Identification of cell type-specific gene targets underlying thousands of rare diseases and subtraits

Kitty B. Murphy[1,2*], Robert Gordon-Smith[1,2*], Jai Chapman[1,2], Momoko Otani [3], Brian M. Schilder[1,2#], Nathan G. Skene[1,2#]

[1] UK Dementia Research Institute, Imperial College London, London, UK

[2] Department of Brain Sciences, Imperial College London, London, UK

[3] National Heart and Lung Institute, Imperial College London, UK

[#] Co-corresponding authors: brian_schilder@alumni.brown.edu ; n.skene@imperial.ac.uk

* Co-first authors

Keywords

Rare diseases, genetics, cell types, genomic medicine, AAV, therapeutics, computational biology, bioinformatics, systems biology

Abstract

Rare diseases (RDs) are uncommon as individual diagnoses, but as a group contribute to an enormous disease burden globally. However, partly due the low prevalence and high diversity of individual RDs, this category of diseases is understudied and under-resourced. The advent of large, standardised genetics databases has enabled high-throughput, comprehensive approaches that uncover new insights into the multi-scale aetiology of thousands of diseases. Here, using the Human Phenotype Ontology (9,677 annotated phenotypes) and multiple single-cell transcriptomic atlases (77 human cell types and 38 mouse cell types), we conducted >688,000 enrichment tests (x100,000 bootstrap iterations each) to identify >13,888 genetically supported cell type-phenotype associations. Our results recapitulate well-known cell type-phenotype relationships, and extend our understanding of these diseases by pinpointing the genes linking phenotypes to specific cell (sub)types. We also reveal novel cell type-phenotype relationships across disparate branches of clinical disease (e.g. the nervous, cardiovascular, and immune systems). Next, we introduce a computational pipeline to prioritise gene targets with high cell type-specificity to minimise off-target effects and maximise therapeutic potential. To broaden the impact of our study, we have released two R packages to fully replicate our analyses, as well as a series of interactive web apps so that stakeholders from a variety of backgrounds may further explore and utilise our findings. Together, we present a promising avenue for systematically and robustly uncovering the multi-scale aetiology of RDs at scale.

Introduction

In aggregate, there are over 10,000 recognised rare diseases (RDs) that affect at least 400 million people globally [1], contributing to an enormous disease burden (1 in 10-20 people) [2]. As over 70% of RDs have a known genetic component [3], the increasing availability of phenotypic and genetic datasets presents an opportunity to apply a systematic approach to investigate many RDs at once. One of the most comprehensive resources for RD research is the Human Phenotype Ontology (HPO), which currently contains over 15,200 human phenotypic abnormalities and subtraits, each assigned to unique identifiers and mapped to hierarchically related terms both within the HPO and across other ontologies [4–6]. Since 2008, the HPO has been continuously updated using knowledge from the medical literature, as well as by

integrating databases of expert validated gene-phenotype relationships, such as OMIM [7–9], Orphanet, and DECIPHER [10]. Currently, the HPO contains gene annotations for over 9,600 phenotypes. In tandem, the single-cell omics technologies have led to an explosion of cell type signature atlases [11].

Here, we strategically combine and extend these resources to identify cell type-specific therapeutic gene targets for thousands of RD with the goal of greatly improving efficacy and limiting side-effects. We provide a fully reproducible framework to systematically and robustly identify the primary cell types associated with RDs and RD-associated phenotypes. Specifically, we used Expression Weighted Cell type Enrichment (*EWCE*) [12] to perform a series of bootstrapped cell types enrichment tests between the gene lists of 6,173 HPO phenotypes (after quality control filtering) and each of the 77 cell types derived from a single-cell RNA seq (scRNA-seq) reference atlas of 15 organs across the developing human body [13]. From this, we identified 8,379 significant cell type-phenotype associations across 2,832 unique phenotypes. Each of the 77 cell types were enriched for at least one phenotype.

We then leveraged our cell type-phenotype results to systematically propose cell type- and gene-specific candidates to target in gene therapies. Specifically, we focused on candidates suitable for recombinant adeno-associated virus (rAAV) vectors to transduce single-stranded DNA into target cells [14–18], which have recently shown success in treating several rare diseases such as Leber's congenital amaurosis (LCA) [19,20], spinal muscular atrophy (SMA) [21], and amyotrophic lateral sclerosis (ALS) [22]. Together, in this study we uncover the molecular aetiology of thousands of phenotypes and diseases at once, and provide a systematic, evidence-based framework for identifying novel therapeutic targets in each of these diseases.

Additionally, we created two accompanying open-source R packages (*HPOExplorer* and *MultiEWCE*) to navigate the HPO data, search and postprocess the enrichment results from this study, and facilitate novel analyses of multiple gene lists in parallel. Finally, to make our data easily accessible we developed a series of interactive web app that allows exploration of all our results: (https://neurogenomics.github.io/rare_disease_celltyping_apps/home/) We hope that these tools will ensure reproducibility and facilitate future analyses as more phenotypic,

genotypic, and transcriptomic data becomes available.

Results

Summary

Within the results using the Descartes cell type signature reference, 8,379 / 475,321 (1.76%) of bootstrapped enrichment tests across 77 cell types and 6,173 phenotypes revealed significant cell type-phenotype associations after multiple-testing correction. Within these results, 2,832 phenotypes were significantly enriched for at least one cell type after multiple testing correction ($q \leq 0.05$) and 5,989 were nominally significant ($p \leq 0.05$). Hereafter we will only refer to the results that were significant after multiple testing correction. The number of enriched cell types per phenotype suggest reasonable specificity of the enrichment strategy (percentages are shown relative to all 77 cell types): min=1 (1.30%), median=2 (2.60%), mean=2.959 (3.84%), max=27 (35.07%). This was also true for the number of enriched phenotypes per cell type (percentages are shown relative to all 6,173 phenotype gene lists): min=1 (0.016%), median=90 (1.46%), mean=108.8 (1.76%), max=338 (5.48%).

With the Tabula Muris enrichment results, 5,509 / 213,028 (2.58%) tests were significant after multiple-testing correction ($q \leq 0.05$), and 20,221 tests were nominally significant ($p \leq 0.05$). All 38 cell types were enriched in at least one phenotype, and 2,579 / 5,509 (46.81%) phenotypes were enriched in at least one cell type. The proportion of enriched cell types per phenotype were comparable to those observed using the Descartes dataset (percentages are shown relative to all 38 cell types): min=1 (2.632%), median=1 (2.632%), mean=2.136 (5.621%), max=17 (44.737%). The number of enriched phenotypes per cell type were slightly greater than that the results using Descartes (percentages are shown relative to all 6,173 phenotype gene lists): min=2 (0.032%), median=119 (1.93%), mean=145 (2.35%), max=506 (8.20 %).

Enrichment tests highlight expected as well as novel cell type-phenotype associations

We first sought to confirm that our enrichment analyses were able to recover well-established cell type-phenotype relationships. We expected the terms “Abnormality of the nervous system”, “Abnormality of the cardiovascular system”, and “Abnormality of the immune system” to be strongly associated with neural cells, heart cells, and immune cells, respectively. Indeed, a hypergeometric test showed that terms related to (i) the “Abnormality of the nervous system” showed an overrepresentation in all nervous system related cell types (21 cell types), with the strongest enrichment in limbic system neurons (*n* enrichments = 194, $p = 5.74^{-44}$; **Fig. 1C**); (ii) the “Abnormality of the cardiovascular system” were overrepresented in 3/4 cardiovascular related cell types with cardiomyocytes being the most enriched (*n* enrichments = 94, $p = 1.23^{-53}$; **Fig. 1C**); (iii) the “Abnormality of the immune system” were overrepresented in 9/10 immune cell types with lymphoid cells being the most enriched (*n* enrichments = 111, $p = 5.23^{-56}$; **Fig. 1C**). Additionally, cell types that hierarchically clustered together (based on transcriptomic similarity) were also significantly associated with a particular term. For instance, the cluster of nervous system related cell types were enriched for terms related to the abnormality of the nervous system (*n* enrichments = 1768, $p < 2.23^{-308}$; **Fig. 1C**).

Somewhat unexpectedly, a significant number of phenotypes related to the “Abnormality of the cardiovascular system” were associated with hepatoblasts (*n* enrichments = 17, p = 0.027; **Fig. 1C**). On closer inspection, these phenotypes were associated with damage to arteries caused by lipid deposition, such as cerebral artery atherosclerosis and myocardial steatosis. Given the prominent role that the liver plays in lipid metabolism [23], it is therefore logical that dysfunction of hepatoblasts would be implicated in abnormal cardiovascular phenotypes.

To further demonstrate that our approach finds expected cell type-phenotype associations, we extracted all the HPO terms enriched within excitatory neurons, cardiomyocytes, and antigen presenting cells, and show that the more significantly associated terms within these cell types were disproportionately related to the expected parent term (**Fig. 1D-F**). Taking excitatory neurons as an example, the more significant the association between the phenotype and

excitatory neuron, the more likely this association was related to the abnormality of the nervous system (r = 0.82, p = 0.03; **Fig. 1d**).

Specific phenotypes are associated with fewer cell types and genes, but higher cell type-specificity of gene expression

We reasoned that lower ontology levels representing more specific phenotypes were likely to be associated with fewer cell types. In contrast, phenotypes with higher ontology levels would tend to be more broad and enriched for a wider variety of cell types. We confirmed that this is the case by counting the number of associated cell types with phenotypes in each ontology level, observing a strong positive correlation between ontology level and the number of associated cell types (Spearman's rank correlation coefficient, r = 0.33, $p < 2.2 \times 10^{16}$; **Fig. 2A**). In addition, lower ontology levels were associated with fewer genes (Spearman's rank correlation coefficient, r = 0.55, $p < 2.2 \times 10^{16}$; **Fig. 2C**) but the cell type-specificity of expression of the associated genes increased (Spearman's rank correlation coefficient, r = -0.65, $p < 2.2 \times 10^{16}$; **Fig. 2B**).

Just as observed for the broader phenotypes (e.g. "abnormality of the immune system"), we expected more specific phenotypes, such as recurrent infections, to also be associated with their expected cell types. Extracting all children terms of recurrent infections, which includes 72 HPO terms at ontology levels ranging from 0 to 3 (relative to each other), we predicted that these would be primarily enriched within immune system-related cell types. As predicted, significant enrichments were found in immune related cell types, but also in less anticipated cell types (**Fig. 3**). "Recurrent staphylococcal infections" were enriched within myeloid cells (p = 0.0098; **Fig. 3B**), an association that has been previously documented in the literature [24–27], whereas "Neisserial infections" highlighted a novel association with hepatoblasts (p = 0.013; **Fig. 3B**). To confirm this association, we repeated the analysis using an independent scRNA-seq dataset from mouse (Tabula Muris) [28] and found a similar enrichment for "Recurrent Neisserial infections" in two hepatic cell types, namely Kupffer cells (p = 0.0094) and hepatoblasts ($p = 2.23 \times 10^{-308}$).

Exemplar results identify known associations while revealing novel multi-scale disease mechanisms

Here we highlight several exemplary results that recapitulate known aspects of disease aetiology while revealing a more comprehensive view of the disease by connecting mechanisms at multiple scales: phenotype ancestors (groups), phenotypes, cell types, genes. One such example is the association between respiratory failure and bronchiolar cells, alveolar epithelial cells, ciliated epithelial cells, and skeletal muscle cells. Specifically, the two airway epithelial cells initiate local and systemic inflammation, which lead to alveolar hypoventilation and eventual respiratory failure [29]. The weakening of the diaphragm, the primary respiratory muscle, can independently lead to life-threatening respiratory failure [30]. Additional cell type specificity filtering and sorting identified the gene *CCNO* acting via ciliated epithelial cells as the most promising target for respiratory failure.

As a second example, "Recurrent Neisserial infections" were significantly enriched for both alpha-Fetoprotein (AFP) / Albumin (ALB) -positive cells (fold-change=11.517, p=0.00010, q=0.00847) and hepatoblasts (fold-change=9.902, p=0.00016, q=0.0125). In both cell types, the associations with the phenotype are mediated by the same set of complement system genes: *C7, C5, C6, C8B, CFB, CFI,* and *MBL2*. Hepatoblasts are the precursor cells to hepatocytes (the primary cell type of the liver). AFP/ALB-positive cells are a canonical biomarker for liver damage or hepatocarcinoma in adults, but are also produced in normally developing foetuses [31].

Third, "Mental deterioration" is a phenotype characterised by "Loss of previously present mental abilities, generally in adults" that is associated with several forms of amyloidosis, leukodystrophy, and a variety of other degenerative neurological conditions (https://hpo.jax.org/app/browse/term/HP:0001268). As expected, "Mental deterioration" was strongly associated with neurons of the central nervous system (excitatory, inhibitory, limbic system, and Purkinje neurons). However, amacrine and ganglion cells of the retina were also significantly enriched, primarily mediated through the genes *SNORD118*, *APOE, CHCHD10 and CSTB*.

Prioritising cell type-specific gene targets for severe disease phenotypes

Next, we identified putative cell type-specific gene targets for several severe disease phenotypes. After all filtering and sorting steps, there remained 62 gene targets associated with 78 phenotypes across 26 cell types (**Fig. 4**). These prioritised targets were then visualised as a directed graph (**Fig. 5**). Grouped by higher-order ontology category, “Abnormality of the nervous system” had the greatest number of enriched phenotypes (26 phenotypes, 36 genes), followed by “Abnormality of the cardiovascular system” (17 phenotypes, 15 genes), “Abnormality of the musculoskeletal system” (8 phenotypes, 16 genes), “Abnormality of the respiratory system” (6 phenotypes, 5 genes), and ”Abnormality of the eye” (5 phenotypes, 15 genes).

Within the “Abnormality of the nervous system” category, 11 different “Abnormality of higher mental function” / “Neurodevelopmental abnormality” phenotypes survived the prioritisation filters, including: "Coma", "Developmental regression", "Global developmental delay", "Intellectual disability", "Intellectual disability, mild", "Intellectual disability, moderate", "Intellectual disability, severe", "Mild global developmental delay", "Neurodevelopmental abnormality", "Neurodevelopmental delay", "Severe global developmental delay". The most common cell types enriched within these phenotypes were excitatory and granule neurons (both enriched in 6/11 phenotypes), followed by Inhibitory neurons (5/11 phenotypes). Across these phenotypes, the most commonly appearing genes were *SOX3* (appearing in 17 cell type-phenotype associations), *SOX2* (12 associations), *POU3F4* (9 associations), and *FOXH1* (8 associations), and . However, none of the “Mental deterioration” targets survived the filters due to the low cell type specificity (median quantile=6/40) and expression levels (median quantile=6/40) of the target genes. Unlike the other phenotypes in these categories, “Coma” was strongly enriched for islet endocrine cells (**Fig. 5E**). This association was mediated through genes critical for glucose regulation, such as *INS* and *KCNJ11.*

The “Abnormality of the nervous system” phenotypes also included non-cognitive phenotypes. Specifically, within the “Abnormality of movement” subcategory, the phenotype “Inability to walk”, which was enriched for both excitatory neurons ($p<2.23 \times 10^{-308}$, $q<2.23 \times 10^{-308}$, fold-change=1.832, prioritised gene target=*FOXG1*) and Schwann cells (p=0.00071, q=0.0421, fold-

change=1.546, prioritised gene target=*NHLRC1*). Second, the “Seizure” subcategory included the phenotype “Status epilepticus”, which was enriched for excitatory neurons ($p<2.23 \times 10^{-308}$, $q<2.23 \times 10^{-308}$, fold-change=2.147, prioritised gene target=*FOXG1*). Finally, there were 13 phenotypes belonging to “Morphological central nervous system abnormality”, which included a variety of neuroanatomical features (e.g. “Cerebellar atrophy”, “Lissencephaly” and “Hypoplasia of the corpus callosum”) and “Stroke” (enriched for cardiomyocytes and stellate cells).

17 phenotypes within the “Abnormality of the cardiovascular system” category remained after the target prioritisation filtering. Of those, “Arrhythmia” showed strong enrichment for Cardiomyocytes ($p<2.23 \times 10^{-308}$, $q<2.23 \times 10^{-308}$, fold-change=2.915; **Fig. 5B**), with six prioritised target genes (*NPPA, TNNC1 NKX2-5, TCAP, KCNA5*). Of those genes, *NKX2-5* is annotated within the HPO as being very frequently associated with arrhythmia (~72% of cases on average). *NKX2-5* is a transcription factor previously demonstrated to have highly specific expression in heart tissue, which is in congruence with the fact that this gene belongs to the top quantile (40) within both our specificity quantile and mean expression quantile metrics. This gene was in fact the first known genetic risk factor for congenital heart disease, and its expression is necessary not only for the development of cardiomyocytes but also the continued functioning of heart cells into adulthood [32,33].

Within the “Abnormality of the musculoskeletal system” there were 9 unique phenotypes that survived the prioritisation pipeline: "Generalized hypotonia", "Spasticity", "Hypotonia", "Distal amyotrophy", "Spastic tetraplegia", "Abnormality of upper limb joint", "Aplasia/hypoplasia of the extremities", "Aplasia/hypoplasia involving bones of the extremities". As an example, “Hypotonia” was highly enriched for a variety of neuronal and glial cell types (**Fig. 5D**).

Finally, for a more comprehensive list of putative targets across a wider variety of phenotypes, we removed or relaxed many of the default arguments in our prioritisation pipeline (see Methods for details). This yielded putative therapeutic targets for 1,307 phenotypes across 37 cell types and 246 genes. Across all phenotypes, excitatory neurons were commonly implicated (236 phenotypes), followed by antigen presenting cells (214 phenotypes),

cardiomyocytes (183 phenotypes), limbic system neurons (173 phenotypes), enteric nervous system (ENS) glia (167 phenotypes), and and ganglion cells (163 phenotypes).

Both the reduced and the extended versions of the prioritised targets network, as well as all code to reproduce them, are available as an interactive report online: https://neurogenomics.github.io/RareDiseasePrioritisation/reports/prioritise_targets

Genetic correlations reveal cross-phenotypic pleiotropy

Pairwise correlations between all phenotypes within the (reduced) prioritised targets based on gene annotation overlap. Phenotypes were then grouped into three k-mean clusters. Based on the phenotypes present in each cluster, cluster 1 appeared to be a mixture of different phenotype categories and included a subcluster of vascular abnormalities across multiple anatomical systems (e.g. “Peripheral arteriovenous fistula”, “Abnormal cerebral vascular morphology”, “Stroke”). Clusters 2 and 3 corresponded closely to “Abnormality of the nervous system” and “Abnormality of the cardiovascular system”, respectively. Within cluster 3 there was a subcluster corresponding to congenital heart conditions (e.g. “Abnormal aortic valve cusp morphology” and “Congenital malformation of the great arteries”).

Discussion

By applying our systematic approach to the HPO and the Descartes cell type atlas, we identified 8,379 significant cell type-phenotype associations across 2,832 unique phenotypes. All 77 cell types were enriched in at least one phenotype. Enriched phenotypes were distributed across all ontology levels within the HPO (**Fig. 1B**). From more general terms such as “Abnormality of the immune system”, down to more specialised (children) terms such as “Mycobacterial infections”. Enrichment tests revealed expected as well as novel cell type-phenotype associations. Whilst terms related to the “Abnormality of the cardiovascular system” highlighted cardiomyocytes, a significant number of these terms were also associated with hepatoblasts. More specific phenotypes with lower ontology levels were associated with fewer cell types and genes, but higher cell type-specificity of gene expression. Taken together, this suggests that the lower ontology level phenotypes may provide more viable avenues for

therapeutics discovery. Being able to model the phenotype using a small set of prioritised genes and a specific cell type will not only benefit the RDs associated to the phenotype in question, but can also guide repurposing of the cell type-specific gene therapy for treatment of similar phenotypes, or even unrelated phenotypes that are underlied by genes and/or cell types with similar functions.

We then applied our computational prioritisation pipeline eto identify putative cell type-specific gene targets with the greatest chance of success as therapeutics (**Fig. 5**). This pipeline was based on a variety of relevant criteria, including disease severity, cell type-specificity, and gene-disease association frequency (**Fig. 4**). These targets spanned 78 severe disease phenotypes from Tiers 1 and 2, as classified by Lazarin et al. [34]. For the phenotype "Respiratory failure", we prioritised the gene *CCNO* acting via ciliated epithelial cells. A review of the literature revealed that Primary Ciliary Dyskinesia (PCD) is known to act via cilia of the respiratory epithelium [35], and is especially severe in patients with mutations to the *CCNO* gene [36–38]. This example result demonstrates our approach is capable of identifying true positive cell type-phenotype relationships. However, based on a search of the literature and *ClinicalTrials.gov* (see **Supplementary Information** for search result links), it appears that therapeutics targeting *CCNO* have yet to be developed.

As a second example, "Coma" was found to be closely associated with islet endocrine cells, which regulate the secretion of insulin and glucagon, hormones that play a role in blood glucose homeostasis [39]. Our target prioritisation procedure narrowed down the results to two gene targets for "Coma"; the insulin gene (*INS)* and potassium inwardly rectifying channel subfamily J member 11 (*KCNJ11*), which encodes for K-ATP channels that trigger insulin release in response to circulating glucose levels. Mutations in either of these genes can cause permanent neonatal diabetes and induce diabetic coma [40], a condition that can occur when blood glucose levels become too high or too low [41–43]. This result provides further evidence that our framework recovers valid relationships between phenotypes, cell types, and genes.

For the phenotype "Mental deterioration" we report expected associations with central nervous system neurons, as well as less expected associations with amacrine and ganglion cells of the

retina. Prior studies have shown that visual impairment or blindness is prevalent in people with intellectual disability (>8.5-fold increased risk) [44–46] and pathological cognitive decline in older individuals [47]. These results suggest that not only are these phenotypes correlated with one another, but they are causally related to the same genetic risk factors. Although "mental deterioration" did not survive the final target gene prioritisation filters, six other intellectual disability phenotypes did (**Table 1**). Within these five phenotypes, the most commonly appearing genes were *SOX3* (12 times across multiple cell types), *SOX2* (10 times), *FOXH1* (7 times), and *POU3F4* (7 times). All four of these genes are transcription factors that play an important role in the development of the nervous system. The disruption of *SOX3* has been implicated in a variety of intellectual disabilities through observations in both patients and experimental models [48,49]. *SOX2* has also been implicated in nervous system development and its disruption can lead to profound deficits in cognition, vision, and motor function [50]. Interestingly, *POU3F4* has primarily been implicated in the development of semicircular canals and inherited deafness [51–54] but has also been linked to deficits in cognition and mental health (including attention deficit hyperactivity and developmental language disorder) which are significantly comorbid with this form of deafness [55]. These non-auditory deficits are more profound than those observed in controls with other forms of deafness, suggesting that *POU3F4* provides a common molecular aetiology underlying aspects of both central and peripheral nervous system development [56,57]. Confirming the relevance of these results, all enriched cell types within intellectual disability phenotypes were neuronal or glial cells. However, given the ethical implications and technical constraints of treating a genetic disorder during gestation, these gene targets should be considered candidates for further preclinical research, rather than therapeutic targets in developing human embryos.

Phenotypes at both higher- and lower levels of the HPO ontology were predominantly associated with their expected cell types (**Figs. 1C-F**). This testified to the credibility of this approach and allowed us to explore novel findings. One such example is the association of "Recurrent Neisserial infections" with hepatoblasts. Whilst unexpected, a convincing explanation involves the complement system, a key driver of innate immune response to Neisserial infections. Hepatocytes, which derive from hepatoblasts, produce the majority of complement proteins [58], and Kupffer cells express complement receptors [59]. In addition,

individuals with deficits in complement are at high risk for Neisserial infections [60,61], and a genome-wide association study in those with a Neisserial infection identified risk variants within complement proteins [62]. Indeed, all seven of the genes mediating this cell type-phenotype association (*C7, C5, C6, C8B, CFB, CFI,* and *MBL2*) are part of the complement system. While the potential of therapeutically targeting complement in RDs (including Neisserial infections) has been proposed previously [63,64], performing this in a gene- and cell type-specific manner may help to improve efficacy and reduce toxicity (e.g. due to off-target effects). Importantly, there are over 56 known genes within the complement system (see **Supplementary Information**) [65], highlighting the need for a systematic, evidence-based approach to identify effective gene targets.

Finally, we interrogated shared genetic mechanisms between our prioritised RDs and other phenotypes (**Fig. 6**). This allowed us to infer which phenotypes tend to co-occur in patients due to pleiotropy, in which mutations in the same gene cause multiple phenotypes. Sometimes the links between the phenotypes are expected due to their being highly related to one another within the HPO, as is the case for multiple phenotypes of abnormal corpus callosum morphology (a subcluster within cluster 2; **Fig 6**). However other phenotypic relationships are less immediately obvious, such as that between “Abnormality of neuronal migration” and “Abnormality of mouth shape”, or between “Severe global developmental delay” and “Optic atrophy”. These insights may help to improve diagnostic criteria for various RDs while simultaneously revealing the cell type-specific genetic mechanisms underlying their clinical comorbidity.

Conclusions

Across the 77 cell types and 6,173 RD-associated phenotypes investigated, more than 8,000 significant cell type-phenotype associations were observed. The examples we have highlighted above align with what is expected, already known, or at least has a plausible biological explanation. Furthermore, he terms “abnormality of the cardiovascular system” and “recurrent Neisserial infections” were both associated with liver cell types, highlighting the potential in investigating and treating RDs collectively. Within the >8,000 enrichments we identified, there

will be many previously understudied or unknown links between RD-associated phenotypes and specific cell types. In addition to prioritising cell type-specific gene targets, our approach presents an opportunity to therapeutically treat multiple phenotypes via the same target. This may be especially effective in patients that express more than one disease phenotype, as is frequently the case [66]. Taken together, this reflects the utility and potential of our approach in advancing understanding, modelling, and treatment of RDs.

For the impact of our results to be fully realised, it was essential that they could be easily accessed and navigated by domain experts, clinicians, and patients alike. To facilitate this, we developed a publicly available interactive web app (https://neurogenomics.github.io/rare_disease_celltyping_apps/home). Importantly, this web app does not require any coding expertise to search for, visualise, and download relevant subsets of our enrichment results. Together with the reproducible workflows available as R packages, we aim to make our high-throughput findings useful to a wide variety of RD stakeholders and facilitate the extension of these analyses as new RD data becomes available over time. Ultimately, we hope that this work will help to overcome some of the difficulties that have hindered RD research in the past and accelerate the development of effective therapeutics across a wide variety of disorders.

Methods

Cell type-phenotype associations

In this study, the gene by cell type specificity matrix was constructed using the Descartes human cell atlas of fetal gene expression, which contains 377,456 cells representing 77 distinct cell types [13]. To independently replicate our findings, we also used the *Tabula Muris* murine whole-body dataset, made up of 100,605 cells representing 38 distinct cell types from 20 organs and tissues [28]. Genes from the Tabula Muris dataset were converted to human orthologs using the *One2One* R package, and genes without 1:1 mouse:human orthologs were dropped. For each cell type, the specificity metric was obtained by dividing the expression of each gene by the sum of the expression of that gene in all cell types. The target gene sets used

here are obtained from the HPO, such that each phenotype has its own associated gene set. If a given gene set is significantly enriched in a cell type, then it is likely that the cell type plays a role in the pathology and therefore may be a valuable target for future research.

We used *EWCE* (v1.1.0) to evaluate significant cell type-phenotype associations [12]. *EWCE* takes as input a gene by cell type specificity matrix, a target gene list of length *n*, referred to as , and a set of background genes referred to as . Where is the specific expression of gene in cell type . is the number of , and is an expression of in cell (indexed from). As *EWCE* requires input genes per test, 6,173 HPO gene lists remained after filtering.

Variable definitions

target gene list

length of target gene list

: background gene list

: gene identity

cell type identity

: cell index

: number of cell types

: specific expression of gene in cell type

: expression of gene in cell

: gene by cell type specificity matrix

: disease-associated phenotype identity

$$e_{g,c} = \frac{\sum_{i=1}^{|L|} F(g,i,c)/N_c}{\sum_{r=1}^{k}(\sum_{i=1}^{|L|} F(g,i,r)/N_r)}$$

Genes with very low expression were considered to be uninformative and were therefore removed before computing the specificity matrix (mean < 0.2 across all cell types).

$$F(g,i,c) = \begin{cases} r_{g,i}, & l_i = c \\ 0, & l_i \neq c \end{cases}$$

We then summed the specificity scores of the genes in to get each gene's total expression specificity score in a given cell (). This is done for all cells, enabling us to quantify the level of specific expression of gene list (indexed by) in each cell type ().

$$\gamma(X, c) = \sum_{g \in X} e_{g,c}$$

Bootstrapping was then used to determine the probability of cell type-specific enrichment for each cell type in target gene list . We used 100,000 bootstrap iterations to ensure robustness and reduce the rate of false positive associations. To do this, the same cell type-specific expression calculation described above is then calculated for 100,000 random gene sets in each cell type . This gives a probability distribution of cell type-specific expression for gene sets of length in any given . The mean and standard deviation of this distribution are normalised (centred to 0 and 1 respectively) and then used to calculate a Z-score. We can then determine the probability of enrichment of in based on the number of bootstrap gene lists that have a higher cell type specific expression than . Gene sets with higher specific expression than most random gene sets of the same length have a high probability of enrichment in a given cell type. This procedure was repeated for each RD-associated phenotype .

$$P(X \text{ enriched for } c) = \frac{\sum_{j=1}^{100000} \begin{cases} 1 & \gamma(X, c) > \gamma(D_j, c) \\ 0 & \gamma(X, c) < \gamma(D_j, c) \end{cases}}{100000}$$

In total, 475,321 *EWCE* enrichment tests were performed using the Descartes cell type signature reference (6,173 phenotypes x 77 cell types). An additional 213,028 tests were performed using the Tabula Muris cell type reference for the phenotypes that had at least four remaining genes after removing genes without 1:1 mouse:human orthologs (5,606 phenotypes x 38 cell types). Within the results from each cell type reference, *EWCE* p-values were corrected with the Benjamini-Hochberg method to produce q-values [67]. To facilitate these analyses and to make them more easily reproducible by others, we created several open-

source R packages. *MultiEWCE* (https://github.com/neurogenomics/MultiEWCE) facilitates the analysis of multiple gene lists across many computing cores in parallel, reducing the time necessary to complete large-scale enrichment testing. *HPOExplorer* (https://github.com/neurogenomics/HPOExplorer) aids in managing and querying the directed acyclic ontology graph within the HPO.

Interactive website

The landing page for the website was made using HTML and CSS, and the web apps were created using the Shiny Web application framework for R and deployed on the ShinyApps server. The website can be accessed here: https://neurogenomics.github.io/rare_disease_celltyping_apps/home

Gene therapy target identification

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

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

1. **start**: All cell type-phenotype association results.
2. **q_threshold**: Keep only results that were significant after multiple-testing correction ($q<0.05$).
3. **fold_threshold**: Keep only results with fold change>=1.
4. **keep_ont_levels**: Keep only phenotypes at certain absolute ontology levels within the HPO.
5. **keep_onsets**: Keep only phenotypes with postnatal age of onsets to circumvent technical and ethical challenges associated with antenatal gene therapeutics delivery.
6. **keep_tiers**: Keep only phenotypes with high severity Tiers.

   a. We used a combination of manual curation and automated text-based substring queries to assign each phenotype a severity Tier as characterised in a survey of healthcare professionals [34].
   b. Tier 1: Diseases that shortened life span in adolescence or earlier or resulted in intellectual disability.
   c. Tier 2: Diseases that shortened lifespan prematurely in adulthood, or resulted in impaired mobility or internal physical malformation.
   d. Tier 3: Diseases causing sensory impairments (hearing, vision, touch, pain, or other), immunodeficiency/cancer, mental illness, or dysmorphic features.
   e. Tier 4: Diseases that reduce fertility. Of the 49 phenotypes that were available in this severity ranking, we selected three that were classified as Tier 1 (the most severe disease category): mental deterioration, coma and respiratory failure.
7. **severity_threshold**: Keep only phenotypes with mean severity score equal to or below the threshold.
   a. Severity scores were computed by assigning each severity modifier term found in the HPO annotations a numerical value. In order of increasing severity:
   b. HP:0012825 "Mild" (Severity_score=4)
   c. HP:0012827 "Borderline" (Severity_score=3)
   d. HP:0012828 "Severe" (Severity_score=2)
   e. HP:0012829 "Profound" (Severity_score=1)
8. **pheno_frequency_threshold**: Keep only phenotypes with mean frequency equal to or above the threshold (i.e. how frequently a phenotype is associated with any diseases in which it occurs).
   a. Keep phenotypes with a mean frequency ≥10% or are NA by default.
9. **keep_celltypes**: Keep only terminally differentiated cell types.
   a. Of the 77 cell types tested in the Descartes cell type reference, the 40 terminally differentiated cell types were identified through a literature search. Of these, three (extravillous trophoblasts, syncytiotrophoblasts and trophoblast giant cells) were excluded as they only played a role in pregnancy [68–70], which would raise additional technical and ethical challenges as rAAV therapy has not yet been used to target foetuses in clinical trials.

10. **keep_seqnames**: Remove genes on non-standard chromosomes.
    a. Only keep chromosomes 1-22, X, and Y.
11. **gene_size**: Keep only genes <4.3kb in length.
    a. Due to limitations in the length of the gene that can be carried by the rAAV vector, genes with a length of >4.3kb were excluded.
12. **keep_biotypes**: Keep only genes belonging to certain biotypes (e.g. "protein_coding", "processed_transcript", "snRNA", "lincRNA", "snoRNA", "IG_C_gene").
    a. Keep all biotypes by default.
13. **gene_frequency_threshold**: Keep only genes at or above a certain mean frequency threshold (i.e. how frequently a gene is associated with a given phenotype when observed within a disease).
    a. Keep genes with a mean frequency ≥10% or are NA by default.
14. **keep_specificity_quantiles**: Keep only genes in top specificity quantiles from the cell type dataset.
    a. To further narrow down genes, we extracted relevant metrics from the Descartes reference for each gene in each cell type. These included mean expression, specificity, and specificity quantiles (using 40 bins). Only genes with the most specific quantiles (39-40) were included for further analysis, as cell type-specific genes may be less likely to have off-target effects in other cell types.
15. **keep_mean_exp_quantiles**: Keep only genes in top mean expression quantiles from the cell type dataset.
16. **top_n**: Sort candidate targets by a preferred order of metrics and only return the top N targets per cell type-phenotype combination.
    a. Finally, results were sorted by the following columns (in order of precedence, where 1=ascending order and -1=descending order):
    b. "tier"=1
    c. “tier_auto"=1
    d. "Severity_score_mean"=1
    e. "q"=1
    f. "fold_change"=-1

g. "specificity_quantile"=-1
h. "mean_exp_quantile"=-1
i. "specificity"=-1
j. "mean_exp"=-1
k. "pheno_freq_mean"=-1
l. "gene_freq_mean"=-1
m. "width"=1

17. **end**: Final table of prioritised cell type- / phenotype-specific gene targets.

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

Phenotype x phenotype genetic correlations

Lastly, we computed genetic correlations between all phenotypes that appeared within the reduced list of prioritised targets. For this analysis, the complete gene lists for each phenotype were extracted from the HPO (not just the genes present in the prioritised targets list) and recast into a binary gene x phenotype matrix, where 0 indicated the absence of a gene-phenotype association and 1 indicated the presence of a gene-phenotype association. Pairwise Pearson correlations were then computed between all phenotypes to generate a phenotype x phenotype matrix. Hierarchical clustering was performed on the resulting correlation matrix and visualised as a heatmap using *MultiEWCE::correlation_heatmap*, which utilises the R package *ComplexHeatmap* [71]. Cluster group assignment was determined using 1,000 iterations of k-means where *k*=3.

Data and Code Availability

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

Human Phenotype Ontology

https://hpo.jax.org

Descartes scRNA-seq atlas

https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development

Tabula Muris scRNA-seq atlas

https://tabula-muris.ds.czbiohub.org

Web app

https://neurogenomics.github.io/rare_disease_celltyping_apps/home

HPOExplorer

https://github.com/neurogenomics/MultiEWCE

MultiEWCE

https://github.com/neurogenomics/HPOExplorer

EWCE

https://doi.org/doi:10.18129/B9.bioc.EWCE

Code to replicate analyses

https://github.com/neurogenomics/rare_disease_celltyping

Results for all enrichment tests

https://github.com/neurogenomics/rare_disease_celltyping/tree/master/results

Cell type-specific gene target prioritisation

https://neurogenomics.github.io/RareDiseasePrioritisation/reports/prioritise_targets

Acknowledgements

For the purpose of open access, the authors has applied a creative commons attribution (CC BY) licence (where permitted by UKRI, 'open government licence' or 'creative commons attribution no-derivatives (CC BY-ND) licence' may be stated instead) to any author accepted manuscript version arising.

Funding

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

Figures

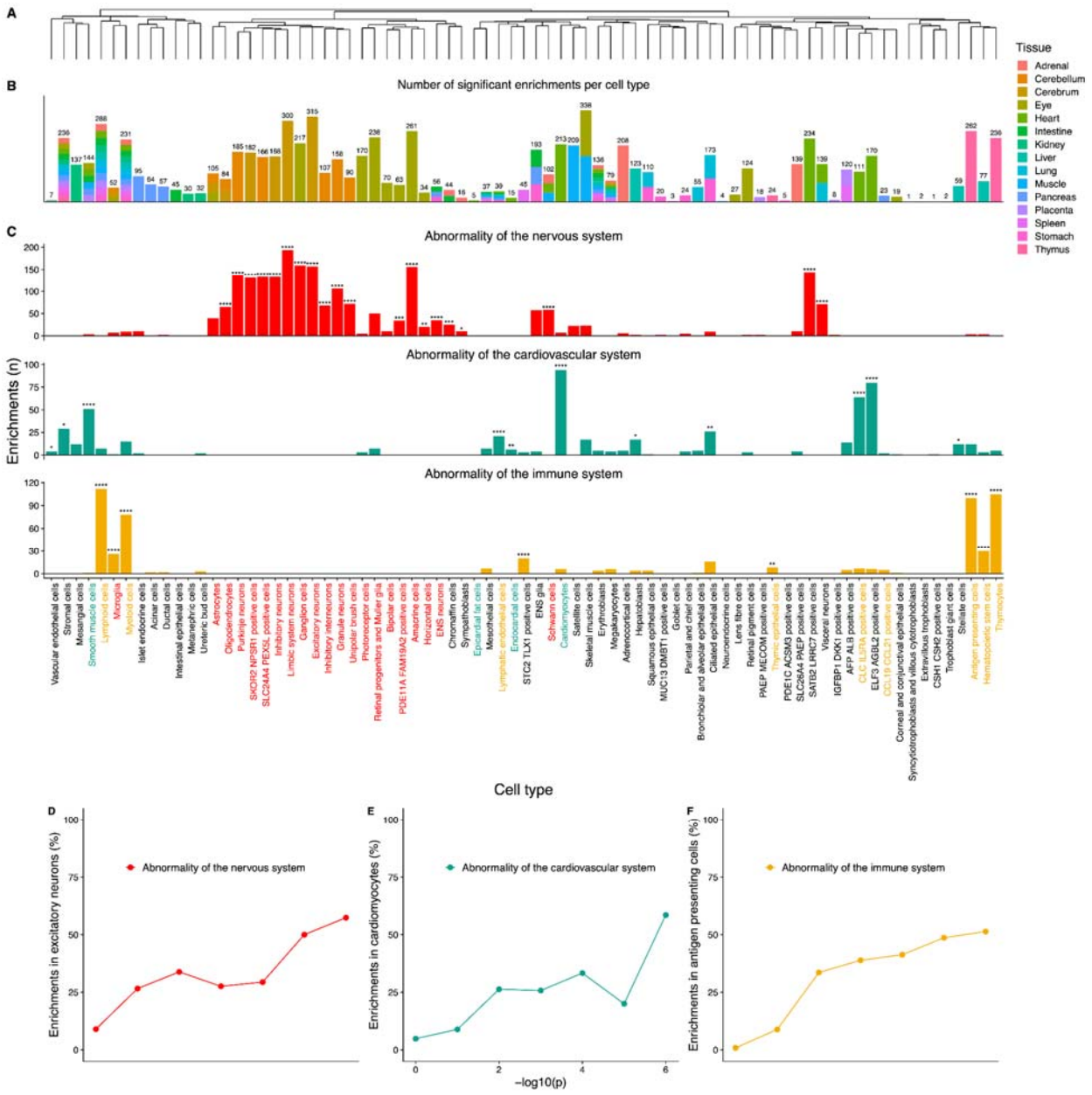


Figure 1. Abnormal nervous system, cardiovascular, and immune phenotypes show expected cell type enrichments.

A. Dendrogram showing the clustering of cell types from the scRNA–seq dataset used. The x-axis is ordered by the dendrogram. **B**. Bar plot showing the number of significant HPO phenotype enrichments for each cell type ($p < 0.05$ and fold enrichment > 1). The colour in

each bar represents the tissue of origin of the cell type. **C**. Bar plot showing the number of phenotype enrichments related to HPO terms abnormality of the nervous system, abnormality of the cardiovascular system, and abnormality of the immune system. A hypergeometric test was used to determine which cell types had significant enrichments. ****,***,**, and *, indicate $p<0.00001$, $p<0.0001$, $p<0.001$, and $p<0.05$, respectively. **D**. Scatter plot of the percentage of phenotype enrichments in excitatory neurons against the enrichment significance threshold. As you decrease the significance threshold, the percentage of phenotype enrichments related to the abnormality of the nervous system increases. **E**. Scatter plot of the percentage of phenotype enrichments in cardiomyocytes against the enrichment significance threshold. As you decrease the significance threshold, the percentage of phenotype enrichments related to the abnormality of the cardiovascular system increases. **F**. Scatter plot of the percentage of phenotype enrichments in antigen presenting cells against the enrichment significance threshold. As you decrease the significance threshold, the percentage of phenotype enrichments related to the abnormality of the immune system increases.

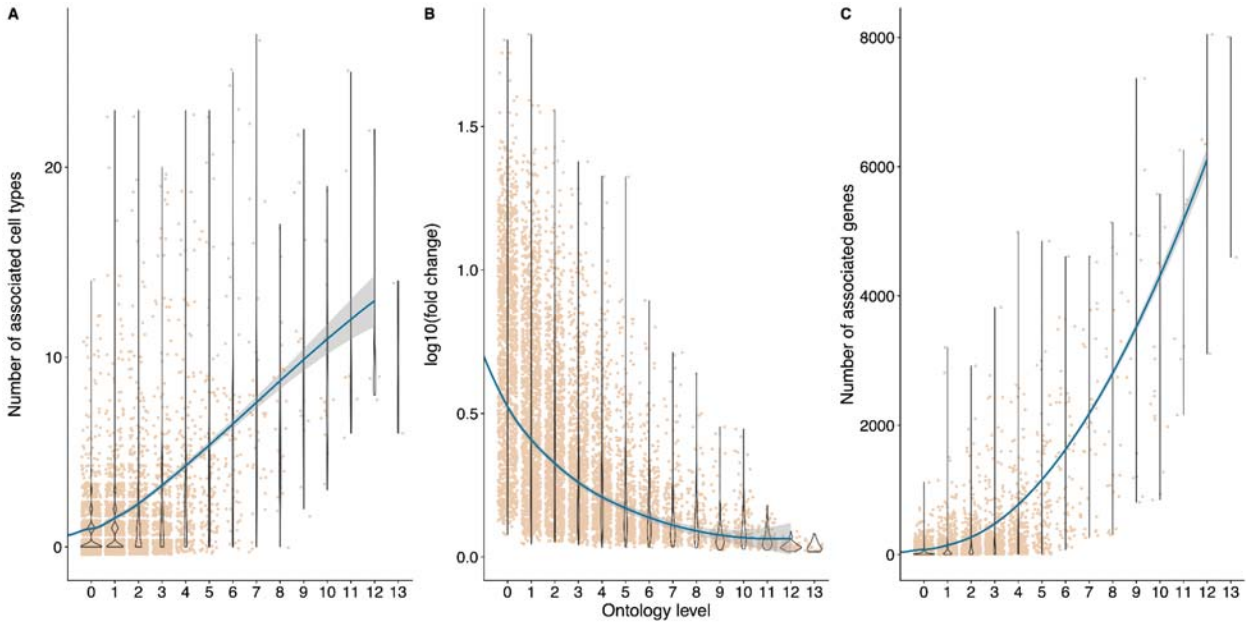


Figure 2. Ontology levels containing more specific phenotypes are associated with a lower number of cell types and genes, but the cell type-specificity of gene expression is higher.

Violin plots showing relationship between HPO ontology level and **A**. the number of associated cell types. **B**. the cell type-specificity of gene expression. **C**. the number of associated genes. Ontology level 12 represents the most broad HPO term: “phenotypic abnormality”.

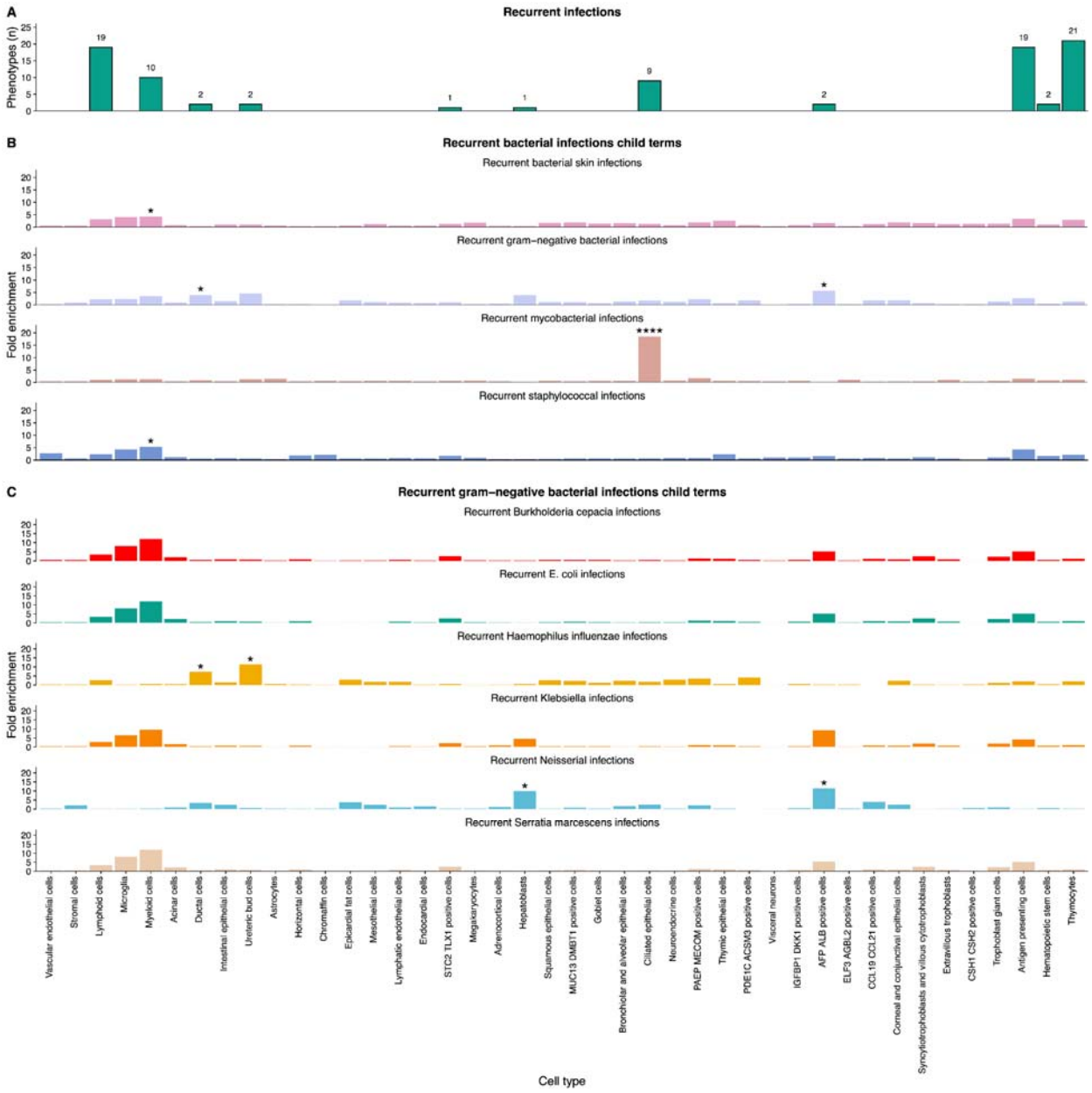


Figure 3. Immune-related phenotypes highlight both expected and unexpected cell type associations.

A. Bar plot showing the number of significant phenotype enrichments related to recurrent infections, for each cell type. **B**. Bar plots showing the number of enrichments related to the child terms of recurrent bacterial infections, for each cell type. **C**. Bar plots showing the number of enrichments related to the child terms of recurrent gram negative bacterial infections. ****,***,**, and *, indicate $p<0.00001$, $p<0.0001$, $p<0.001$, and $p<0.05$, respectively.

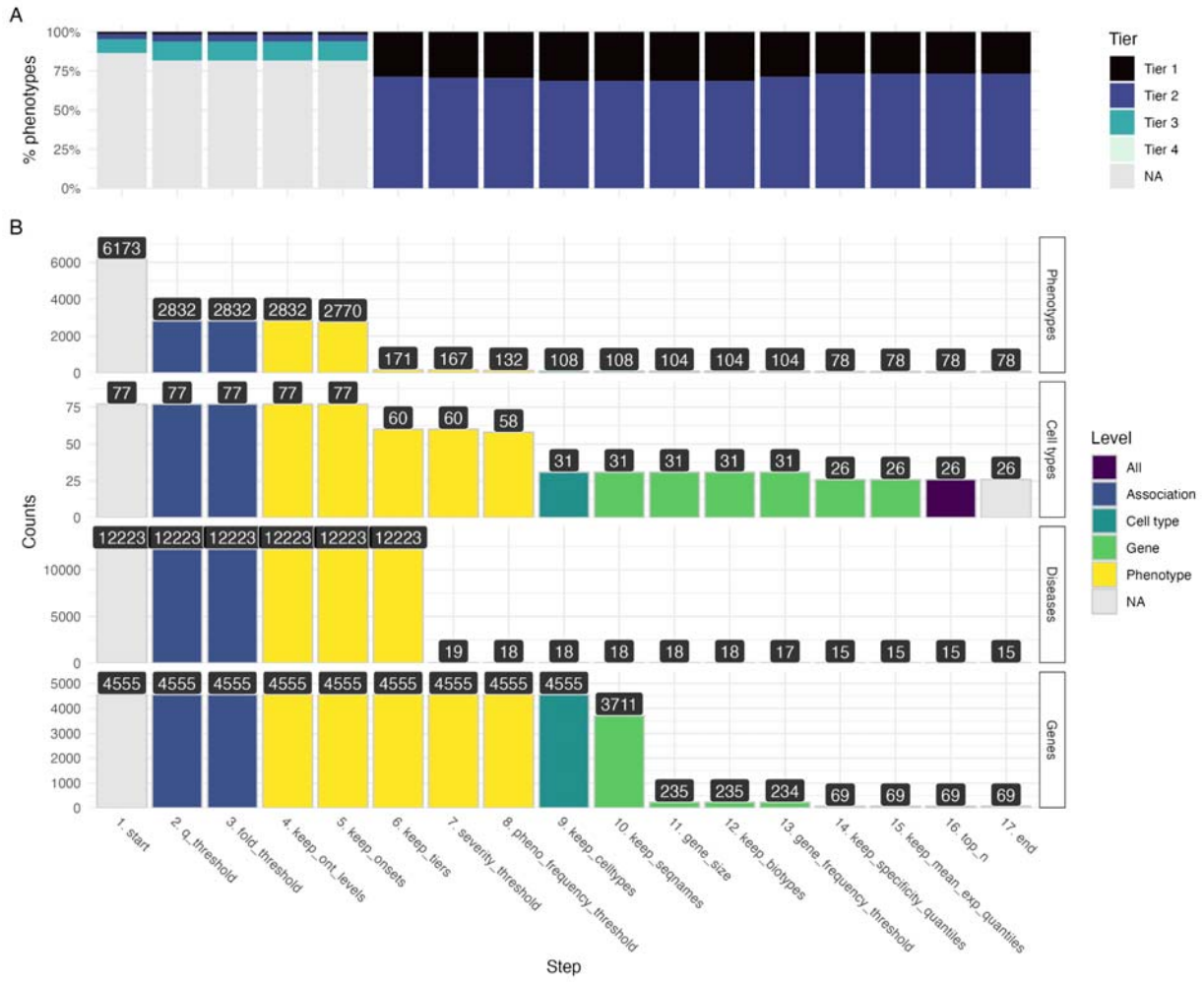


Figure 4. Prioritised target filtering steps

This plot visualises prioritised targets using the default parameters in *MultiEWCE::prioritise_targets* and is fully reproducible using the *MultiEWCE::report_plot* function. Each step in the pipeline can be easily adjusted according to user preference and use case. See **Methods** for descriptions and criterion of each filtering step. **A**. The percentage of phenotypes belonging to each severity Tier after each filtering step (Tier 1 being the most severe). **B**. The number of phenotypes, cell types, associated diseases and genes remaining after each filtering step during the gene prioritisation pipeline.

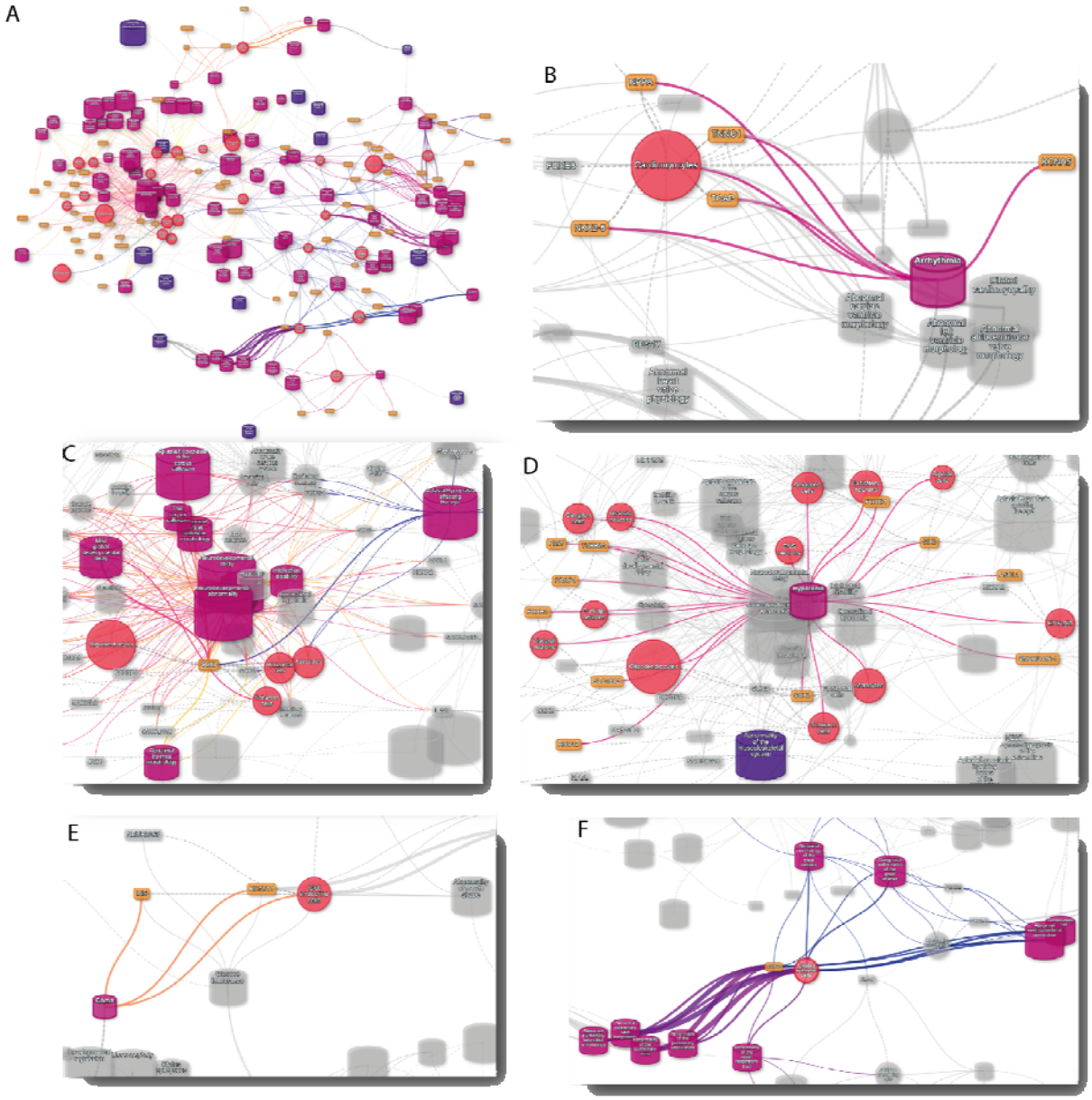


Figure 5. Network of prioritised cell type-specific gene targets

The above network illustrates phenotypes/cell types/genes identified by our target prioritisation strategy. Tier 1/2 phenotypes are connected significantly associated cell types via mediating genes. Each RD phenotype (purple cylinders) is connected to their respective causal cell-types (red circles). RD phenotypes are classified by the higher-order phenotypes to which they belong in the HPO (blue cylinders). Each cell type is in turn connected to the prioritised gene targets (gold boxes) that are driving the cell type-phenotype association, show highly cell type-specific RNA expression, and meet our criterion for rAAV therapeutic applications. The

thickness of the edges connecting the nodes represent the (mean) fold-change from the bootstrapped enrichment tests. Edge colour indicates which phenotype it connects to (grey dotted lines are used for edges that do not connect directly to a phenotype). Nodes were spatially arranged using the Kamada-Kawai algorithm [72].

A. A zoomed out view of the full network. Subsequent subplots are zoomed in sections of this full network. **B**. Nodes connected to the phenotype "Arrhythmia". C. Nodes connected to the gene *SOX3*. **D**. Nodes connected to the phenotype "Hypotonia". **E**. Nodes connected to the phenotype "coma". **F**. Nodes connected to ciliated epithelial cells.

An interactive version of this plot and all code to fully reproduce this plot can be further explored online: : https://neurogenomics.github.io/RareDiseasePrioritisation/reports/prioritise_targets

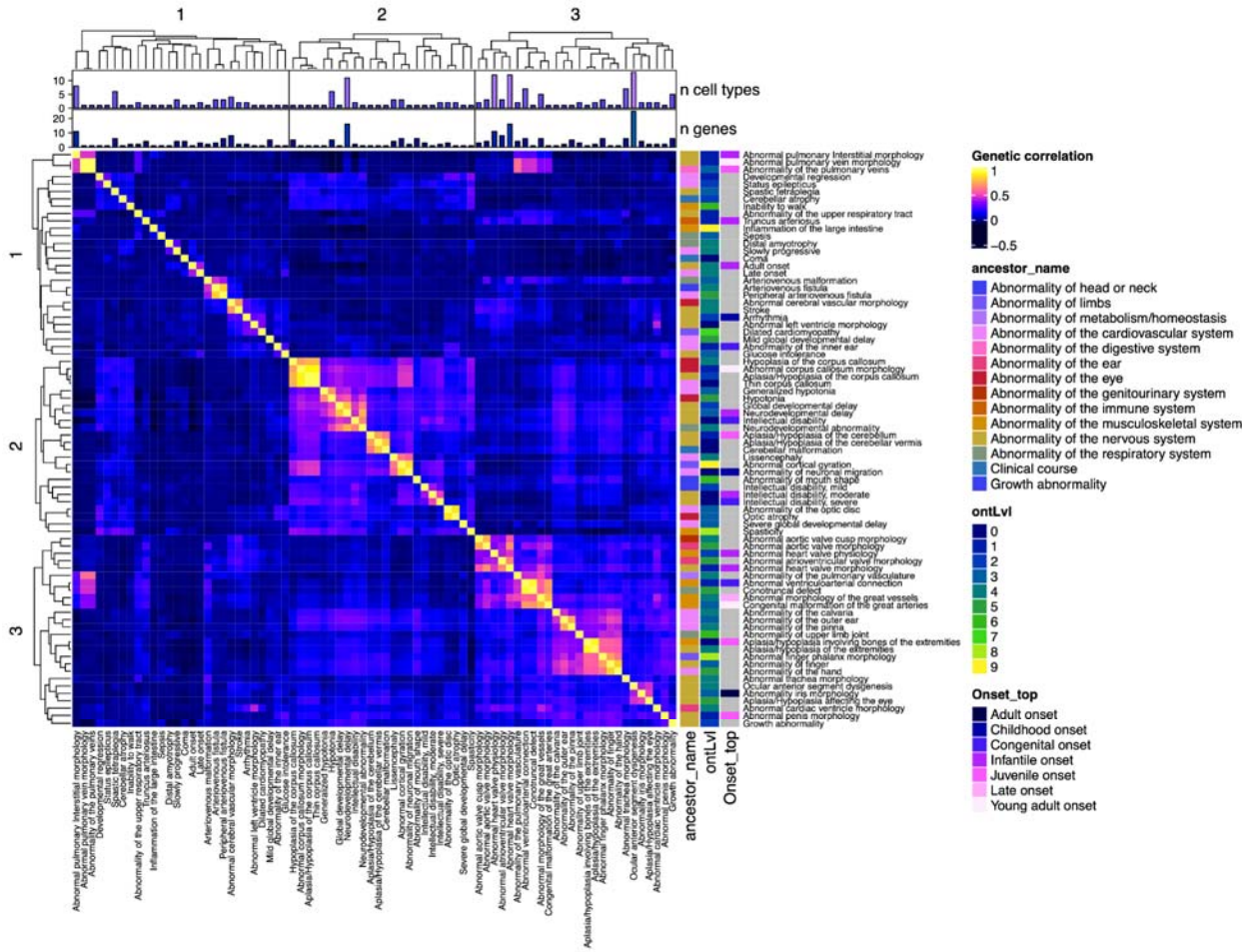


Figure 6. Genetic correlation map between prioritised phenotypes

Heatmap of phenotype-phenotype correlations based on the presence or absence of known associations with genes in the HPO. Rows and columns are hierarchically clusters to identify genetically related groups of phenotypes. Metadata on the top shows the number of unique cell types and genes associated with each phenotype on the x-axis after the prioritisation filtering pipeline was applied. Metadata on the right side indicate the ancestral phenotype to which each phenotype belongs (ancestor_name), and the most frequent age of onset for a given phenotype (Onset_top). This plot was generated using *MultiEWCE::correlation_heatmap.*

Supplementary Information

Links

ClinicalTrials.gov search for "Primary Ciliary dyskinesia":
https://clinicaltrials.gov/ct2/results?cond=primary+ciliary+dyskinesia

ClinicalTrials.gov search for "CCNO":
https://clinicaltrials.gov/ct2/results?cond=&term=ccno

Complement system gene list:
https://www.genenames.org/data/genegroup/#!/group/492

Supplementary Tables

Table S1. Prioritised targets

Cell type- and gene- specific targets for each phenotype. Targets were prioritised using the filtering and sorting procedure implemented in the *MultiEWCE::prioritise_targets* function.
An interactive version of this table (with sorting, searching, and downloading features) is available online:
https://neurogenomics.github.io/RareDiseasePrioritisation/reports/prioritise_targets

Table S2. Cell type groupings

Cell type groupings for testing overrepresentation of nervous system related cell types, immune related cell types, and cardiovascular related cell types from the Descartes dataset in the HPO branches "Abnormality of the nervous system", "Abnormality of the immune system", and "Abnormality of the cardiovascular system", respectively.

Table S3. Cell type-branch enrichment tests

Hypergeometric test results for overrepresentation of cell type-phenotype associations by HPO branch. The selected branches were children terms of "Phenotypic abnormality" and each

phenotype was annotated to a branch if it was a descendant of the branch e.g. recurrent infections was annotated to “Abnormality of the immune system”. Terms that were not descendants of “Phenotypic abnormality” e.g. those related to “Mode of inheritance”, were not included in this analysis. Hypergeometric p-values were corrected with the Benjamini-Hochberg method [67] to produce q-values.

References

1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. A* **179**, 885–892 (2019).
2. Zhu, Q. *et al.* An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). *J. Biomed. Semantics* **11**, 13 (2020).
3. 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).
4. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
5. 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).
6. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
7. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
8. Amberger, J. S. & Hamosh, A. Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. *Curr. Protoc. Bioinformatics* **58**, 1.2.1–1.2.12 (2017).

9. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).

10. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).

11. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, (2020).

12. 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).

13. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).

14. Naso, M. F., Tomkowicz, B., Perry, W. L., 3rd & Strohl, W. R. Adeno-Associated Virus (AAV) as a Vector for Gene Therapy. *BioDrugs* **31**, 317–334 (2017).

15. Flotte, T. R. Size does matter: overcoming the adeno-associated virus packaging limit. *Respir. Res.* **1**, 16–18 (2000).

16. Dong, J. Y., Fan, P. D. & Frizzell, R. A. Quantitative analysis of the packaging capacity of recombinant adeno-associated virus. *Hum. Gene Ther.* **7**, 2101–2112 (1996).

17. Russell, D. W. & Kay, M. A. Adeno-associated virus vectors and hematology. *Blood* **94**, 864–874 (1999).

18. Wörner, T. P. *et al.* Adeno-associated virus capsid assembly is divergent and stochastic. *Nat. Commun.* **12**, 1642 (2021).

19. Darrow, J. J. Luxturna: FDA documents reveal the value of a costly gene therapy. *Drug Discov. Today* **24**, 949–954 (2019).

20. Russell, S. *et al.* Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in

patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet* **390**, 849–860 (2017).

21. Mendell, J. R. *et al.* Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy. *N. Engl. J. Med.* **377**, 1713–1722 (2017).
22. Mueller, C. *et al.* 5 Year Expression and Neutrophil Defect Repair after Gene Therapy in Alpha-1 Antitrypsin Deficiency. *Mol. Ther.* **25**, 1387–1394 (2017).
23. Nguyen, P. *et al.* Liver lipid metabolism. *J. Anim. Physiol. Anim. Nutr.* **92**, 272–283 (2008).
24. Heim, C. E. *et al.* Myeloid-derived suppressor cells contribute to Staphylococcus aureus orthopedic biofilm infection. *J. Immunol.* **192**, 3778–3792 (2014).
25. 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).
26. Stoll, H. *et al.* Staphylococcal Enterotoxins Dose-Dependently Modulate the Generation of Myeloid-Derived Suppressor Cells. *Front. Cell. Infect. Microbiol.* **8**, 321 (2018).
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).
28. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
29. Wang, X., Adler, K. B., Erjefalt, J. & Bai, C. Airway epithelial dysfunction in the development of acute lung injury and acute respiratory distress syndrome. *Expert Rev. Respir. Med.* **1**, 149–155 (2007).
30. Dubé, B.-P. & Dres, M. Diaphragm Dysfunction: Diagnostic Approaches and Management Strategies. *J. Clin. Med. Res.* **5**, (2016).

31. Sell, S. Alpha-fetoprotein, stem cells and cancer: how study of the production of alpha-fetoprotein during chemical hepatocarcinogenesis led to reaffirmation of the stem cell theory of cancer. *Tumour Biol.* **29**, 161–180 (2008).
32. Dixit, R. *et al.* Functional analysis of novel genetic variants of NKX2-5 associated with nonsyndromic congenital heart disease. *Am. J. Med. Genet. A* **185**, 3644–3663 (2021).
33. Schott, J. J. *et al.* Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science* **281**, 108–111 (1998).
34. Lazarin, G. A. *et al.* Systematic Classification of Disease Severity for Evaluation of Expanded Carrier Screening Panels. *PLoS One* **9**, e114391 (2014).
35. Paff, T., Omran, H., Nielsen, K. G. & Haarman, E. G. Current and Future Treatments in Primary Ciliary Dyskinesia. *Int. J. Mol. Sci.* **22**, (2021).
36. Amirav, I. *et al.* Systematic Analysis of CCNO Variants in a Defined Population: Implications for Clinical Phenotype and Differential Diagnosis. *Hum. Mutat.* **37**, 396–405 (2016).
37. Wallmeier, J. *et al.* Mutations in CCNO result in congenital mucociliary clearance disorder with reduced generation of multiple motile cilia. *Nat. Genet.* **46**, 646–651 (2014).
38. Henriques, A. R. *et al.* Primary ciliary dyskinesia due to CCNO mutations-A genotype-phenotype correlation contribution. *Pediatr. Pulmonol.* **56**, 2776–2779 (2021).
39. Hughes, J. W., Ustione, A., Lavagnino, Z. & Piston, D. W. Regulation of islet glucagon secretion: Beyond calcium. *Diabetes Obes. Metab.* **20 Suppl 2**, 127–136 (2018).
40. De Franco, E. *et al.* The effect of early, comprehensive genomic testing on clinical care in neonatal diabetes: an international cohort study. *Lancet* **386**, 957–963 (2015).
41. Karslioglu French, E., Donihi, A. C. & Korytkowski, M. T. Diabetic ketoacidosis and

hyperosmolar hyperglycemic syndrome: review of acute decompensated diabetes in adult patients. *BMJ* **365**, l1114 (2019).

42. Hockaday, T. D. & Alberti, K. G. Diabetic coma. *Clin. Endocrinol. Metab.* **1**, 751–788 (1972).
43. Guthrie, R. A. & Guthrie, D. W. Pathophysiology of diabetes mellitus. *Crit. Care Nurs. Q.* **27**, 113–125 (2004).
44. Warburg, M. Visual impairment in adult people with intellectual disability: literature review. *J. Intellect. Disabil. Res.* **45**, 424–438 (2001).
45. Bowman, R. The importance of assessing vision in disabled children - and how to do it. *Community Eye Health* **29**, 12–13 (2016).
46. Kiani, R. & Miller, H. Sensory impairment and intellectual disability. *Advances in Psychiatric Treatment* **16**, 228–235 (2010).
47. Nagarajan, N. *et al.* Vision impairment and cognitive decline among older adults: a systematic review. *BMJ Open* **12**, e047929 (2022).
48. Tahira, A. C. *et al.* Chapter 13 - Linking SOX3, SRY, and disorders of neurodevelopment. in *Factors Affecting Neurodevelopment* (eds. Martin, C. R., Preedy, V. R. & Rajendram, R.) 143–156 (Academic Press, 2021).
49. Tahira, A. C. *et al.* Putative contributions of the sex chromosome proteins SOX3 and SRY to neurodevelopmental disorders. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **180**, 390–414 (2019).
50. Mercurio, S., Serra, L., Pagin, M. & Nicolis, S. K. Deconstructing Sox2 Function in Brain Development and Disease. *Cells* **11**, (2022).
51. de Kok, Y. J. *et al.* Association between X-linked mixed deafness and mutations in the

POU domain gene POU3F4. *Science* **267**, 685–688 (1995).

52. Su, Y. *et al.* Clinical and molecular characterization of POU3F4 mutations in multiple DFNX2 Chinese families. *BMC Med. Genet.* **19**, 157 (2018).
53. Bernardinelli, E. *et al.* Novel POU3F4 variants identified in patients with inner ear malformations exhibit aberrant cellular distribution and lack of SLC6A20 transcriptional upregulation. *Front. Mol. Neurosci.* **15**, 999833 (2022).
54. Lee, H. K. *et al.* Clinical and molecular characterizations of novel POU3F4 mutations reveal that DFN3 is due to null function of POU3F4 protein. *Physiol. Genomics* **39**, 195–201 (2009).
55. Smeds, H. *et al.* X-linked Malformation Deafness: Neurodevelopmental Symptoms Are Common in Children With IP3 Malformation and Mutation in POU3F4. *Ear Hear.* **43**, 53–69 (2022).
56. Giannantonio, S. *et al.* Genetic identification and molecular modeling characterization of a novel POU3F4 variant in two Italian deaf brothers. *Int. J. Pediatr. Otorhinolaryngol.* **129**, 109790 (2020).
57. Carlson, D. L. & Reeh, H. L. X-linked mixed hearing loss with stapes fixation: case reports. *J. Am. Acad. Audiol.* **4**, 420–425 (1993).
58. Zhou, Z., Xu, M.-J. & Gao, B. Hepatocytes: a key cell type for innate immunity. *Cell. Mol. Immunol.* **13**, 301–315 (2016).
59. Dixon, L. J., Barnes, M., Tang, H., Pritchard, M. T. & Nagy, L. E. Kupffer cells in the liver. *Compr. Physiol.* **3**, 785–797 (2013).
60. Rosain, J. *et al.* Strains Responsible for Invasive Meningococcal Disease in Patients With Terminal Complement Pathway Deficiencies. *J. Infect. Dis.* **215**, 1331–1338 (2017).

61. Ladhani, S. N. *et al.* Invasive meningococcal disease in patients with complement deficiencies: a case series (2008-2017). *BMC Infect. Dis.* **19**, 522 (2019).

62. The International Meningococcal Genetics Consortium. Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nature Genetics* **42**, 772–776 (2010).

63. Reis, E. S. *et al.* Applying complement therapeutics to rare diseases. *Clin. Immunol.* **161**, 225–240 (2015).

64. Lung, T. *et al.* The complement system in liver diseases: Evidence-based approach and therapeutic options. *J Transl Autoimmun* **2**, 100017 (2019).

65. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).

66. Díaz-Santiago, E. *et al.* Phenotype-genotype comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. *PLoS Genet.* **16**, e1009054 (2020).

67. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* (1995).

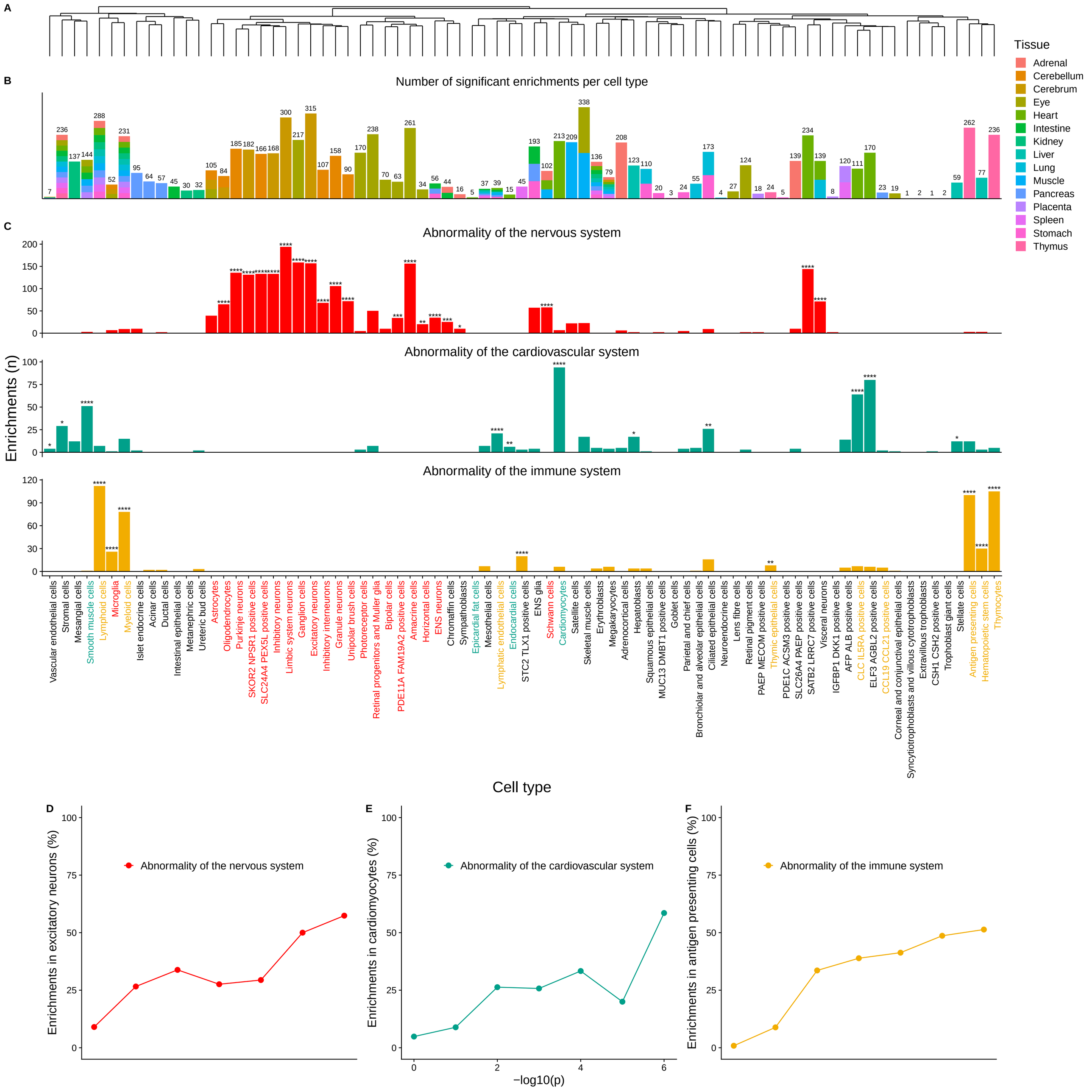
68. Chang, C.-W., Wakeland, A. K. & Parast, M. M. Trophoblast lineage specification, differentiation and their regulation by oxygen tension. *J. Endocrinol.* **236**, R43–R56 (2018).

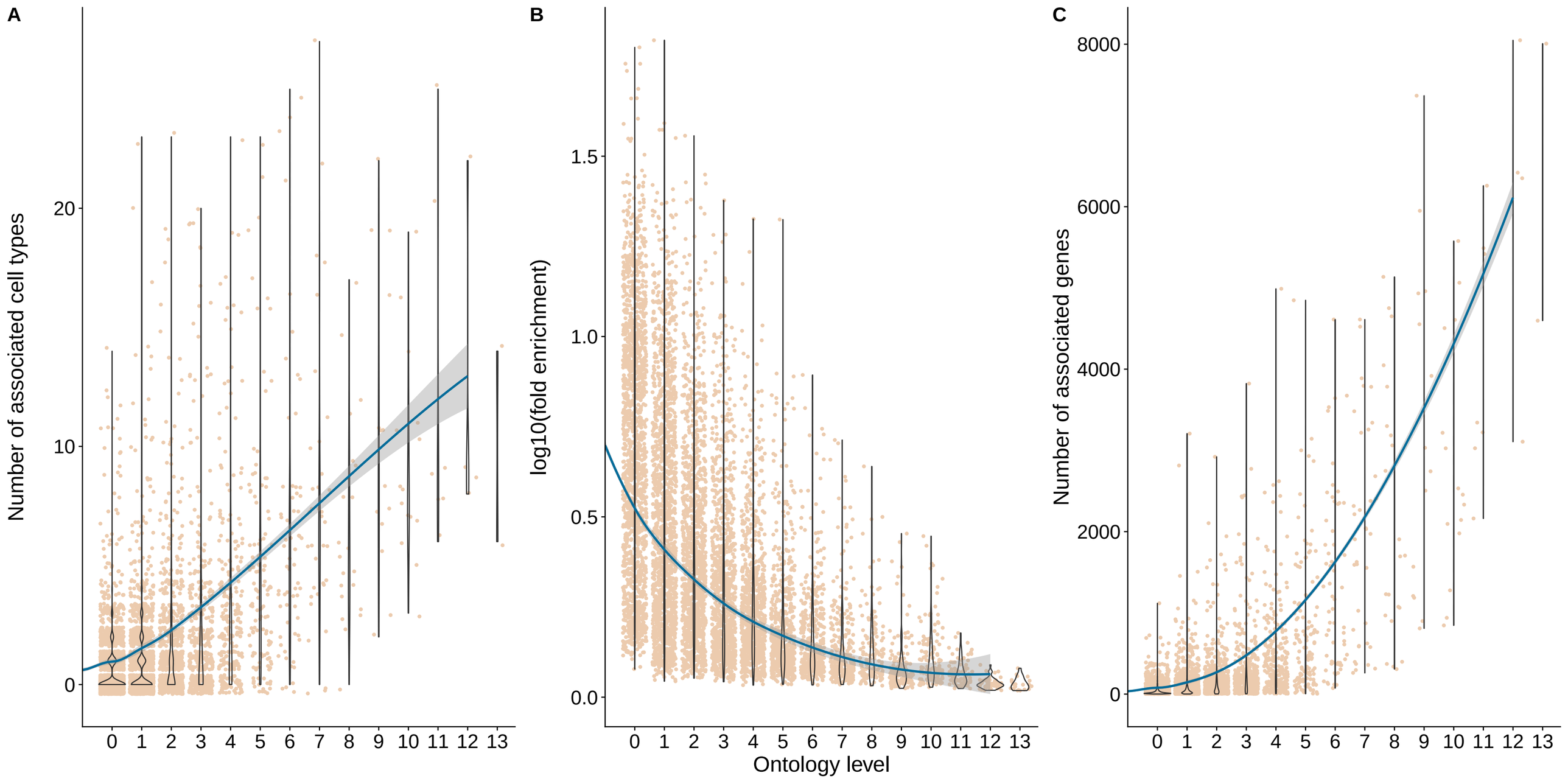
69. Fogarty, N. M. E., Mayhew, T. M., Ferguson-Smith, A. C. & Burton, G. J. A quantitative analysis of transcriptionally active syncytiotrophoblast nuclei across human gestation. *J. Anat.* **219**, 601–610 (2011).

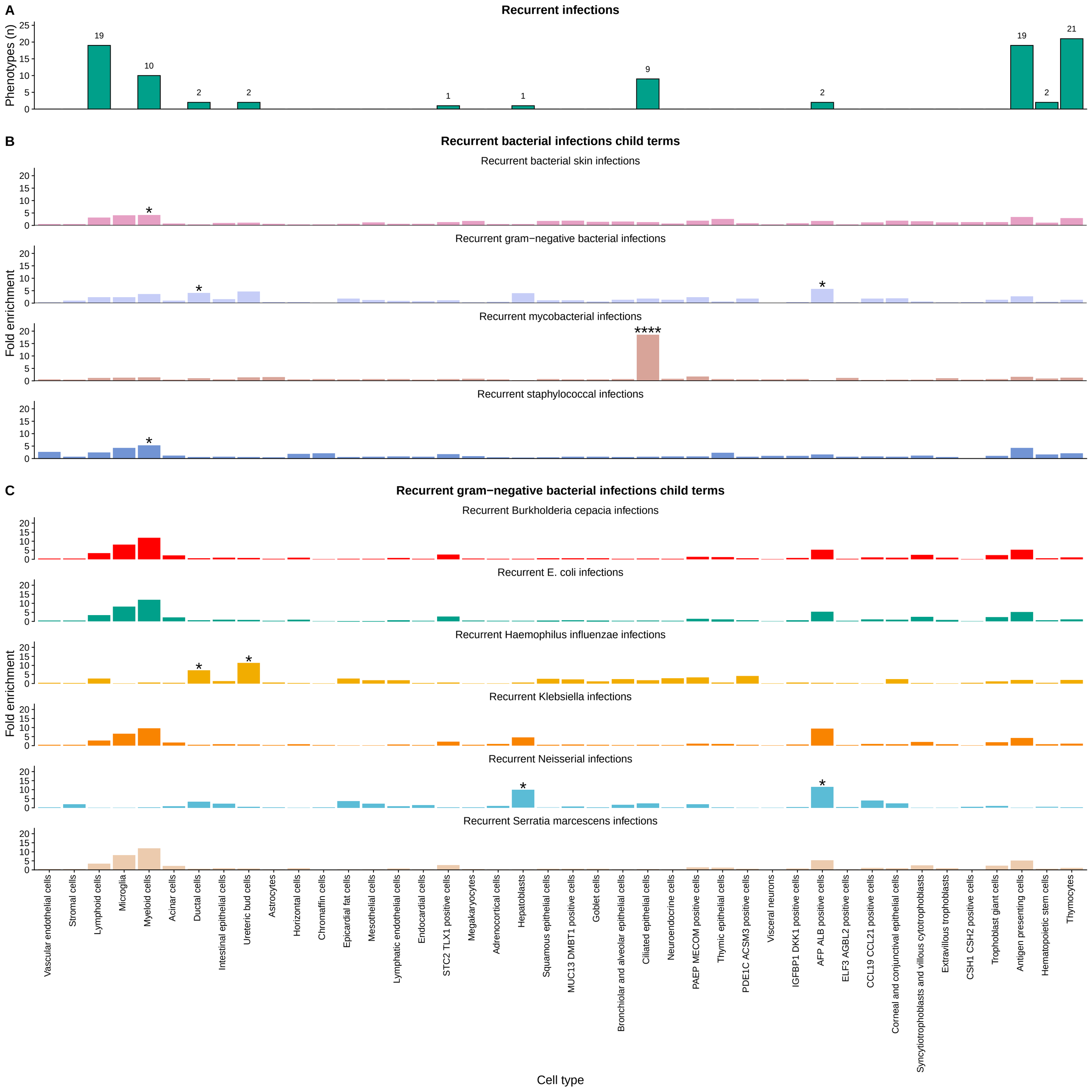
70. Hu, D. & Cross, J. C. Development and function of trophoblast giant cells in the rodent placenta. *Int. J. Dev. Biol.* **54**, 341–354 (2010).

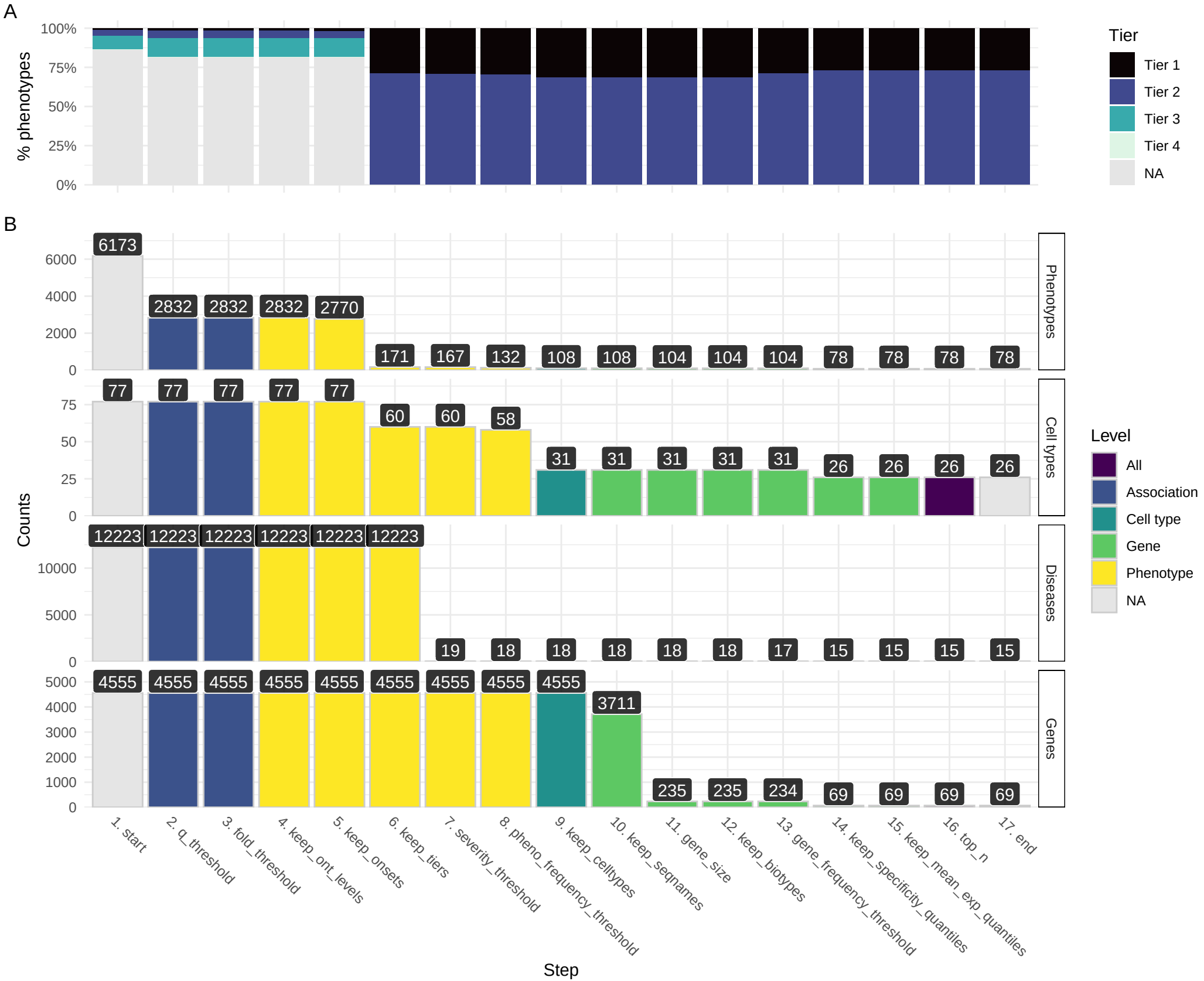
71. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

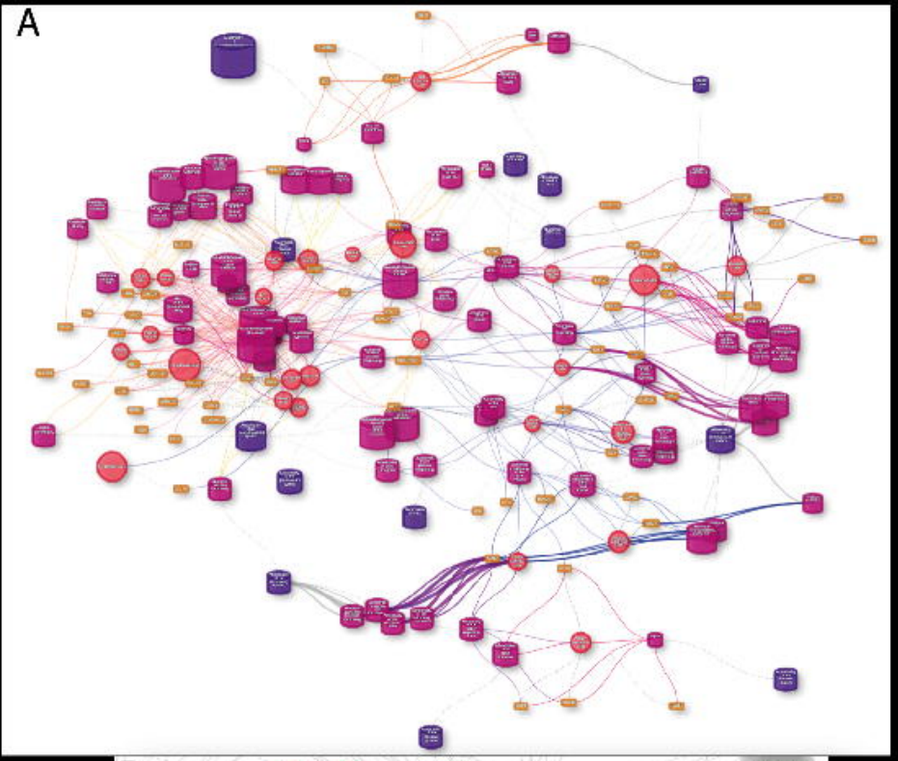
72. Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **31**, 7–15 (1989).

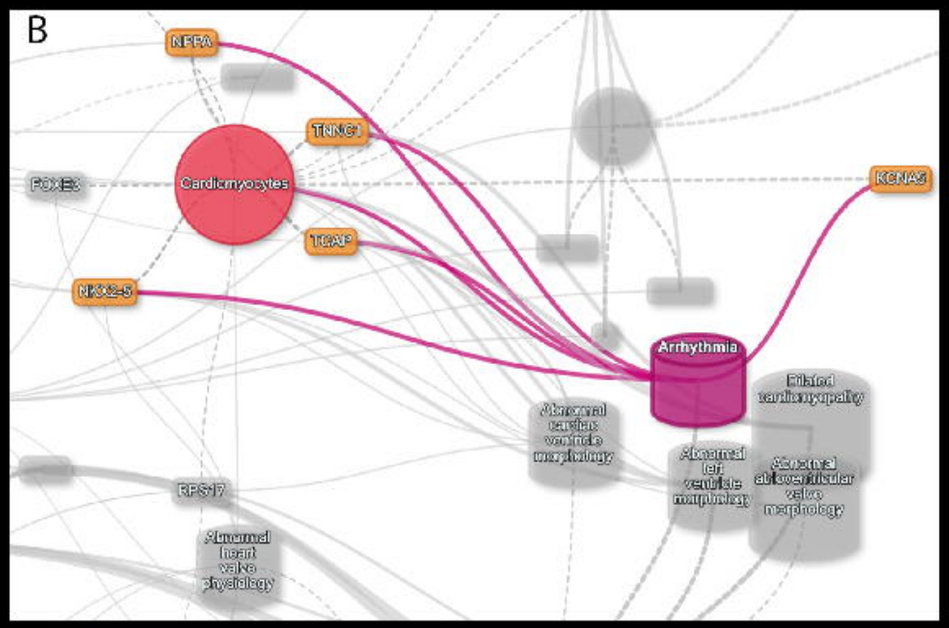

A
B
Number of significant enrichments per cell type
Tissue
Adrenal
Cerebellum
Cerebrum
Eye
Heart
Intestine
Kidney
Liver
Lung
Muscle
Pancreas
Placenta
Spleen
Stomach
Thymus
C
Abnormality of the nervous system
Abnormality of the cardiovascular system
Abnormality of the immune system
Enrichments (n)
Vascular endothelial cells
Stromal cells
Mesangial cells
Smooth muscle cells
Lymphoid cells
Microglia
Myeloid cells
Islet endocrine cells
Acinar cells
Ductal cells
Intestinal epithelial cells
Metanephric cells
Ureteric bud cells
Astrocytes
Oligodendrocytes
Purkinje neurons
SKOR2 NPSR1 positive cells
SLC24A4 PEX5L positive cells
Inhibitory neurons
Limbic system neurons
Ganglion cells
Excitatory neurons
Inhibitory interneurons
Granule neurons
Unipolar brush cells
Photoreceptor cells
Retinal progenitors and Muller glia
Bipolar cells
PDE11A FAM19A2 positive cells
Amacrine cells
Horizontal cells
ENS neurons
Chromaffin cells
Sympathoblasts
Epicardial fat cells
Mesothelial cells
Lymphatic endothelial cells
Endocardial cells
STC2 TLX1 positive cells
ENS glia
Schwann cells
Cardiomyocytes
Satellite cells
Skeletal muscle cells
Erythroblasts
Megakaryocytes
Adrenocortical cells
Hepatoblasts
Squamous epithelial cells
MUC13 DMBT1 positive cells
Goblet cells
Parietal and chief cells
Bronchiolar and alveolar epithelial cells
Ciliated epithelial cells
Neuroendocrine cells
Lens fibre cells
Retinal pigment cells
PAEP MECOM positive cells
Thymic epithelial cells
PDE1C ACSM3 positive cells
SLC26A4 PAEP positive cells
SATB2 LRRC7 positive cells
Visceral neurons
IGFBP1 DKK1 positive cells
AFP ALB positive cells
CLC IL5RA positive cells
ELF3 AGBL2 positive cells
CCL19 CCL21 positive cells
Corneal and conjunctival epithelial cells
Syncytiotrophoblasts and villous cytotrophoblasts
Extravillous trophoblasts
CSH1 CSH2 positive cells
Trophoblast giant cells
Stellate cells
Antigen presenting cells
Hematopoietic stem cells
Thymocytes
Cell type
D
Enrichments in excitatory neurons (%)
Abnormality of the nervous system
E
Enrichments in cardiomyocytes (%)
Abnormality of the cardiovascular system
F
Enrichments in antigen presenting cells (%)
Abnormality of the immune system
−log10(p)

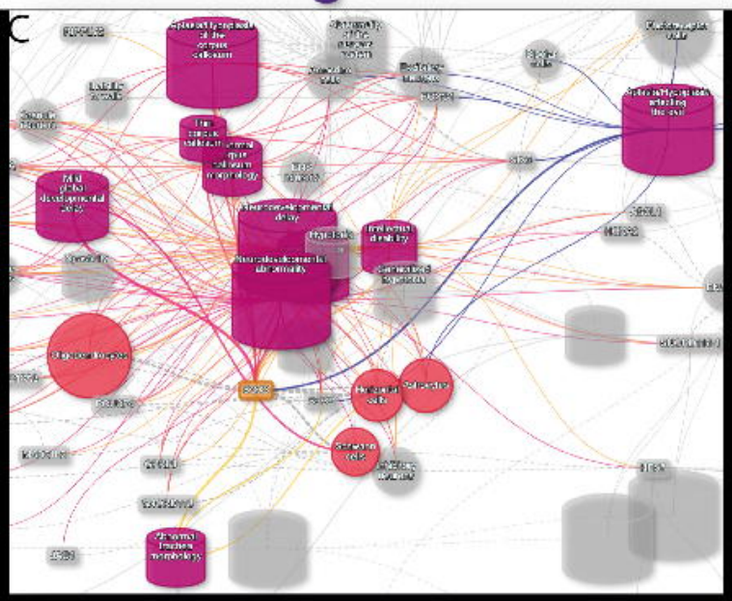

A
Number of associated cell types
0
10
20
0 1 2 3 4 5 6 7 8 9 10 11 12 13
B
log10(fold enrichment)
0.0
0.5
1.0
1.5
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Ontology level
C
Number of associated genes
0
2000
4000
6000
8000
0 1 2 3 4 5 6 7 8 9 10 11 12 13

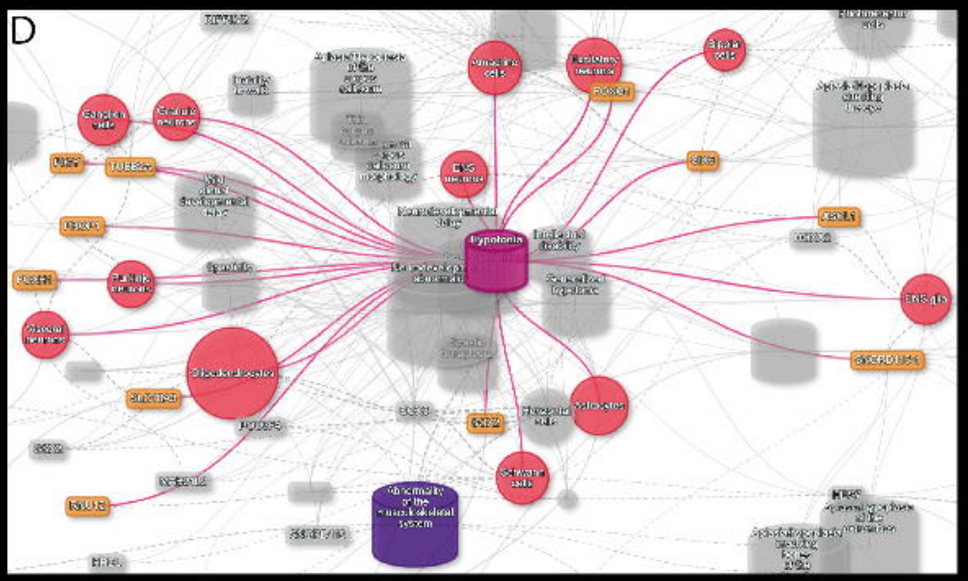

A
Recurrent infections
Phenotypes (n)
19
10
2
2
1
1
9
2
19
2
21
B
Recurrent bacterial infections child terms
Recurrent bacterial skin infections
Recurrent gram−negative bacterial infections
Recurrent mycobacterial infections
Recurrent staphylococcal infections
Fold enrichment
C
Recurrent gram−negative bacterial infections child terms
Recurrent Burkholderia cepacia infections
Recurrent E. coli infections
Recurrent Haemophilus influenzae infections
Recurrent Klebsiella infections
Recurrent Neisserial infections
Recurrent Serratia marcescens infections
Vascular endothelial cells
Stromal cells
Lymphoid cells
Microglia
Myeloid cells
Acinar cells
Ductal cells
Intestinal epithelial cells
Ureteric bud cells
Astrocytes
Horizontal cells
Chromaffin cells
Epicardial fat cells
Mesothelial cells
Lymphatic endothelial cells
Endocardial cells
STC2 TLX1 positive cells
Megakaryocytes
Adrenocortical cells
Hepatoblasts
Squamous epithelial cells
MUC13 DMBT1 positive cells
Goblet cells
Bronchiolar and alveolar epithelial cells
Ciliated epithelial cells
Neuroendocrine cells
PAEP MECOM positive cells
Thymic epithelial cells
PDE1C ACSM3 positive cells
Visceral neurons
IGFBP1 DKK1 positive cells
AFP ALB positive cells
ELF3 AGBL2 positive cells
CCL19 CCL21 positive cells
Corneal and conjunctival epithelial cells
Syncytiotrophoblasts and villous cytotrophoblasts
Extravillous trophoblasts
CSH1 CSH2 positive cells
Trophoblast giant cells
Antigen presenting cells
Hematopoietic stem cells
Thymocytes
Cell type

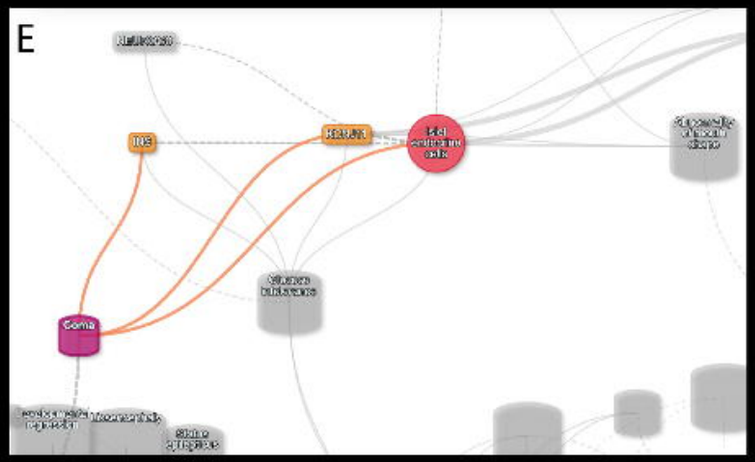

A
% phenotypes
100%
75%
50%
25%
0%
Tier
Tier 1
Tier 2
Tier 3
Tier 4
NA
B
Counts
Phenotypes
Cell types
Diseases
Genes
6173
2832
2832
2832
2770
171
167
132
108
108
104
104
104
78
78
78
78
77
77
77
77
77
60
60
58
31
31
31
31
31
26
26
26
26
12223
12223
12223
12223
12223
12223
19
18
18
18
18
18
17
15
15
15
15
4555
4555
4555
4555
4555
4555
4555
4555
4555
3711
235
235
234
69
69
69
69
Level
All
Association
Cell type
Gene
Phenotype
NA
1. start
2. q_threshold
3. fold_threshold
4. keep_ont_levels
5. keep_onsets
6. keep_tiers
7. severity_threshold
8. pheno_frequency_threshold
9. keep_celltypes
10. keep_seqnames
11. gene_size
12. keep_biotypes
13. gene_frequency_threshold
14. keep_specificity_quantiles
15. keep_mean_exp_quantiles
16. top_n
17. end
Step

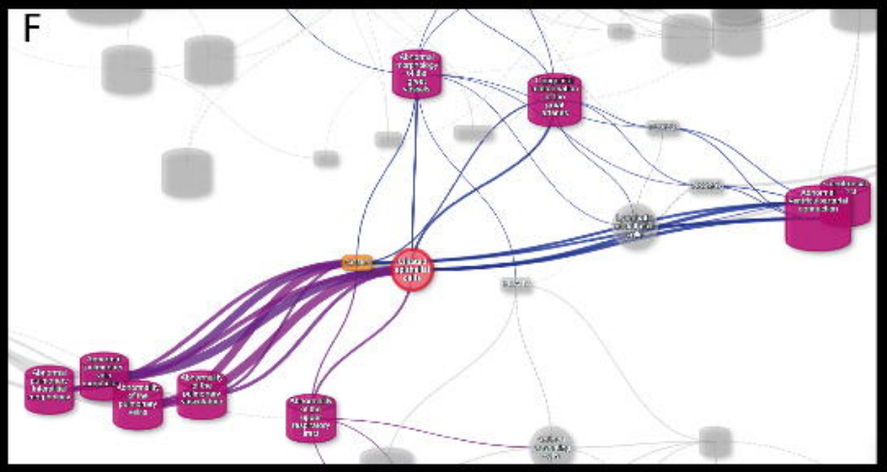

A

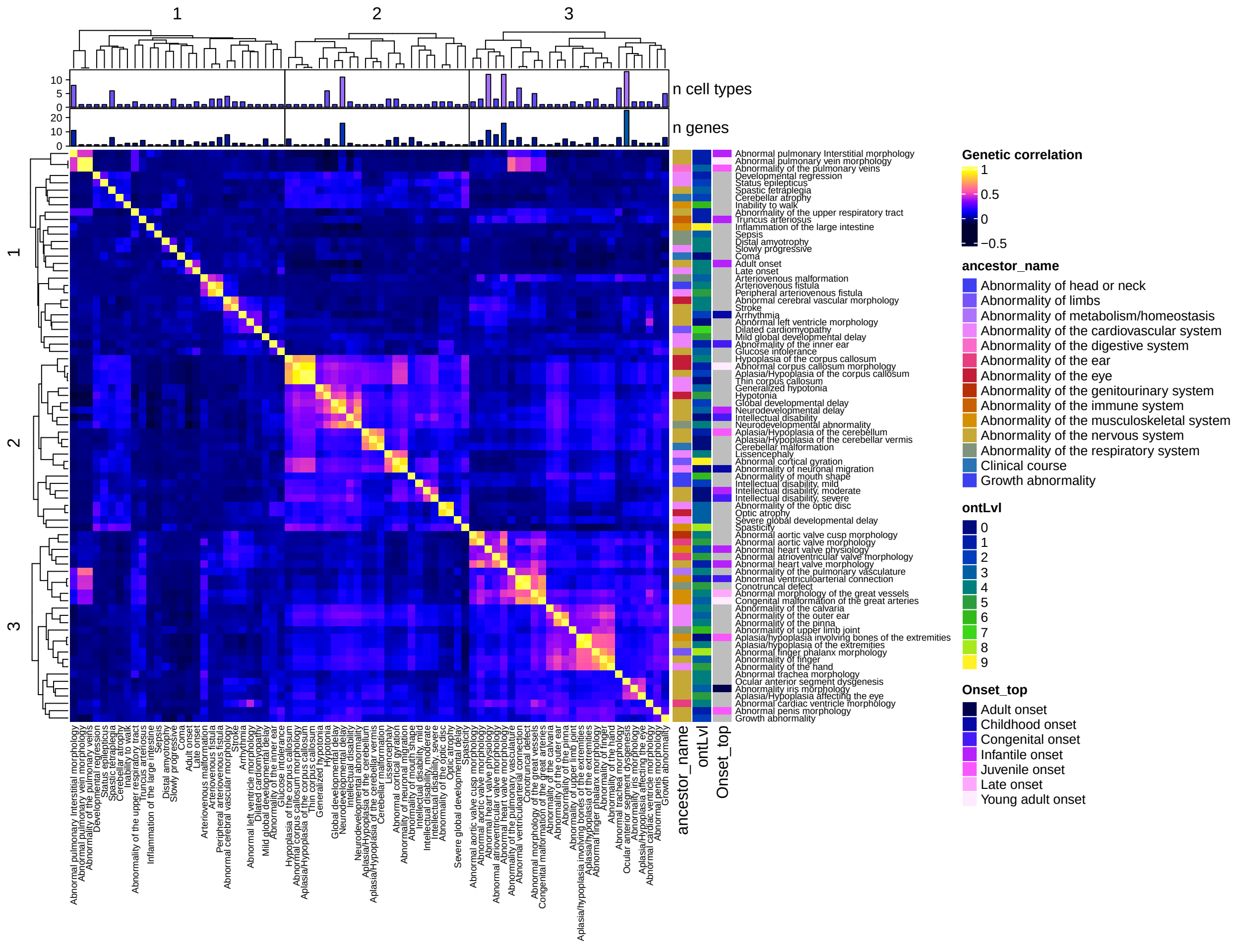

B
NPPA
TNNC1
FOXE3
Cardiomyocytes
KCNA5
TCAP
NKX2-5
Arrhythmia
Dilated cardiomyopathy
Abnormal cardiac ventricle morphology
Abnormal left ventricle morphology
Abnormal atrioventricular valve morphology
RPS17
Abnormal heart valve physiology


C


D
Hypotonia


E
INS
Coma


F


1
2
3
n cell types
n genes
Genetic correlation
1
0.5
0
−0.5
ancestor_name
Abnormality of head or neck
Abnormality of limbs
Abnormality of metabolism/homeostasis
Abnormality of the cardiovascular system
Abnormality of the digestive system
Abnormality of the ear
Abnormality of the eye
Abnormality of the genitourinary system
Abnormality of the immune system
Abnormality of the musculoskeletal system
Abnormality of the nervous system
Abnormality of the respiratory system
Clinical course
Growth abnormality
ontLvl
0
1
2
3
4
5
6
7
8
9
Onset_top
Adult onset
Childhood onset
Congenital onset
Infantile onset
Juvenile onset
Late onset
Young adult onset