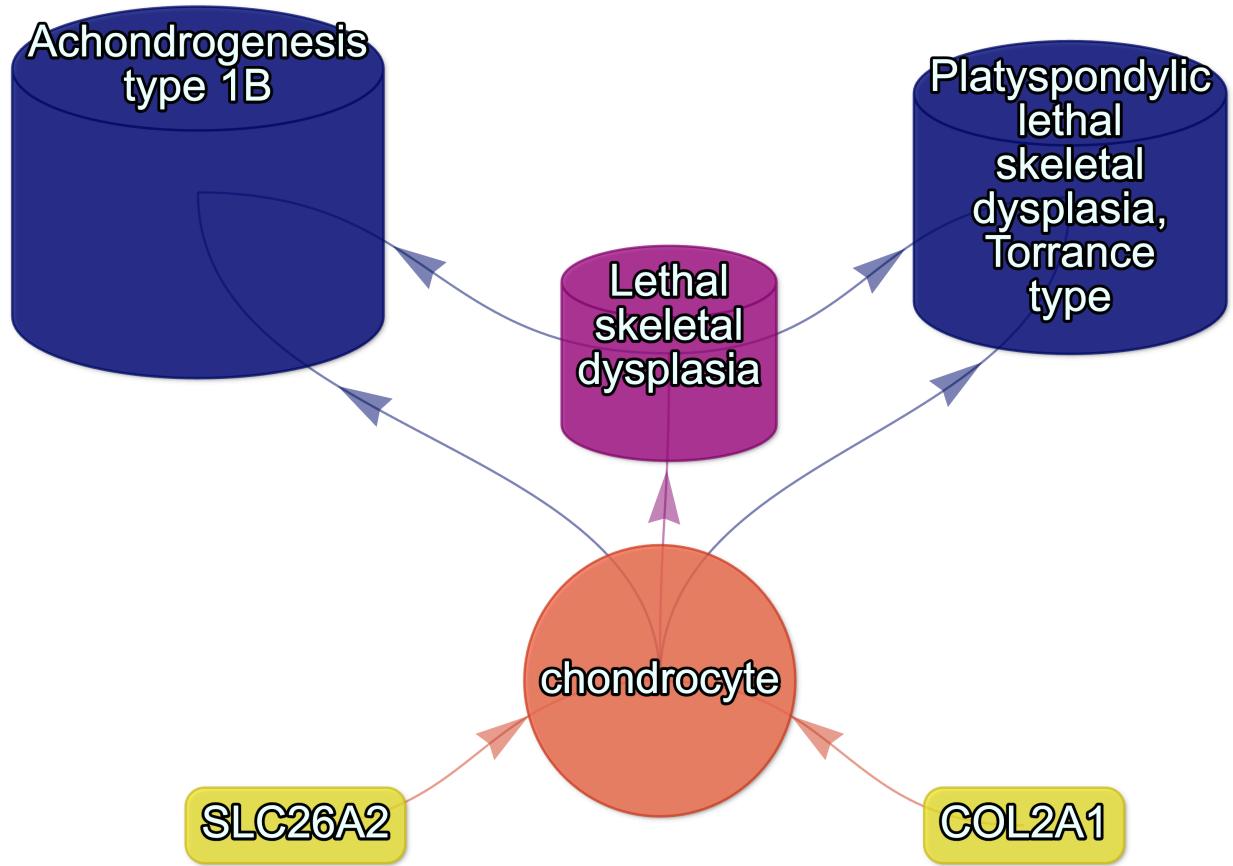


Figure 21: Causal multi-scale network for the phenotype lethal skeletal dysplasia.

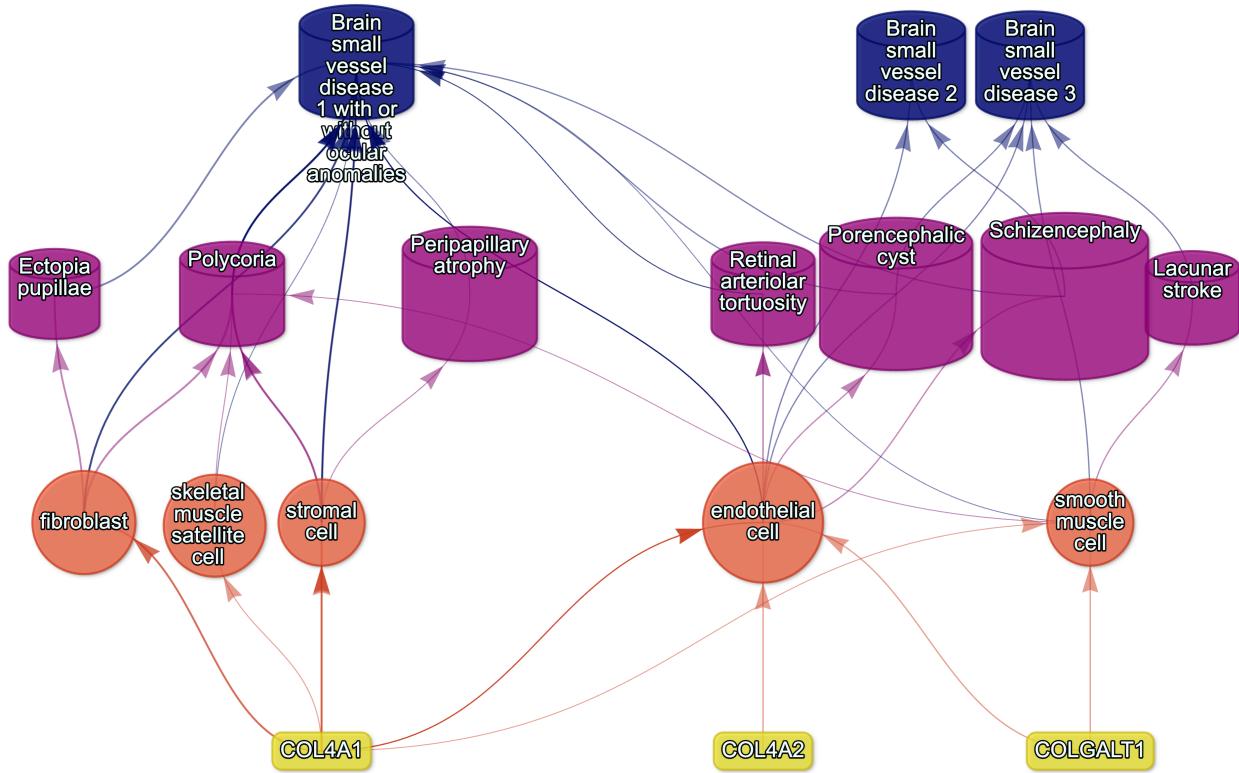


Figure 22: Causal multi-scale network for phenotypes associated with small vessel disease.

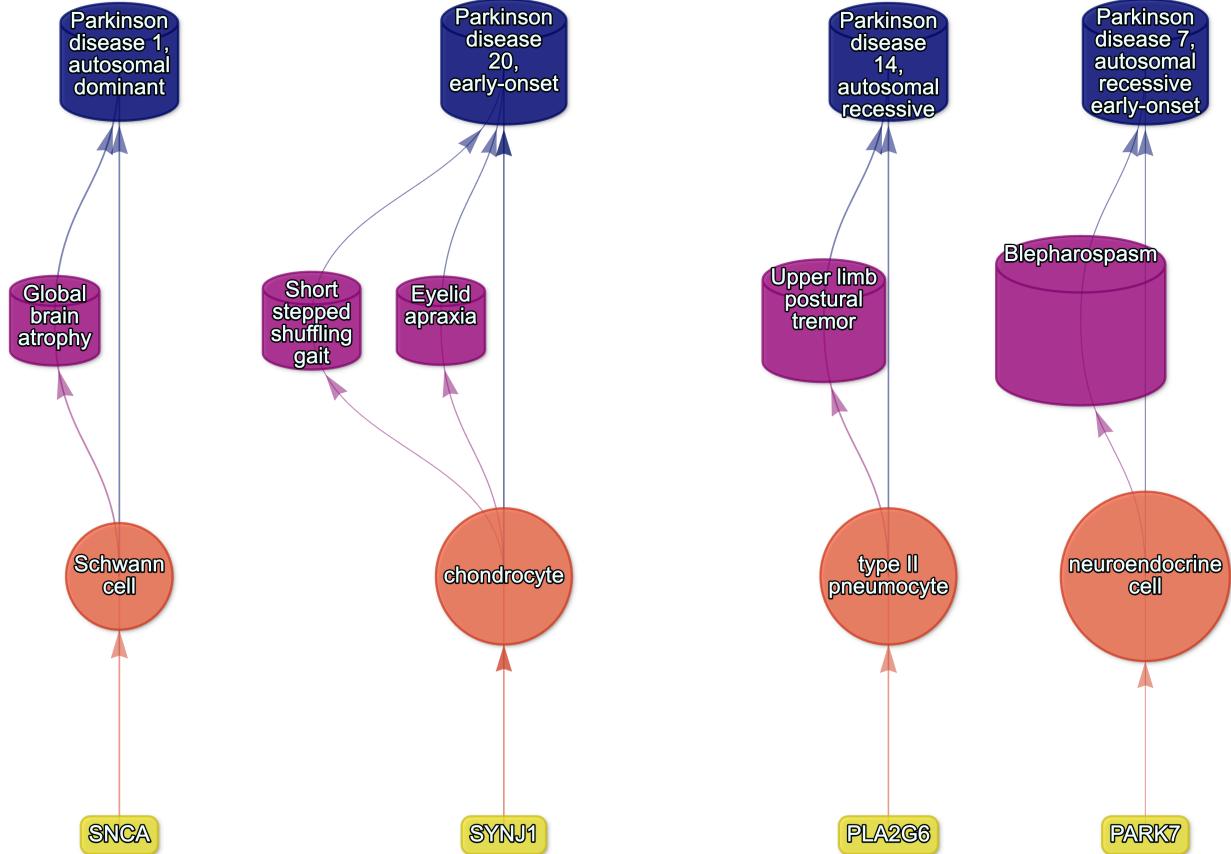


Figure 23: Causal multi-scale network for phenotypes associated with various subtypes of Parkinson's disease.

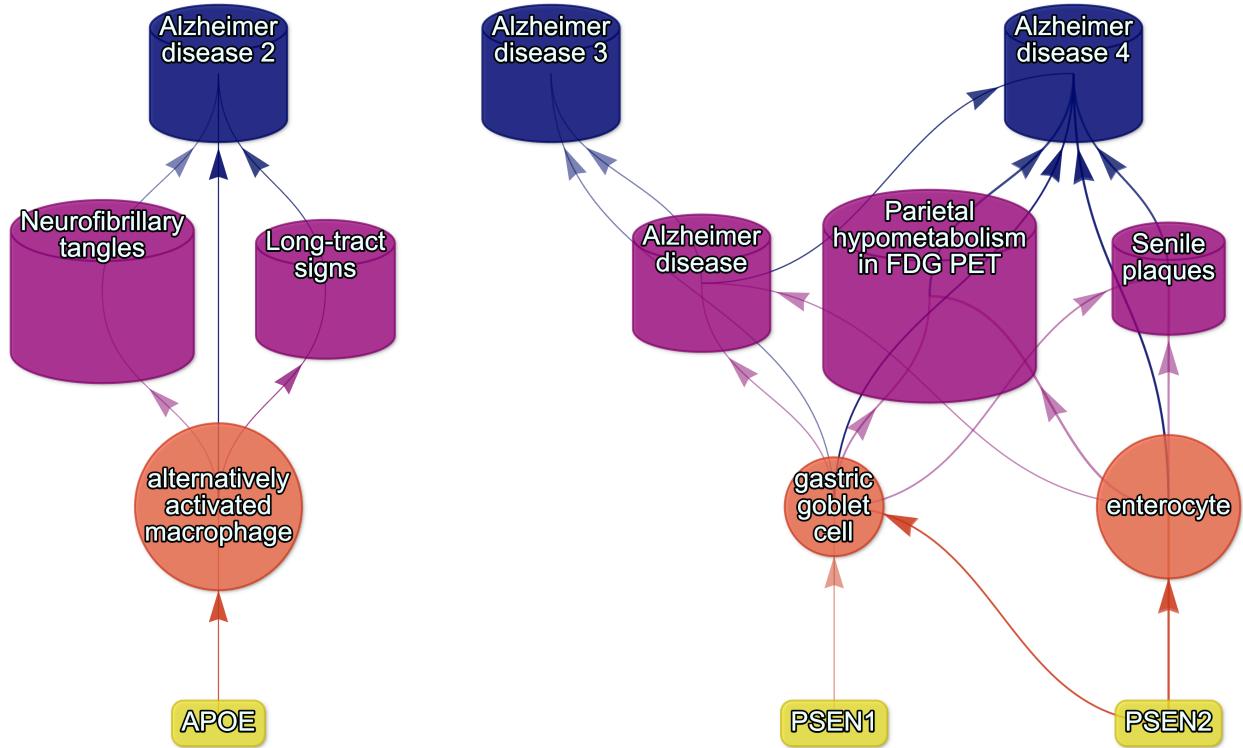


Figure 24: Causal multi-scale network for phenotypes associated with various subtypes of Alzheimer's disease.

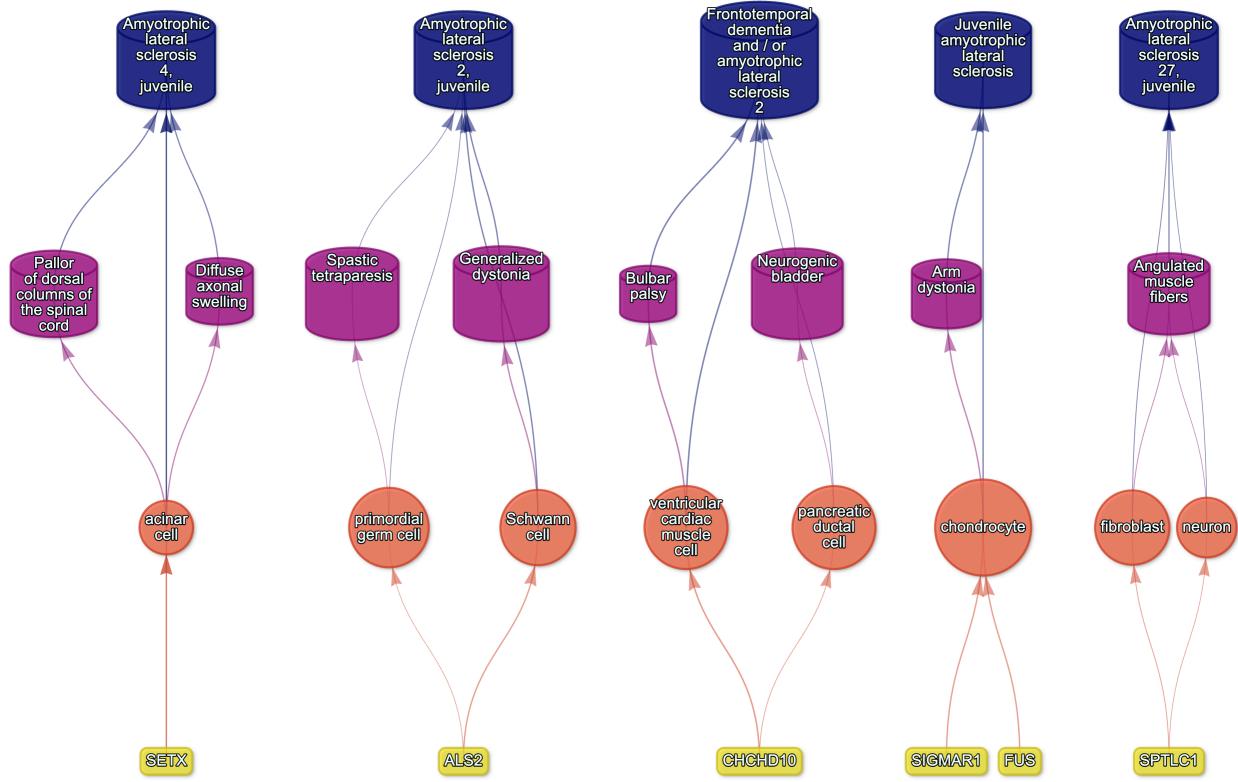


Figure 25: Causal multi-scale network for phenotypes associated with Amyotrophic Lateral Sclerosis (ALS).

855 **Supplementary Tables**

Table 1: **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 3: **Description of each filtering step performed in the multi-scale therapeutic target prioritisation pipeline.** ‘level’ indicates the biological scale at which the step is applied to.

level	step	description
NA	1. start	NA
Cell type	2. q threshold	Keep only cell type-phenotype association results at $q \leq 0.05$.
Phenotype	3. keep descendants	Remove phenotypes belonging to a certain branch of the HPO, as defined by an ancestor term.
Phenotype	4. info content threshold	Keep only phenotypes with a minimum information criterion score (computed from the HPO).
Phenotype	5. severity threshold	Keep only phenotypes with mean Severity equal to or below the threshold.
Symptom	6. pheno frequency threshold	Keep only phenotypes with mean frequency equal to or above the threshold (i.e. how frequently a phenotype is associated with any diseases in which it occurs).
Gene	7. symptom gene overlap	Ensure that genes nominated at the phenotype-level also appear in the genes overlapping at the cell type-specific symptom-level.
Gene	8. evidence score threshold	Remove genes that are below an aggregate phenotype-gene evidence score threshold.

Table 3: Description of each filtering step performed in the multi-scale therapeutic target prioritisation pipeline. ‘level’ indicates the biological scale at which the step is applied to.

level	step	description
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 2: **Summary statistics of enrichment results stratified by single-cell atlas.** Summary statistics at multiple levels (tests, cell types, phenotypes, diseases, cell types per phenotype, phenotypes per cell type) stratified by the single-cell atlas that was used as a cell type signature reference (Descartes Human or Human Cell 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 4: **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 6.

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 5: **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 6: **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
Abnormality of the cardiovascular system		
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
Abnormality of the endocrine system		
endocrine cell	endocrine cell	CL:0000163
endocrine cell	neuroendocrine cell	CL:0000165
endocrine cell	chromaffin cell	CL:0000166
Abnormality of the eye		
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
Abnormality of the immune system		
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
Abnormality of the musculoskeletal system		
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
Abnormality of the nervous system		
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
Abnormality of the respiratory system		
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 7: 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
Foetal-biased				
HP:0005105	Abnormal nasal morphology	0.00	4.5	6
HP:0010938	Abnormal external nose morphology	0.00	5.4	7
HP:0000366	Abnormality of the nose	0.00	3.8	5
HP:0000055	Abnormal female external genitalia morphology	0.00	5.2	6
HP:0000271	Abnormality of the face	0.00	1.9	4
HP:0000234	Abnormality of the head	0.00	1.7	3
HP:0000152	Abnormality of head or neck	0.00	1.6	2
HP:0010460	Abnormality of the female genitalia	0.03	2.8	5
HP:0000811	Abnormal external genitalia	0.03	2.8	5
HP:0000078	Abnormality of the genital system	0.03	1.9	3
Adult-biased				
HP:0010647	Abnormal elasticity of skin	0.00	6.0	5
HP:0008067	Abnormally lax or hyperextensible skin	0.00	6.0	6
HP:0011121	Abnormal skin morphology	0.00	2.4	4
HP:0000951	Abnormality of the skin	0.00	2.1	3
HP:0001574	Abnormality of the integument	0.01	1.6	2
HP:0001626	Abnormality of the cardiovascular system	0.02	1.4	2
HP:0030680	Abnormal cardiovascular system morphology	0.02	1.7	3
HP:0025015	Abnormal vascular morphology	0.04	1.9	4
HP:0030962	Abnormal morphology of the great vessels	0.04	2.7	6

Table 8: **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
Foetal-biased				
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
Adult-biased				
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