Identification of cell types involved in rare disease-associated human phenotypes

Imperial college - Molecular and cellular biosciences project 2
(6500 words)

Robert Gordon-Smith, Molecular and Cellular Biosciences MRes

Dr Nathan Skene, Department of Brain Sciences, Imperial College London

Abstract

Despite the name, rare diseases (RD) contribute to a significant burden of disease globally. The increasing accessibility of genomic and biomedical datasets, and high throughput analytical techniques, has made it possible to uncover new insights into the genetic susceptibility and cell types involved. With the low prevalence of individual RDs, they have often not had the resource allocation they need. Here we attempt to overcome this by tackling the problem as a whole. The Human Phenotype Ontology (HPO) contains disease phenotypes annotated with associated risk genes. Here, these disease-associated gene lists, combined with human single-cell RNA sequence data, were used to identify the primary cell types involved in the HPO disease phenotypes. Expression weighted cell type enrichment (EWCE) was used for the analysis with 100,000 bootstrap reps. Over 8000 significant cell-phenotype associations were found (where $q<0.05$). The results were shown to be overrepresented by expected enrichments. For example, immune cells were more frequently enriched for phenotypes from the "Abnormalities of the immune system" ontology branch than all other cell types combined ($p<0.05$). The analysis was able to reproduce many previously known cell-phenotype relationships documented in the literature. There were also many novel and unexpected results, on examination, many of these were shown to have have a plausible mechanistic explanation. We can be confident that within these results are many previously unknown insights that could provide targets to future research. As too many results were produced to present in a single paper, the Rare Disease EWCE web app was developed to make the analysis available to clinicians and researchers, without the need for specialist knowledge and computational resources (https://ovrhuman.github.io/ewce_website/).

Contents

Abbreviations

Application program interface (API)

Benjamini-Hochberg (BH)

Cell type data file (CTD)

Expression weighted cell type enrichment (EWCE)

False discovery rate (FDR)

Genome wide association study (GWAS)

High powered computing (HPC)

Human phenotype ontology (HPO)

Population-specific expression analysis (PSEA)

Rare disease (RD)

Random-access memory (RAM)

Secure shell connection (SSH)

Single-cell RNA sequencing (scRNA-seq)

t-distributed stochastic neighbor embedding (t-SNE)

Uniform manifold approximation and projection (UMAP)

Introduction

Rare diseases

Although rare diseases (RD) have a prevalence of less than 1 in 2000, they are over 6000 in number (Rath et al., 2012). As a conservative estimate, between 263 and 466 million patients suffer from RDs globally, contributing to a significant disease burden. For individual RDs, economic incentives are often not aligned with the need for research and drug development. However, despite the large number of distinct RDs, approximately 72% are genetic disorders, which may provide opportunities to approach them collectively (Nguengang Wakap et al., 2020). Given the rise of genome-wide association studies (GWAS), electronic healthcare records, and transcriptomic data, it is increasingly feasible to utilise high-throughput computational methods. Not only does increased understanding of RDs benefit patients directly, but RDs can also be used as disease models and they can give valuable insights into complex polygenetic conditions, further extending the impact of RD research beyond what is implied by the term “rare” (Peltonen et al., 2006).

Here we utilise human single cell transcriptomic data from the Descartes human cell atlas, and gene-phenotype relationships from the Human Phenotype Ontology (HPO) to identify cell types associated with the primary disease pathology for 9677 RD phenotypes (Cao et al., 2020; Köhler et al., 2020). Information scarcity and poorly defined case definitions have historically presented significant obstacles to the diagnosis, treatment, and research of RDs. RDs patients often find their diagnosis too late or not at all. Additionally, given the broad scope of this research, it is impossible to present all of the results in a single paper. In light of these points, a crucial part of this project is to develop a publicly available web-based application for retrieval of results relevant to the users field of study.

Human Phenotype Ontology

The Human Phenotype Ontology (HPO) is an initiative founded in 2008 that consists of an ontology of $13\,000$ annotated clinically relevant phenotypes. They are connected in a directed acyclic graph with edges representing transitive "is-a" connections (Köhler et al., 2020). Each phenotype is a subclass of its parent phenotypes, all the way up to the root node of all phenotypes, which is "Phenotypic abnormality." This allows for computationally efficient and human interpretable representation of the complex relationships between phenotypes. For example, Figure 1 shows all ancestor terms for Attention Deficit Hyperactivity Disorder (ADHD), which is a subclass of both hyperactivity and short attention span, which themselves are subclasses of two nervous system disorders, movement and behavioural abnormalities, respectively.

As well as conveying the relationship between nodes (phenotypes), the description logic of bio-ontologies can also describe properties of the node, including synonyms, descriptions, and associated genes and diseases. This data integration allows for implicit knowledge hidden in the data to become explicit and retrievable (Haendel et al., 2018). Currently, much of the HPO focuses on rare Mendelian diseases, many of which have extensive annotations taken from electronic health records and genotype data. The aggregation of data from heterogeneous sources makes the HPO an indispensable tool in phenotype-driven genomic research, precision medicine, and diagnostics.

Descartes Human-Cell atlas

The Descartes human-cell atlas is a single-cell gene expression reference atlas for the human body, created using three-level combinatorial indexing (sci-RNA-seq3) (Cao et al., 2020). It includes 4 million cells representing 15 different organs from 121 human fetal samples. Uniform manifold approximation and projection (UMAP) and Louvain clustering

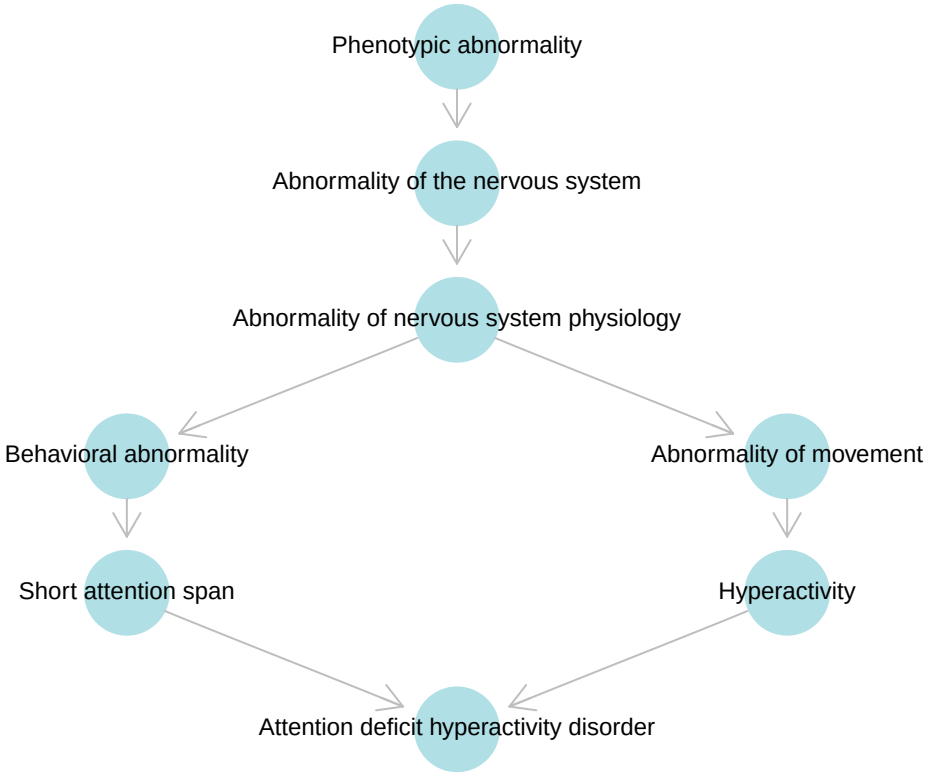

Figure 1: **ADHD ancestor terms.** The ancestor terms of ADHD (HPO Id: HP:0007018) in the HPO. It is 5 generations down from the root node, "Phenotypic abnormality." The arrows represent "is-a" relationships, meaning that a child term is a sub-class of the term above it. This was plotted using R version 4.05 and the ontolgyX package (Greene et al., 2017; R Core Team, 2021).

were applied to the transcriptomic data. After accounting for common annotations across tissues, it clustered into 77 main cell types. Additionally, the Tabula Muris whole-body mouse scRNA dataset was analysed, which clustered into 38 distinct cell types (The Tabula Muris Consortium et al., 2018). The Descartes results are the primary focus of this study.

Expression weighted cell type enrichment (EWCE)

EWCE is a technique developed by Skene and Grant (2016) that utilises single-cell gene expression data to determine whether a set of genes is significantly associated with a particular cell type with respect to expression. Unlike other enrichment analysis techniques, like population-specific expression analysis (PSEA), developed by Kuhn et al. (2011), EWCE does not rely on quantitative gene expression data from disease tissue samples or predefined cell marker genes. Instead, EWCE takes a set of genes found to be associated with a phenotype or disease and, using scRNA-seq data, it identifies cells that express that gene set more than could be expected by chance from a random gene set of the same length.

Using scRNA-seq data allows for the use of techniques such as UMAP and t-distributed stochastic neighbour embedding (t-SNE) to classify the cell types involved. This allows EWCE to identify cell sub-types at a higher resolution than is possible with the predefined marker genes used in PSEA. Another advantage of EWCE is that the gene lists can be obtained from genome-wide association studies (GWAS) and therefore are involved in the primary genetic susceptibility. Conversely, PSEA results can be confounded by secondary "reactive" expression. For example, in PSEA you may find that inflammatory cytokines are overexpressed in disease tissue sample due to immune activation. It would then be unclear if this is part of the primary pathomechanism or is a protective effect, brought about by an appropriate immune response to the condition. Additionally, the lack of need

for quantitative expression data from disease samples makes it applicable to many more publicly available data sets.

In this study, all 9677 RD associated gene lists were taken from the HPO and, using the human whole-body scRNA-seq data from Descartes, EWCE was run to identify the key cell types involved in the phenotypes. This task would not have been feasible with other methods and without these large publicly available data sets. There are multiple objectives to this project. The first is to create a program, written in R, to pull data from the HPO and the Descartes Human cell atlas, clean and format the data, and then run EWCE to identify significantly enriched cell types for the RD phenotypes (R Core Team, 2021). A secondary objective is to make the code written for this project available for future research and follow up studies by storing it on the Neurogenomics lab GitHub, and submitting modifications to the EWCE R package that is published on Bioconductor, an open-source repository of software for bioinformatics. It is common practice to submit improvements in open source software development and helps to drive the improvement of future versions. Finally, a web-based application must be developed for exploring and retrieving the results as there are too many to be presented in one paper. This will allow domain experts to view results most relevant to their field, which will bring gene set enrichment analysis to a wider audience and enable the full impact of the study to be realised.

Materials and methods

EWCE overview

The EWCE R package from Bioconductor (version 1.1.0) is used for the analysis (Skene, 2021). To briefly summarise the technique, as described by Skene and Grant (2016), EWCE takes a target gene list of length n, referred to as T, and a set of background genes referred to as B. A scRNA data set is then used to calculate a gene-cell specificity score for each gene in every cell type. Simply put, this is obtained by dividing the expression of a gene within a cell type by the total expression of the gene in all cell types.

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

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

Where $e_{g,c}$ is the specific expression of gene g in cell type c. N_c is the number of c, and $r_{g,i}$ is expression of g in cell i (indexed from c). We can then sum the specificity scores of the genes in T to get its total expression specificity score in a given cell (γ). This is done for all cells, enabling us to quantify the level of specific expression of gene list T (indexed by X) in each cell type (c).

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

Bootstrapping is then used to determine if the level of expression of T in a given cell is statistically significant, in other words, the probability of enrichment. To do this, $100\,000$ random gene sets of length n are then randomly sampled without replacement from B.

The same cell-specific expression calculation, as described above, is then calculated for all $100\,000$ random gene sets in each cell type "c." This gives a probability distribution of cell-specific expression for gene sets of length n in any given c. The mean and standard deviation of this distribution is normalised to 0 and 1 respectively and is used to calculate a Z score. We can then determine the probability of enrichment of T in c based on the number of bootstrap gene lists that have a higher cell type specific expression than T.

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

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. The target gene sets used here are obtained from the HPO. Each gene set is known to be associated with a rare disease phenotype. If one of these gene sets is significantly enriched in a cell type, then it is likely that the cell type plays a role in the pathology, possibly making them valuable targets for future research.

Gene-cell specificity scores, rather than absolute expression, is used because it allows us to weight the expression of a gene by how commonly it is expressed in other cell types. In other words, genes that are only expressed highly in a few cell types are weighted higher and contribute more to the probability of enrichment than common genes. However, we must exclude genes with very low expression (mean < 0.2 in all cell types), as a small number of reads of the gene could make it appear highly specific. The Descartes human cell atlas scRNA dataset was used. It contained 377456 cell samples that clustered into 77 distinct cell types. The RD phenotype-associated gene lists were obtained from the HPO. There are 9677 phenotypes in the HPO, however, phenotypes with less than 4 associated genes were excluded from analysis, leaving 6173 gene lists. The fold change of

enrichment is calculated by dividing the specific expression of T by the mean expression of the bootstrapped random gene sets.

The EWCE package has a function for visualising results (`ewce_plot`) which returns a plot of the fold change of expression for a gene set in each cell type. It also gives a dendrogram of the clustering of the cell types in the plot. The dendrogram must then be manually aligned with the bar chart using image editing software in the current version. Given the many thousands of results that need to be visualised for this study, manual alignment of the plots would not be possible. A modified version of the `ewce_plot` function was created to enable automatic alignment of the dendrogram to the bar chart. This solution was then submitted to the EWCE developers and the modified `ewce_plot` function has now been published on Bioconductor.org in the latest version of the EWCE package.

Multiple testing adjustment

Due to the high multiple testing burden of this analysis, the Benjamini-Hochberg (BH) method was used to limit the false discovery rate (FDR). Traditional family-wise error rate based methods, like Bonferroni corrections, are too conservative when performing these types of high-throughput analyses, resulting in the rejection of many positive results. Instead, BH corrects the proportion of significant findings down to an acceptable FDR (Benjamini and Hochberg, 1995). The p-values that have been adjusted for FDR are referred to as q-values.

Analysis Makefile

This analysis involved multiple steps, including pulling the data from online resources, cleaning and formatting it and running the analysis. A makefile (`Makefile.R`) was created to automate the whole process. This RD EWCE analysis makefile was then pushed

to the UKDRI Neurogenomics lab GitHub, a remote host for software development and version control. This allows it to be reused in follow up analysis, when more phenotypic data becomes available, as well as to be developed further for future studies by other researchers.

First, the makefile calls a function, `gen_ctd`, which is stored in the `source` directory. This downloads the latest scRNA-seq data set from Descartes and produces the cell type data (CTD) file by calculating specificity scores for each gene and cell in the scRNA data set. The CTD is structured as a list of matrices of normalised specificity scores, with columns representing the cell type and rows representing gene. The matrices are generated at various levels of cell-clustering resolution, so that the user can decide how specific they wish to be in distinguishing between cell sub-types. In this case, the levels generated were Level 1 (Cell lineage), Level 2 (Organ Cell lineage), Level 3 (Cell lineage Developmental day), and Level 4 (Organ Cell lineage Developmental day). The level 1 data was used as it provides sufficient resolution without over-complicating the results or increasing the multiple testing burden unnecessarily. The `gen_results` function is then called, which pulls the phenotype gene lists from the HPO website and runs EWCE on each, using the CTD file gene-cell specificity scores.

In order to run the EWCE analysis with $100\,000$ bootstraps on 9677 phenotypes, significant computational resources were required. To make this possible, the program was written to support parallel processing. A secure shell connection (SSH) was then established with the Imperial College High Performance computing (HPC) service; and an 8 CPU machine with 96 GB of random access memory (RAM) was requested for an interactive session using the following command:

```
qsub -I -l select=01:ncpus=8:mem=96gb -l walltime=08:00:00
```

The Rare Disease EWCE R project was then cloned from the GitHub repository onto the HPC. There is a time limit of 8 hours per HPC job, which was not sufficient to complete the

whole analysis. To get around this, the individual results generated for each phenotype were output as separate `.RDS` files. Another script (`completed_phenotypes.R`) was created to identify which phenotypes had already been analysed by reading the file names in the output directory. It then removes the completed phenotypes from the input data set, making it possible to continue the analysis from where it left off after the previous job was terminated. As this had to be repeated many times, a shell script was created. It loads the Anaconda development environment, and then executes the the `completed_phenotype.R` and `Makefile.R` scripts ("Anaconda software distribution," 2020). The shell script was placed in a `while` loop so that, in cases where the analysis prematurely terminated due to error, it would restart automatically (see Code example 1).

```
#!/bin/bash
n=1
while (($n <= 20))
do
  echo "Running script (attempt = $n)"
  n=$(( n+1 ))
  module load anaconda3/personal
  source activate pyre
  Rscript completed_phenotypes.R
  Rscript Makefile.R
done
```

Code example 1: **Rare Disease EWCE HPC shell script.** This script loads the development environment and executes the code to run the analysis. It is placed in a while loop so that it automatically restarts if it terminates prematurely.

The analysis ran at a rate of approximately 90 phenotypes per hour and took a little over 100 hours to complete. A terminal multiplexer (Tmux) was used so that the session would continue to run in the event of disconnection.

Rare disease web application

A web-based application was made so that the study can have its maximum impact. It is important for researchers and clinicians to be able to retrieve relevant results without the need for substantial computational resources and specialist knowledge of enrichment analysis. As the results give cell-phenotype relationships, they need to be retrievable by both fields (cell type and phenotype), and it must also be possible to subset them by significance and effect size (q-value and fold change). Due to the heavy computation required to generate the figures and subset the results, the cell selection and phenotype selection features were developed as separate web applications to reduce loading times. The apps were created using the Shiny Web application Framework for R and deployed on the ShinyApps server (Chang et al., 2021). A landing page was created using HTML and CSS to provide links to the apps, data sources, and code repositories; as well as to give general information about the project (see Figure 2).

A demonstration of the interactive plot feature from the cell-select app can be seen in Figure 3 below. The ggnetowrk R package was used to give x and y coordinates to each phenotype in the network and the plotting was done using the ggplot2 R package (Briatte, 2020; Wickham, 2016). Mapping of the ontology level of the phenotypes to the size of the node was done so that phenotypes higher up the HPO hierarchy appear larger than those below (within a connected component of the graph). This is done by checking if a phenotypes parent node is present in the component to be plotted. If the parent is not present, the function returns zero, if a parent node is present, the function returns $1 + f(parent)$. In other words, the function then calls its self recursively to check if the parent phenotype also has a parent phenotype present. This happens repeatedly until the last parent is found and the result can be returned (see Code example 2).

The box on the interactive plot, seen in Figure 3, shows information about each phenotype when the mouse hovers above it. It includes the results for the selected phenotype, as

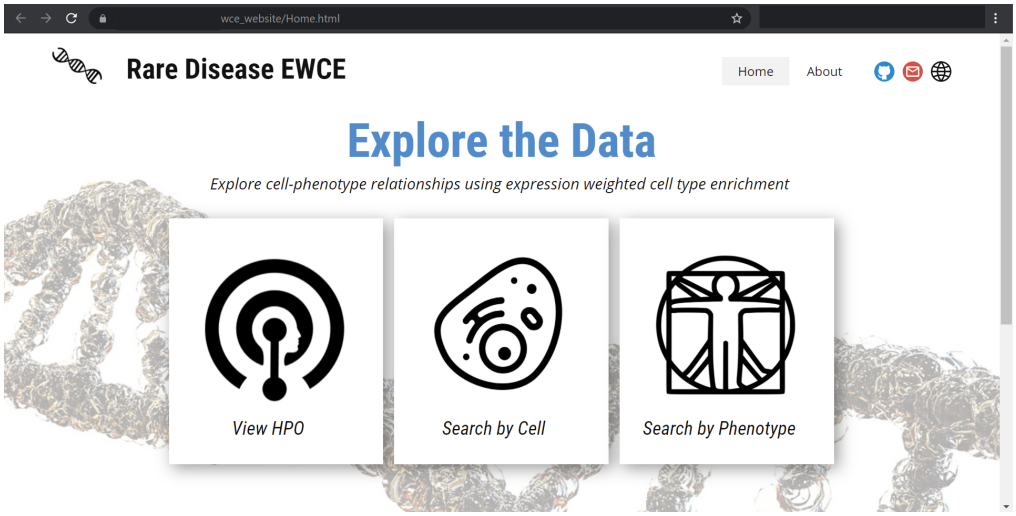

Figure 2: **Rare disease EWCE homepage.** The website homepage has links to the Cell sellect and phenotype select apps, as well as to the HPO. The about page contains further information on sources and relevant links. It was created using the markup languages, HTML and CSS. More applications and features for the website are currently in production.

well as a disease description pulled directly from the HPO via an application program interface (API). This was done for all terms. New lines ("\n") were added to the disease descriptions using a custom built function to prevent the box getting too wide when the description is long. A `.RDS` file containing the hover box information was stored on the ShinyApps server.

```
find_parent <- function (phenotype,phenoAdj,hpo){
  pos_parents = hpo$parents[phenotype]
  phenotypes = rownames(phenoAdj)
  paths = list()
  for (p in phenotypes){
    if (phenoAdj[p,phenotype] == 1) {
      if (p %in% pos_parents) {
        paths[p] = 1 + find_parent(p,phenoAdj,hpo) # <- recursion
      }}}
  if (length(paths) == 0) {
    return (0)
  } else {
    parents = 0
    for (i in seq(length(paths))) {
      if (paths[[i]] > parents) {
        parents = paths[[i]]
      }}}
  return (parents)}
```

Code example 2: **Find number of generations of ancestors for an HPO term.** This function finds the number of generations of ancestor terms present for each term in a connected component of a subset of the results. This is used to map the relative ontology level of terms to node size in the interactive plot. `phenoAdj` is an adjacency matrix of all the phenotypes to be plotted where 0 and 1 is used to indicate if `phenotype[i]` is a parent of `phenotype[j]`. `hpo` is the HPO ontology data object.

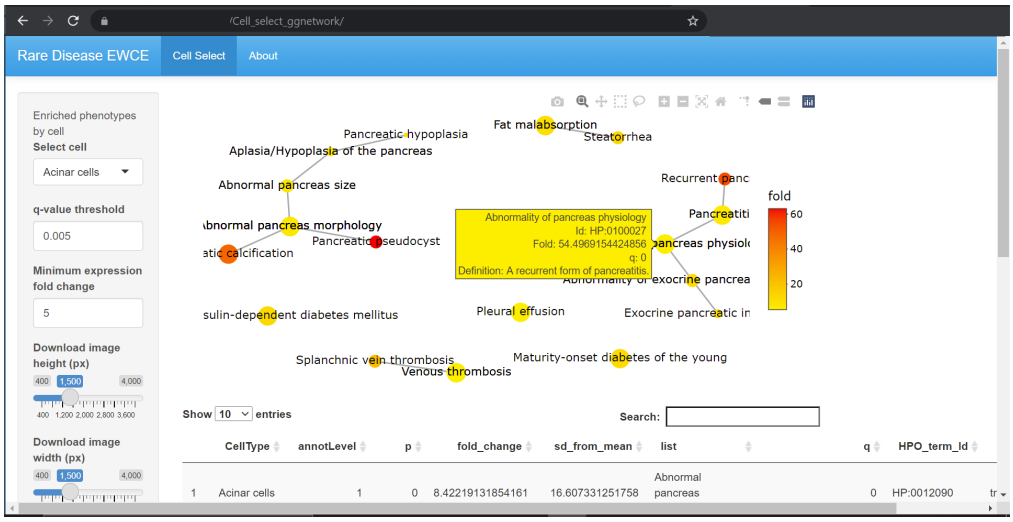

Figure 3: **Interactive web app for exploring significant phenotype enrichments for a given cell.** This is the interactive RD EWCE app for searching for significant phenotype enrichments within a selected cell type. The figure is generated using a combination of bespoke functions and the ggplot2 and ggnetwork packages (Briatte, 2020; Wickham, 2016). The size of the nodes represents the direction of the "is_a" relationship between nodes, where parent terms are larger than their child terms. Expression fold change is represented by the colour. In this example, the user has selected Acinar cells, with a significance threshold of q<0.005 and a fold change threshold of > 5. This is designed for dynamically interacting with the data to find interesting results. The hover box was created by connecting to the HPO website API and it provides a description of the phenotype when the cursor hovers over it.

The interactive plot works well for exploring the data. However, it is less clear when printed as a static image. A second, more print-friendly version, was also created and users may download the figure from the app. This was created using the ontologyX R package `onto_plot` function, modified to allow a heat map of fold change. An example of the figures that can be created is shown in Figure 4. Again, as in Figure 3, the user has selected acinar cells, q<0.005 and fold change > 5 (Greene et al., 2017). The heat

map represents fold change in specific expression of the gene list (compared to the mean expression of bootstrapped gene lists of the same length). Note, it can be seen that higher fold changes (red) are typically found towards the leaf nodes, where phenotypes are more specific (this will be covered in more detail later, see Figure 9 A).

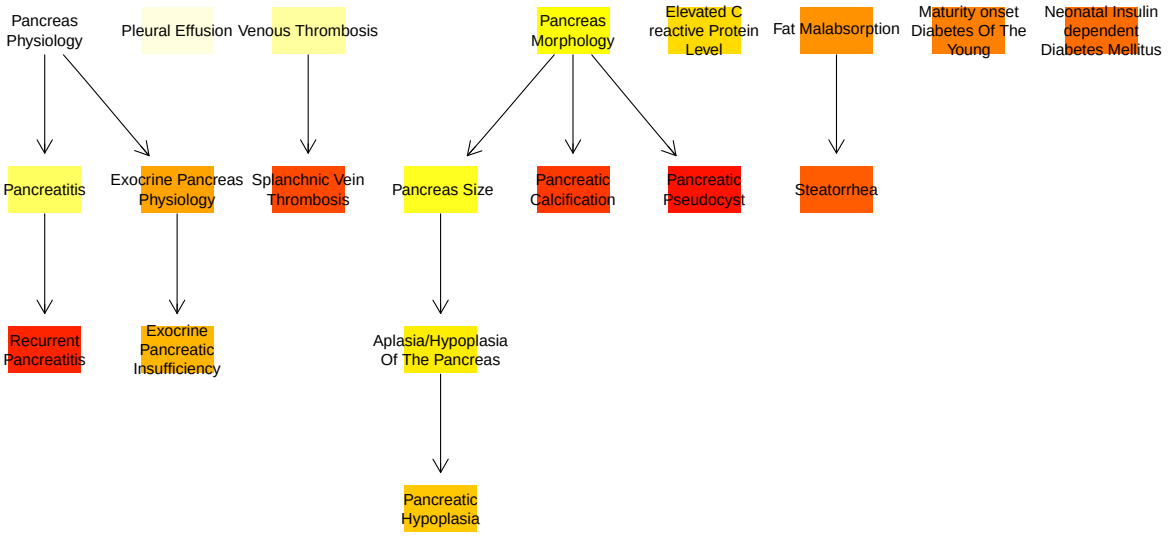

Figure 4: **Print-friendly version of the cell select app.** This is a demonstration of the figures produced by the print-friendly version of the cell select app, created using the modified functions from the ontolgyX R package (Greene et al., 2017). The same user settings have been selected as in the interactive plot demonstration (Acinar cells where q<0.05 and fold change > 5). This version of the figure is more well suited to printing as it represents the HPO hierarchy vertically and has less overlapping text.

Results

The Descartes human scRNA data clustered into 77 unique cell types. There were 8379 significant cell-phenotype associations across all cells (q<0.05). The number of significant phenotype associations for each individual cell type can be seen in Figure 5.

The greatest number of significant phenotypes enriched for a single cell type was 338, seen in Skeletal muscle cells, (where q<0.05 and fold change > 1).

Expected cell types are associated with the main HPO branches

The root HPO term, phenotypic abnormality (HP:0000118), has 23 child phenotypes that represent the main classes of phenotypic abnormality in the HPO. These 23 branches can be seen in Figure 6. Abnormality of the nervous system, abnormality of the cardiovascular system, and abnormality of the immune system have been highlighted, as these branches have clear cell types in which there is expected to be high numbers of enrichments (neurons, cardiomyocytes, and immune cells, respectively). These expected associations will be used as a means of validating the results.

As can be seen in Figure 7, the phenotypes from these branches tend to have a greater number of significant enrichments in the expected cell types. A hypergeometric test was used to determine if the number of significantly enriched phenotypes for a cell type within a branch is significantly elevated. BH correction was used to account for multiple comparisons. A significant result from the hypergeometric test suggests that the branch as a whole is significantly associated with those cell types (the hypergeometric test results, seen in Fig. 7, are denoted by ****, *** **, and * which indicate that q<0.00001, 0.0001, 0.001, and 0.05, respectively). The cells are ordered by dendrogram grouping, so it can be seen that the large cluster of significant results in the “Abnormality of the nervous system” branch, ranging from astrocytes to sympathoblasts, are cells of the nervous system.

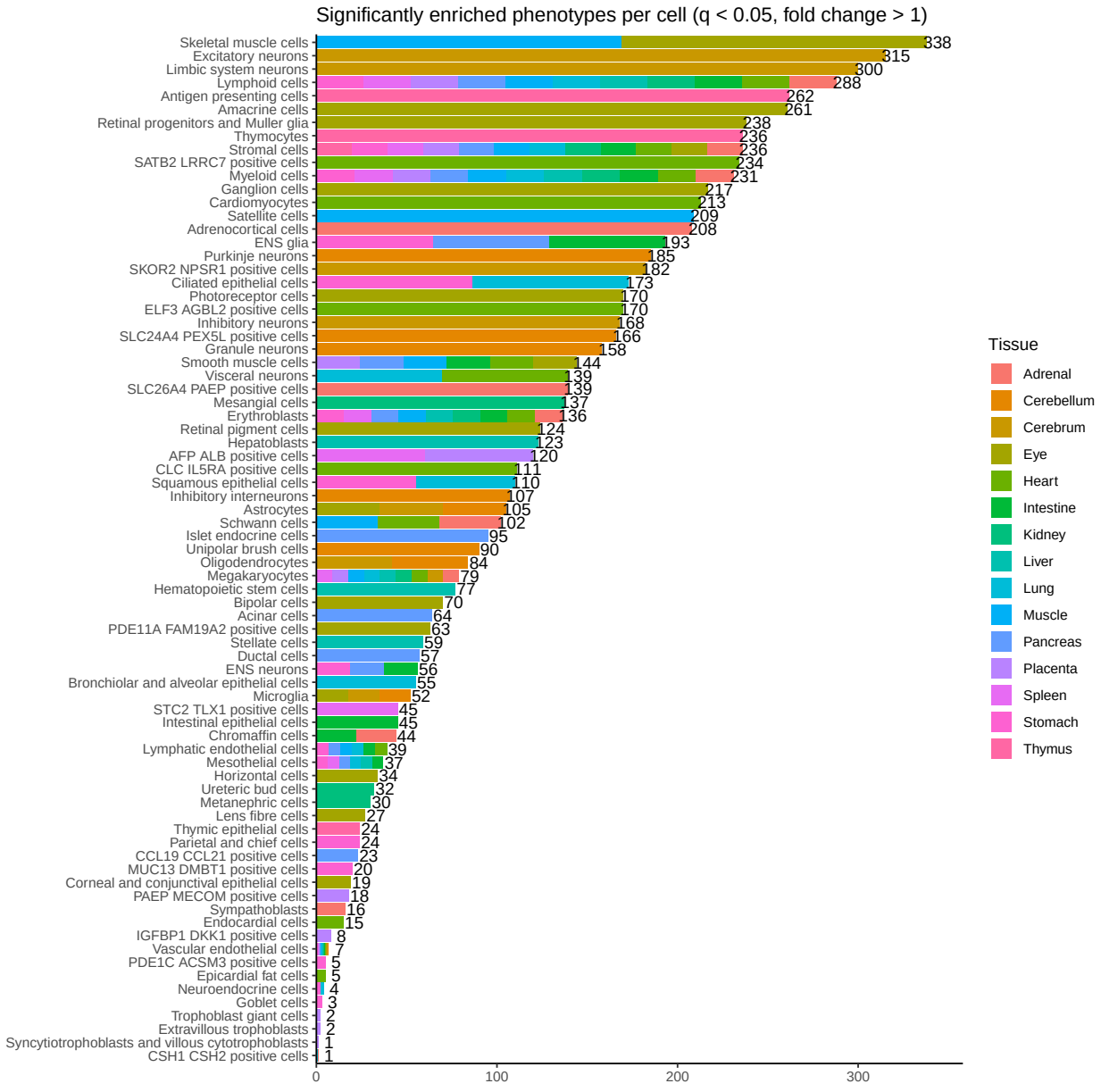

Figure 5: **Number of significant enrichments per cell.** This figure shows how many significant phenotype enrichments were found for each cell type, where q<0.05 and fold change > 1.

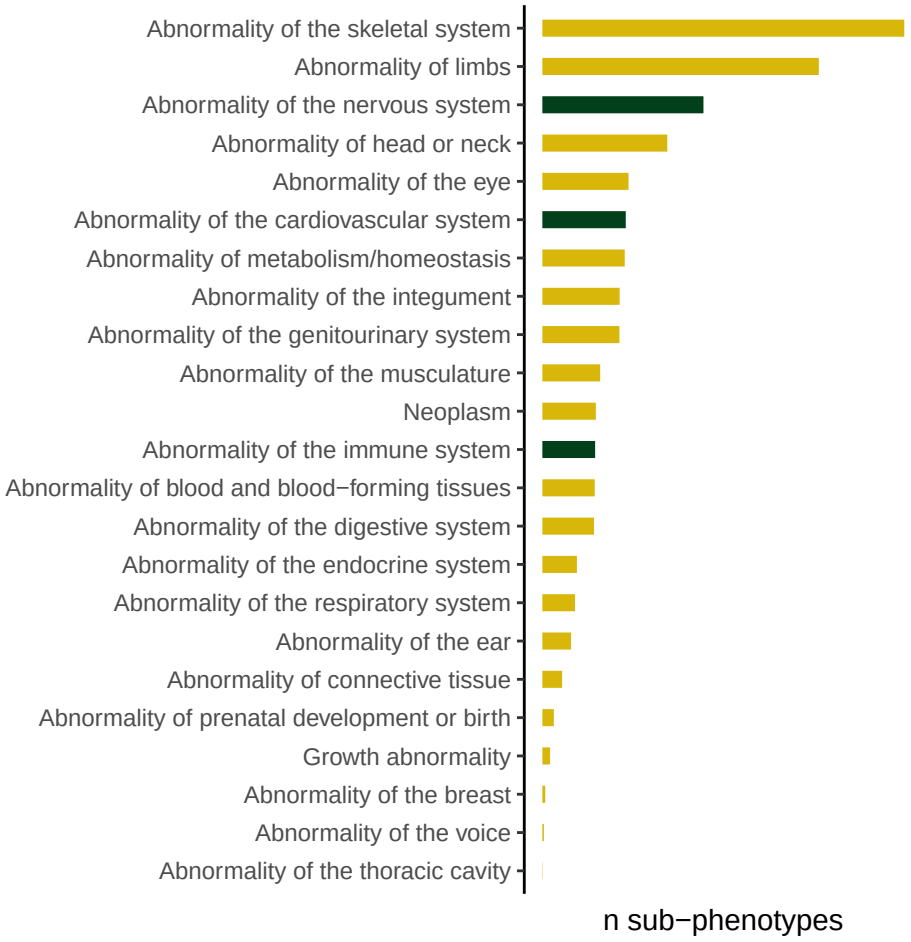

Figure 6: **Main branches of the HPO.** This shows all the child phenotypes under "Phenotypic abnormality" (ID: HP:0000118) in the HPO. These can be considered as the main classes of phenotypic abnormality, or the main branches of the HPO.

The largest number of disease phenotypes in the nervous system branch are associated with limbic system neurons. In the immune system branch, there were a large number of significant enrichments in immune cells such as lymphoid cells, myeloid cells and antigen presenting cells. Finally, the cell type most significantly associated with abnormality of the cardiovascular system is cardiomyocytes (q<0.00001). These findings are in agreement with the predicted cell-phenotype relationships.

Additionally, there are many novel results that may provoke new lines of research. For example, a significant number of sub-phenotypes of "Abnormality of the cardiovascular system" are enriched in hepatoblasts.

To further demonstrate that the analysis finds expected phenotype-cell relationships, excitatory neurons, cardiomyocytes, and antigen presenting cells were again used. Figure 8 shows that if we take all results for a particular cell, the more significantly associated phenotypes disproportionately come from the expected branch of the HPO.

In other words, Figure 8 shows that as more stringent significance thresholds are used, the proportion of enriched phenotypes from the expected HPO branch increases (r_{19}=0.7877836, p=2.2383734×10^{-5}).

Low ontology level terms have high specific expression in enriched cells

It was then hypothesised that significant enrichments would have higher specific expression towards the leaf nodes of the HPO, compared to terms closer to the root. We quantify specific expression by fold change. This denotes how many times greater the specific expression is in a gene list, relative to the mean expression of random bootstrapped gene lists of the same length, in a given cell type (see Figure 9 A). In other words, the strength of the enrichment. Ontology level represents the position of a term in the HPO heirar-

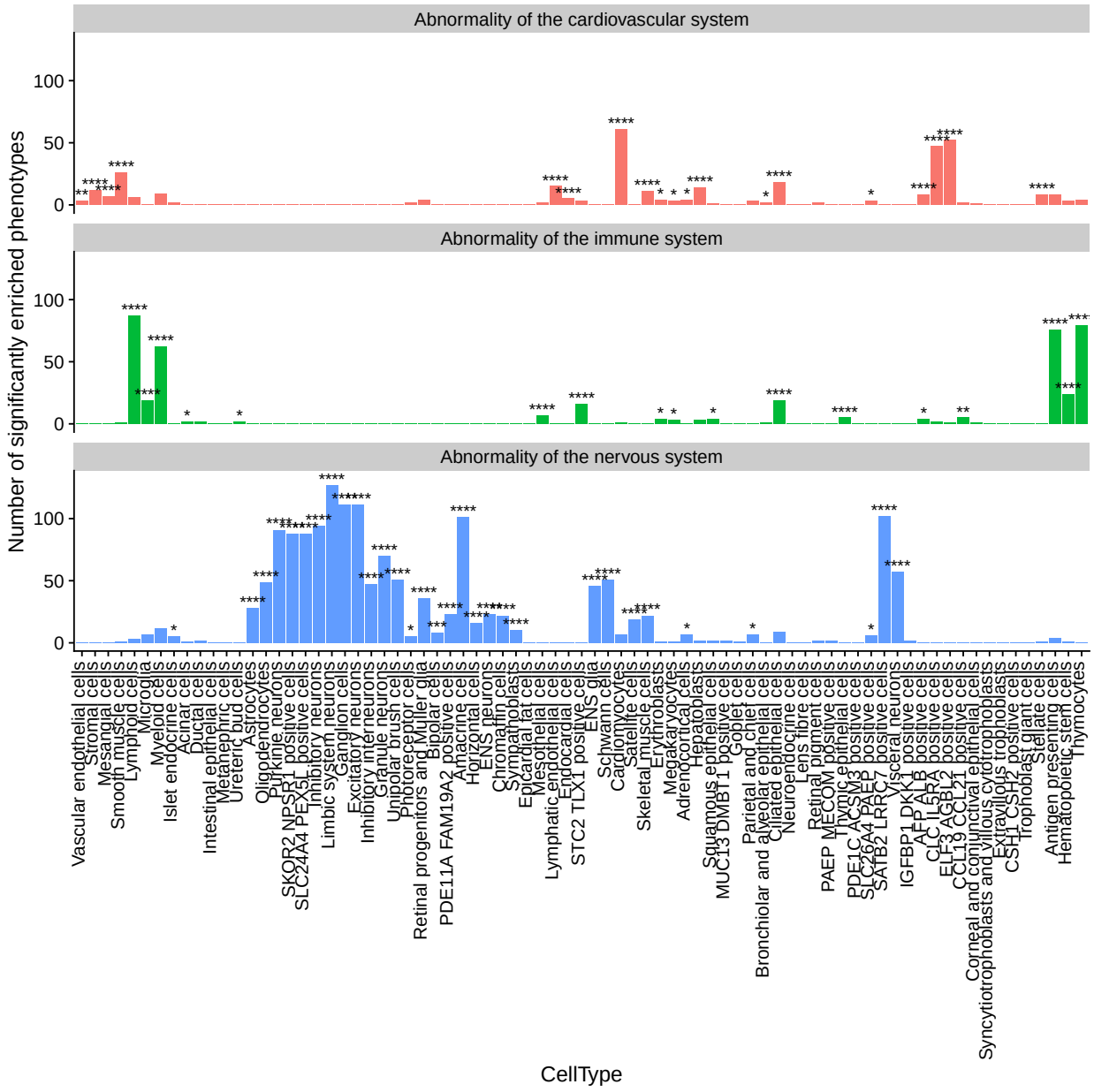

Figure 7: **The number of significant enrichments per cell in three main branches of the HPO.** It can be seen that the greatest number of significant enrichments are seen in the cell types that we would expect for that branch. An additional hypergeometric test was done to show where the number of significant enrichments in a branch is significantly elevated, significant cases are highlighted with asterics ****, *** **, and * which indicate that q<0.00001, 0.0001, 0.001, and 0.05, respectively. The cells are ordered by their dendrogram grouping

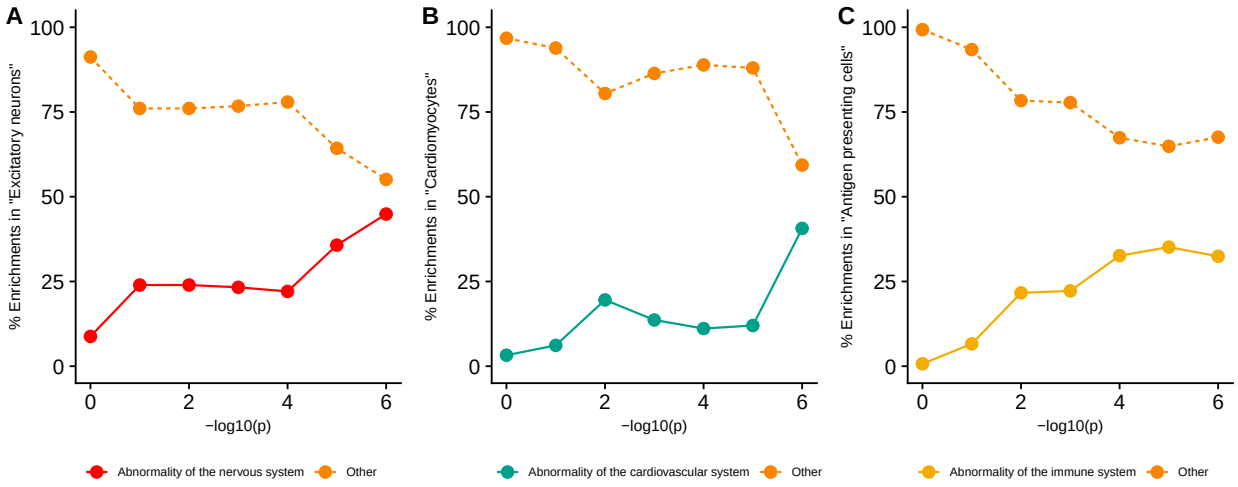

Figure 8: **The relationship between significance threshold and the proportion of enrichments found in the expected HPO branch.** As more stringent significance thresholds for enrichment in a particular cell type are used, the proportion of enrichments from an expected HPO branch increases. This validates the method as it shows that many of the strong phenotype-cell associations are in agreement with our current understanding. **A** shows the proportion of enrichments for abnormality of the nervous system descendants in excitatory neurons as significance threshold is increased, **B** looks at the cardiovascular system branch and cardiomyocytes, and **C** looks at the immune system branch and antigen presenting cells.

chy. It denotes, how many generations of sub-phenotypes (descendants) a term has. For example, an ontology level 4 term like “Recurrent infections” has 4 generations of descendant terms below it, including “Recurrent gram-negative bacterial infections,” which in turn has 2 generations of descendants, so it is at ontology level 2. The root term of the HPO (Phenotypic abnormality) is at ontology level 13.

The relationship between ontology level and fold change shows that the phenotypes at higher ontology levels, which encompass a broad range of traits, generally a have lower specific expression. Conversely, the significant associations at lower HPO levels tend to have higher specific expression of the phenotype gene list in the associated cell type. Figure 9 B shows that higher ontology level phenotypes also tend to have longer gene lists. Conversely, the more narrowly defined phenotypes at lower ontology levels are generally associated with only a few genes. This further supports the idea that the gene lists at lower levels tend to be more specific.

Low ontology level terms are enriched in expected and novel cell types

While figures 7 and 8 show that high-level HPO terms are generally associated with expected cell types, figure 10 shows that expected cell-phenotype relationships are also found at the more specific phenotypes at lower HPO levels. As an example, we look at all descendant terms of Recurrent infections (HP:0002719), which includes 72 HPO terms at ontology levels ranging from 0 to 3. The hypothesis is that they will be primarily enriched in cells involved in the immune system. It was predicted that, from these sub-phenotypes, the number of significant enrichments in immune system associated cell types would be greater than the number of significant results in all other cell types combined.

As predicted, the majority of enrichments were found in cells associated with the immune system (lymphoid cells, myeloid cells, antigen presenting cells, thymocytes, hematopoietic

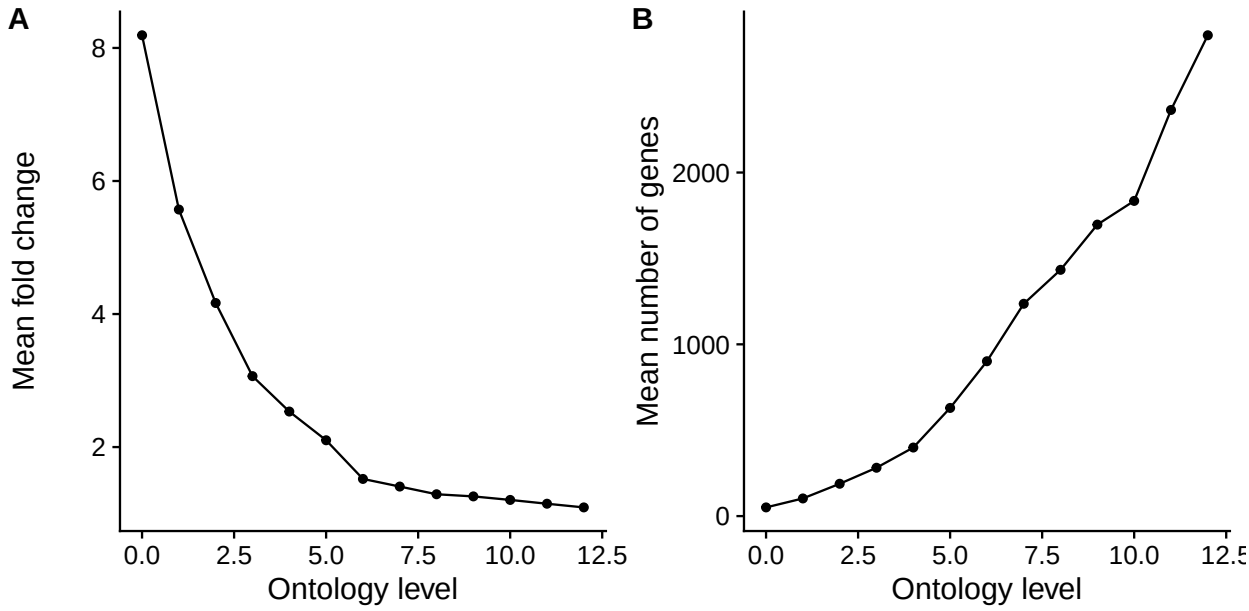

Figure 9: **Relationship between ontology level and expression specificity. A:** More narrowly defined phenotypes, at lower HPO levels, show more specific expression of their gene lists in significantly associated cells than phenotypes at higher HPO levels. Ontology level represents how many generations of sub-phenotypes a term has in the HPO. As we progress out towards these leaf-nodes (ontology level 0), the specific expression of the gene list tends to be higher in significantly enriched cell types. The level of specific expression is represented as fold change here. **B:** At higher ontology levels, the phenotypes encompass a larger number of traits and are associated with larger lists of genes. The lower ontology level phenotypes are more narrowly defined and are often associated with only a few genes, possibly making them good targets for therapy and research.

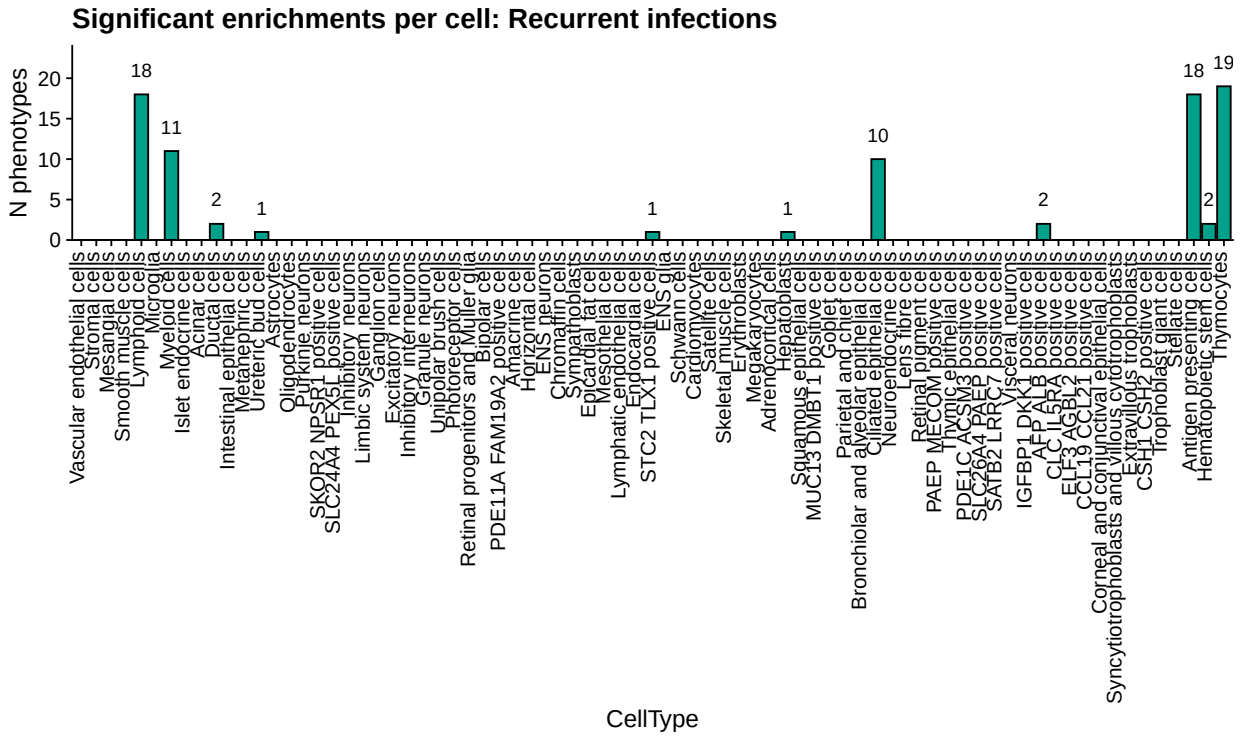

Figure 10: **Significant enrichments per cell for all descendant terms of Recurrent infections.** This shows the number of significantly enriched phenotypes per cell for descendant terms of “Recurrent infections.” As expected, the majority of enrichments are found in cells involved in the immune system, including antigen presenting cells, hematopoietic cells, yhymocytes, myeloid cells, and lymphoid cells. These collectively account for 68 significant enrichments. We also find novel enrichments in ciliated epithelial cells, ductal cells, ureteric bud cells, hepatoblasts, and some undefined cells. These may warrant further investigation.

stem cells). 68 enrichments were found in immune system related cells, and only 17 enrichments were found in other cell types ($t_{4.0160216}$=4.1279362, p=0.014399).

A selection of the sub-phenotypes of "Recurrent infections" were then isolated, to see where the significant effects lie, and to investigate some of the unexpected and novel associations. Figure 11 shows the adjacent child phenotypes of "Recurrent bacterial infections," which is its self a child term of "Recurrent infections."

The analysis was able to reproduce a well known association with myeloid cells, a class of phagocytic immune cell often involved in defence against bacterial pathogens. Interestingly, a very strong association is seen between recurrent mycobacterial infection and ciliated epithelial cells (q<0.0001, fold change=18.52).

Figure 12 provides a closer look at the child phenotypes of Recurrent gram-negative bacterial infections. Recurrent neisserial infections were found to be enriched in Hepatoblasts (q=0.0125167, fold change=9.9). As this is a potentially novel finding, this phenotype was also checked in the Tabula Muris data. Similarly, significant enrichment in hepatic cells (Kupfer cells and Hepatocytes) were found (q<0.0001).

To further validate the ability to find enrichments for specific RD phenotypes, an example from the abnormality of the nervous system branch was also examined (Impaired social interactions). This can be seen in figure 13 (this figure also demonstrates the modified version of the `ewce_plot` function that was published in the EWCE Bioconductor package).

The child terms of impaired social interactions were then examined to get a higher resolution on the surprising association with amacrine cells. They were found to be significantly enriched for the Poor eye contact sub-term (see figure 14).

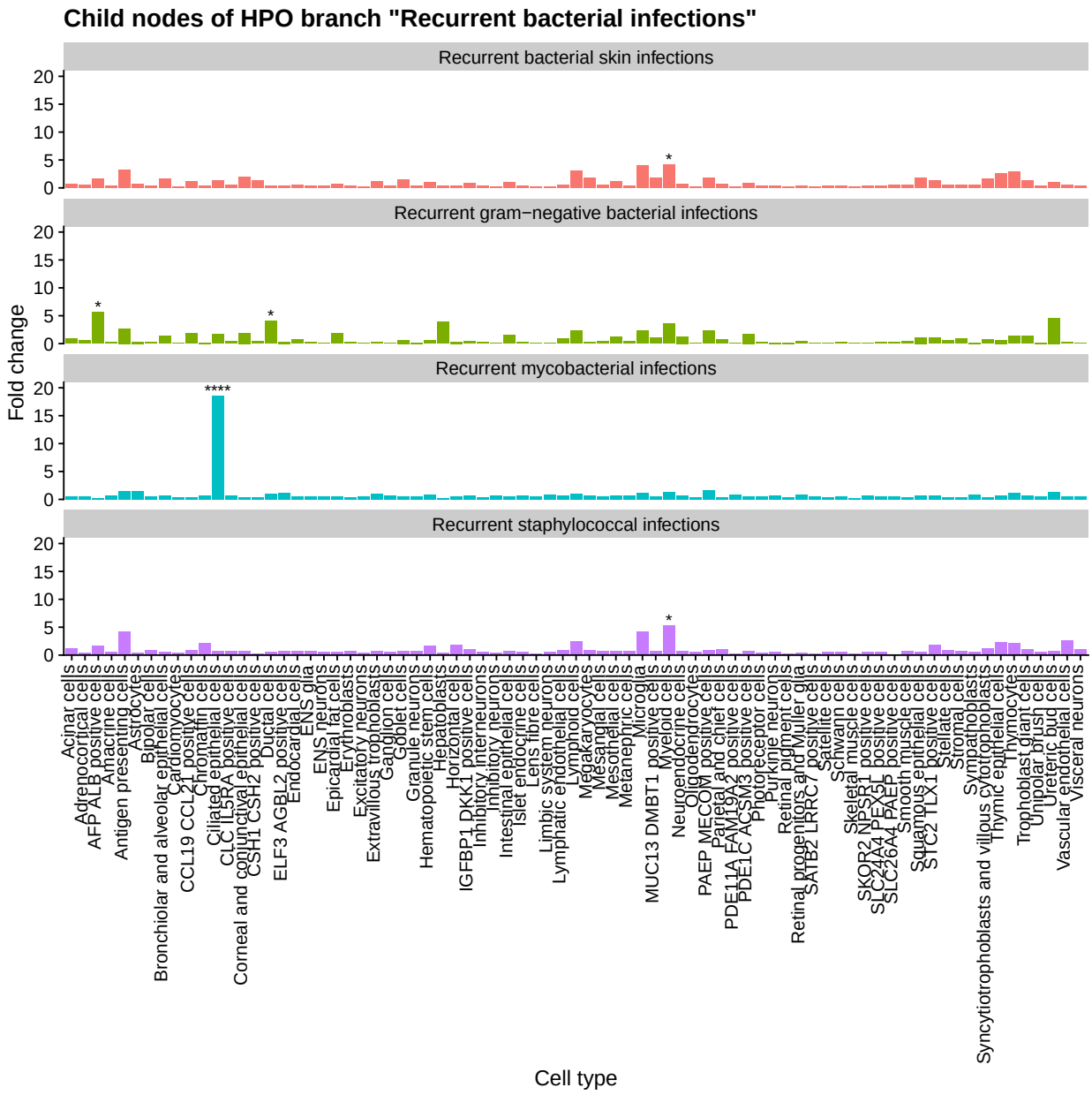

Figure 11: **Cell-phenotype associations for child terms of Recurrent bacterial infections.** Here we can see all cell-phenotype relationships for child phenotypes of the HPO term “Recurrent bacterial infections.” Significance is indicated with asterics where *, **, *** , and **** represent textit{q} less than 0.05, 0.001, 0.0001, and 0.00001. respectively.

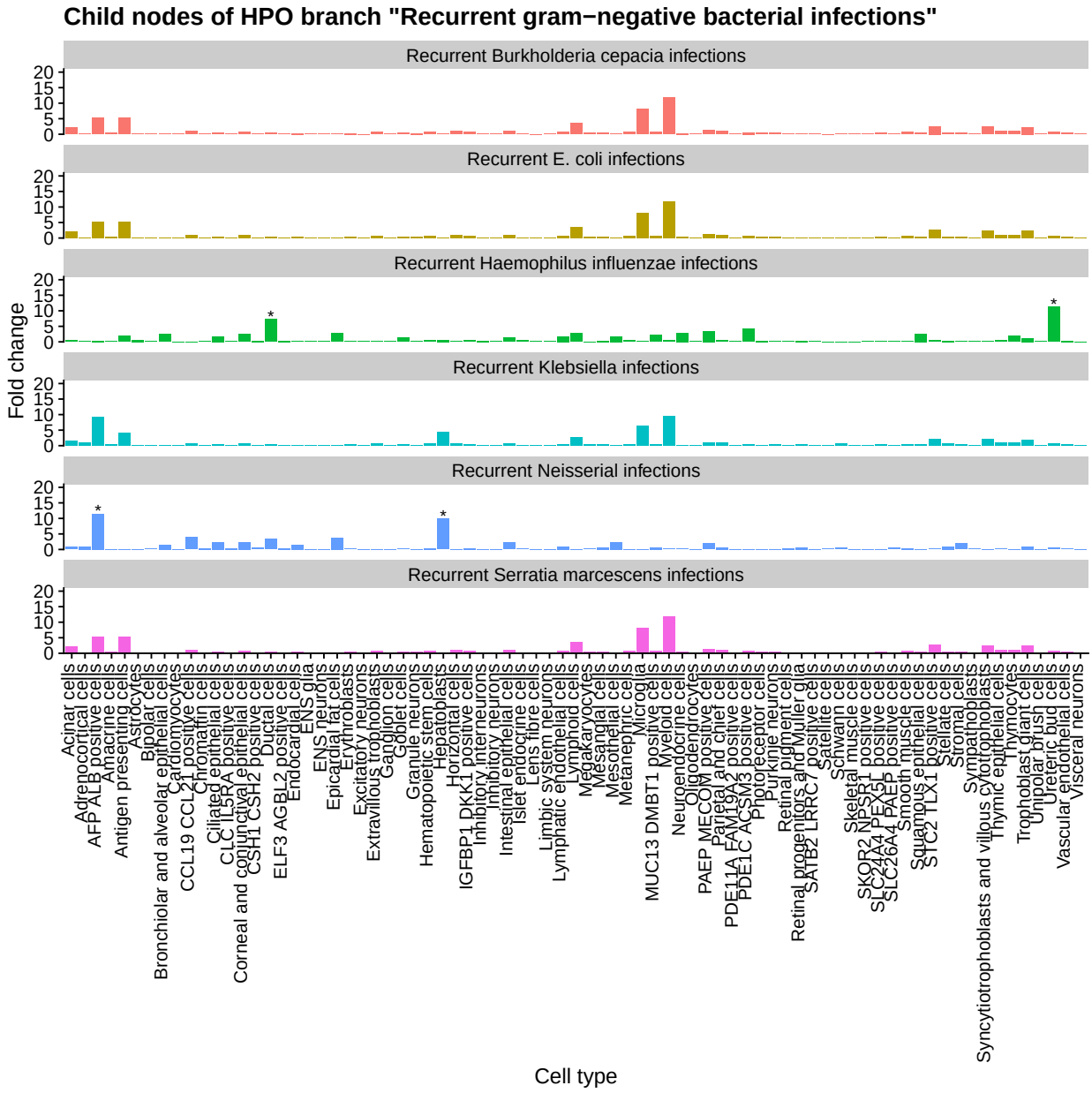

Figure 12: **Enrichments for child phenotypes of "Recurrent gram-negative bacterial infections.** Here we can see all cell-phenotype relationships for child phenotypes of the HPO term “Recurrant gram-negative bacterial infections.”

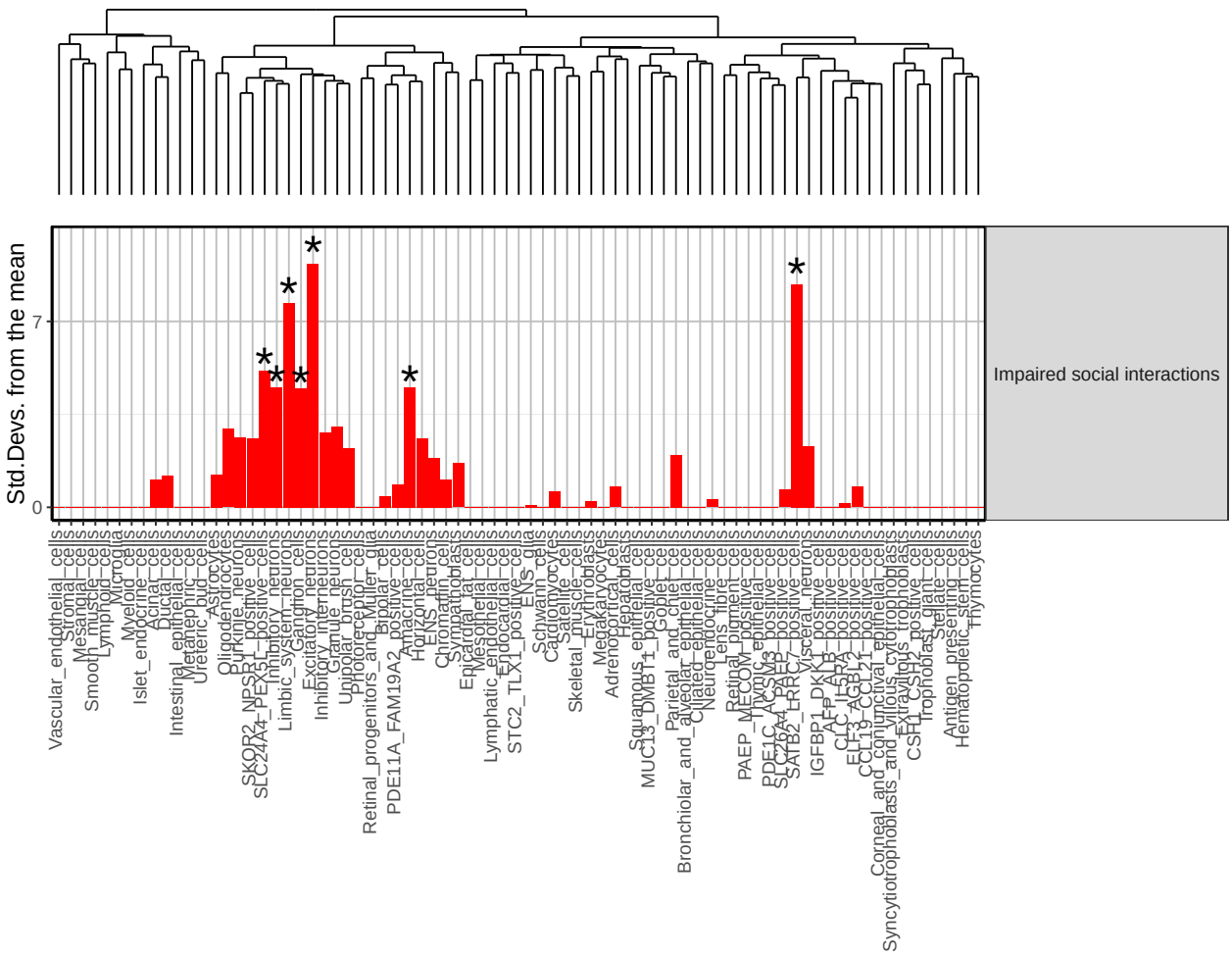

Figure 13: **Cell type enriched for Impared social interactions.** This shows the cell types that are significnatly enriched for the "Impared social interactions" gene list. The majority of significant enrichments are seen in cells of the nervous system, with particularly high specific expression in limbic system neurons and excitatory neurons. Interestingly, there is also a strong association with Amacrine cells, a retinal interneuron. Significance is denoted with asterics, where * signifies q<0.05

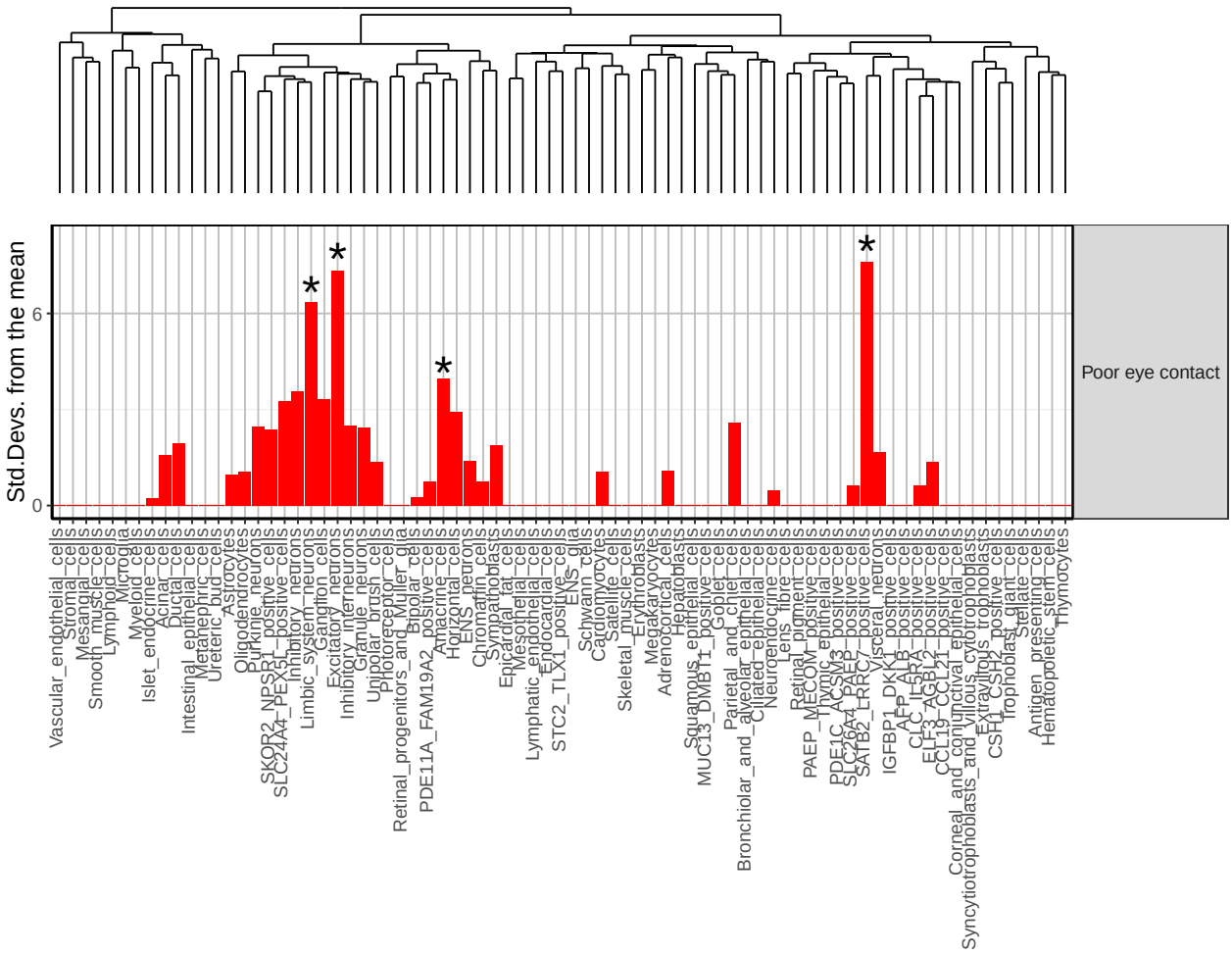

Figure 14: **Cell types enriched for Poor eye contact.** This shows the cells enriched for poor eye contact. It is interesting to note the surprising association of impaired social interaction with Amacrine cells, which are retinal interneurons, was found to come exclusively from a sub-phenotype related to the eye, “poor eye contact.” Significance is denoted with astericks, where * signifies q<0.05

Discussion

More than 8000 significant enrichments were found. These are spread across all HPO branches. High-level HPO terms, encompassing a broad range of traits, were significantly associated with expected cell types. This validates the method as it shows that the majority of cell-phenotype associations are in agreement with accepted knowledge. The number of expected enrichments also increases as more stringent significance thresholds are used, as the more well-understood cell-phenotype relationships tend to have a stronger enrichment.

It was found that, in significant cell-phenotype associations, the strength of enrichment (measured by fold change specific expression) tends to be higher at lower ontology levels. These more narrowly defined disease phenotypes also tend to have shorter, more specific gene lists. This potentially makes many of these low ontology level terms great targets for research. One of the roadblocks of gene therapy is the lack of understanding of the pleiotropic effects of many genes (Bulaklak and Gersbach, 2020). Being able to isolate the problem down to a small set of genes in a very specific cell type goes a long way to alleviating this problem. This could also provide a way to prioritise different research possibilities. The results could be subset to show all low ontology level HPO terms that are severe enough to warrant treatment and have a strong association with a particular cell type. Other parameters could also be used in this way. For example, to prioritise long-lived cell types that may be the best candidates for gene therapy.

Figure 7 showed that the majority of significant enrichments for terms from the main branches of the HPO are in expected cell types. For example, terms that are a subclass of “Abnormality of the cardiovascular system” are primarily enriched in cardiomyocytes. It was also noted that there are a number of enrichments in less obvious cell types. For example, a significant number of terms in the cardiovascular branch were enriched in hepatoblasts (p<0.0001). To explore this further, we retrieved all results from

the cardiovascular branch that are enriched in hepatoblasts and have a fold change > 7, and $q < 0.05$. The phenotypes were primarily ones that often involve damage to arteries caused by lipid deposition (cerebral artery atherosclerosis, joint hemorrhage, myocardial steatosis, precocious atherosclerosis, premature arteriosclerosis). Given the large role that the liver plays in lipid metabolism, it is unsurprising that dysfunction of hepatocytes may be implicated in these cardiovascular diseases. Additionally, a brief literature search finds many recent studies exploring the link between hepatic cells and cardiovascular diseases, such as the paper by Xu et al. (2021) which shows that hepatocyte ATF3 is protective against atherosclerosis by regulating high-density lipoprotein metabolism. Furhter, Bell et al. (2018) showed that higher circulating hepatocyte growth factor is associated with elevated atherosclerosis progression measures.

More specific HPO terms from lower ontology levels were also primarily enriched in expected cell types. For example, it was found that the majority of significant enrichments for descendants of “Recurrent bacterial infections” were in immune cells. Again some less obvious enrichments were also found, such as ciliated epithelial cells. To investigate this further, the EWCE results for descendant terms of “Recurrent bacterial infections,” that were also enriched in ciliated epithelial cells, were retrieved individually. It was found that the enrichments are primarily associated with the recurrent respiratory infections subbranch (chronic bronchitis, recurrent bacterial infections, recurrent bronchitis, recurrent infections, recurrent mycobacterial infections, recurrent otitis media, recurrent respiratory infections, recurrent sinopulmonary infections, recurrent sinusitis, recurrent upper respiratory tract infections). This relationship between structural defects of the cilia and recurrent respiratory tract infections has been documented in the literature (Eliasson et al., 1977). The strongest association was found in mycobacterial infections (which includes tuberculosis), and this also primarily affects the lungs (fold change=18.52, $q<0.00001$). The only non-pulmonary phenotype was associated with ciliated cells was recurrent otitis media infections, which is also known to be a problem for people with immotile-cilia (Mossberg

et al., 1983).

Another potentially novel finding was the significant enrichment in hepatoblasts for a term in the “Recurrent infections” branch. This was found to be specific to recurrent neisserial infections. For further confirmation, this phenotype was also analysed in the Tabula Muris data (mouse scRNA dataset), where it was also significantly enriched in hepatic cells (kupffer cells and hepatocytes). Whilst this is a seemingly surprising association, we found a number of plausible explanations for it in the literature. Fitz-Hugh-Curtis syndrome is a RD characterised by inflammation of the peritoneum and the tissues surrounding the liver, caused by *Neisseria gonorrhoeae* infection (Rueda et al., 2017). It is possible that an abnormality of hepatic cells leaves patients particularly susceptible to this problem. Though Fitz-Hugh_Curtis syndrome is a sub-phenotype of recurrent neisserial infections, at the time of writing, there was no entry for it in the HPO, so it was not possible to check if the enrichment was specific to this sub-phenotype. Perhaps the most likely explanation involves the complement system, a part of the innate immune system that plays a key role in defense against neisserial infections. Complement is synthesised primarily in the liver, and it was found that people with deficits in complement are at high risk for Neisserial infection (Fijen et al., 1994; Lewis and Ram, 2020).

Significant enrichments for “Impaired social interactions” were then examined to take an example from another branch of the HPO. The expectation was that it would primarily involve neuronal cell types. There were strong enrichments in excitatory neurons (fold change=2.366, q<0.00001) and limbic system neurons (fold change=2.119, q<0.00001). While we may need to use higher resolution cell type data to explore the excitatory neurons result, which could encompass a lot of different cells, the association with limbic system neurons was found to be supported in the literature. The basolateral circuit encodes information about social signals, and the amygdala and cingulate gyrus are involved in social cognition, all of which are limbic system structures (Frith, 1996).

More surprisingly, an association between impaired social interaction and amacrine cells was found (fold change=1.6, q=0.0156625). These are a type of interneuron found in the retina. At first glance, this seemed like it may have been a false positive, possibly due to similarities to other cell types. In cases like this, it is possible to do a conditional enrichment analysis with EWCE that controls for expression in other cells, which may be worthwhile when investigating the more novel results in the future. However, to explore the association further, the descendant terms of impaired social interactions were checked to see if there was a more specific association between amacrine cells and one of the sub-phenotypes (impaired ability to form peer relationships, impaired use of nonverbal behaviors, poor eye contact, lack of peer relationships, no social interaction). Interestingly, the significant effect was found exclusively in the Poor eye contact (HP:0000817) sub-term (fold change=1.6248474, q<0.05). These results may imply that, in some cases, poor eye contact could be caused by a physical problem with structures in the eye, rather than with the psychological origins normally associated with impaired social interaction related phenotypes.

From these examples, it can be seen that when scrutinised, many of the more surprising cell-phenotype relationships are either previously known or at least have a plausible mechanistic basis. It is very likely that within the over 8000 results presented here, there are many previously unknown links between disease phenotypes and specific cell types that could lead to advances in the understanding and treatment of RDs. For the impact of the results to be fully realised, it was essential that they could be easily explored by domain experts and clinicians without the need for extensive computational resources and specialist knowledge of enrichment analysis and programming. The web-based applications were developed to facilitate this. Although further work is needed, a preliminary version has been deployed here [https://ovrhuman.github.io/ewce_website/]. Increased efficiency is needed in the cell select app, as the large data set and complex graph algorithms result in slow loading times.

This study would not have been possible with other methods such as hypergeometric tests and PSEA, which would not account for the specificity of gene expression, or would need quantitative expression data from the disease state, making them unable to distinguish between secondary "reactive" expression and primary expression associated with the main genetic susceptibility. The lack of requirements for disease tissue expression data made it possible to utilise large, publicly available lists of simple phenotype-gene associations. One disadvantage was the substantial amount of computation required for the analysis. The interactive HPC sessions had a time limit of 8 hours which was not long enough to complete the analysis in one pass, causing it to take substantially longer. In future analysis, we plan to use a 32 Core Threadripper, acquired by the Neurogenomics lab, which will be much faster and has no time limit.

Another issue was that the HPO and Descartes data are updated and changed over time. Directly pulling the data from these resources for each analysis caused some problems with reproducing previous analysis. This was resolved by privately hosting the data in its current state, whilst still including an option in the makefile to download the latest data.

One of the goals of bio-ontology data structures is the integration of information and making implicit knowledge become explicit. To this point, all of the cell-phenotype associations found in this study were already present in pre-existing data but had not yet been made accessible or explicitly stated. A future plan is to continue to bring out this type of information, thereby extending the description logic the bio-ontology. For example, the phenotypes are currently organised as a network of "is-a" connections, which is useful for understanding how the physical attributes of diseases are related. But, the phenotypes could also be connected in other ways; for example, the associated gene lists for each phenotype could be used to uncover genetic associations between phenotypes from completely separate "is-a" branches. As a proof of concept, a bootstrap enrichment analysis was done with 10,000 reps to test which phenotype gene lists are enriched for the "Abnormal lip morphology" gene list. In other words, to find phenotypes that are geneti-

cally similar or related to abnormal lip morphology. Though still in early development, the algorithm created for this phenotype gene list similarity analysis is shown below in code example 3.

```
similar_phenotypes <- function(phenotype,phenotype_to_genes,bootstrap=1000) {
  phenotype_similarity_scores = data.frame()
  # Primary Phenotype gene list
  pheno_genes = unique(phenotype_to_genes[
    phenotype_to_genes$Phenotype == phenotype, "Gene"])
  for (p in unique(phenotype_to_genes$Phenotype)) {
      # Comparison Phenotype gene list
      comparison_pheno_genes=unique(
        phenotype_to_genes[phenotype_to_genes$Phenotype==p,"Gene"])
      if (length(comparison_pheno_genes)>4) {
        n_matches=length(pheno_genes[pheno_genes%in%comparison_pheno_genes])
        # Create bootstrap distribution for current comparison phenotype
        all_genes = unique(phenotype_to_genes$Gene)
        bootstrap_distribution = c()
        for (i in seq(1:bootstrap)){
          bootstrap_genelist=sample(
            all_genes,size=length(comparison_pheno_genes),replace=FALSE)
          bootstrap_n_matches=length(
            pheno_genes[pheno_genes %in% bootstrap_genelist])
          bootstrap_distribution=append(
            bootstrap_distribution,bootstrap_n_matches)}
        # calculate Z score for the comparison
        Z_score=(n_matches-mean(bootstrap_distribution))/sd(
          bootstrap_distribution)
        p_val = 2*(length(bootstrap_distribution[
          bootstrap_distribution>=n_matches])/length(bootstrap_distribution))
      } else {
        Z_score = NA
        p_val = NA }
      phenotype_similarity_scores = rbind(phenotype_similarity_scores,
        data.frame("Primary_Phenotype"=phenotype,"Comparison_Phenotype"=p,
                   "Z_score"=Z_score,"p_val" = p_val)) }
  return(phenotype_similarity_scores)}
```

Code example 3: **Phenotype-phenotype bootstrap enrichment analysis function.** This is the algorithm currently being developed for the bootstrap enrichment analysis. It determines whether one gene list is enriched in another. Bootstrapping is used rather than hypergeometric test because it accounts for the specificity of genes to phenotypes. Simply put, it allows us to identify genetically similar RDs whilst accounting for how frequently individual genes are associated with RDs.

It was predicted that “Abnormal lip morphology” would have many genetically similar phenotypes in the “Abnormality of the nervous system” branch. This is because abnormal

lip morphology can be caused by abnormal migration of neural crest cells during development and is often associated with abnormalities of the central nervous system. Additionally, both share an ectodermal origin (Mueller et al., 2007). Again for hypothesis testing purposes, we also looked at our other two example branches (immune and cardiovascular system). We found that there were 245 significantly similar gene lists in the nervous system branch, and only 77 and 29 in the cardiovascular and immune system branches, respectively. Additionally, the slightly more related phenotypes we found in the cardiovascular branch may be explained by the fact that neural tube defects have also been associated with congenital heart diseases (Gardner, 1981).

The nervous system phenotypes that were found to be genetically similar to "Abnormal lip morphology" included many that would be very difficult to diagnose at birth, such as ADHD, Delayed speech and language development, and intellectual disability, and many other neurological abnormalities. These examples were all highly significant (q<0.0001 and fold change>15). These findings appear to be supported by the literature. For example, research by Conrad et al. (2008) showed that children with isolated cleft lip and/or palate had significantly higher levels of neurological soft signs than controls. It was also found that the known association between abnormal lip morphology and abnormal neuronal migration was reproduced by this analysis (gene list similarity between "Abnormal lip morphology" and "Abnormal neuronal migration": Z=12.9, q<0.00001).

The next step would be to perform the same enrichment analysis with 100,000 reps for all pairwise phenotype-phenotype relationships. This could extend the description logic of the HPO bio-ontology to have "has-a" connections, representing relationships between phenotypes that have a significantly similar gene list, rather than just the "is-a" connections that are currently available. The full impact of doing this is hard to predict, but an example use-case could be to find all instances of easily diagnosed diseases that are genetically similar to ones that are harder to diagnose. This could help to target diagnostic and screening resources to patients at higher risk of certain conditions, possibly

leading to faster and more effective treatment. For example, it was found that the “Autism” gene list was significantly enriched for the “Abnormal lip morphology” gene list (Z=13.8, q<0.00001). Abnormal lip morphology is clearly easier to diagnose at birth than autism, and it has been shown that early autism spectrum disorder intervention (age < 3) provides significant benefit (Zwaigenbaum et al., 2015). Further work would need to be done in this case to determine if abnormal lip morphology is predictive of autism in human patients. However, it is easy to see how these kinds of findings could provide beneficial insights for RD patients, especially given that diagnosis is such a significant problem for many of them.

In addition to these plans to extend the analysis, much of the code written for the RD EWCE analysis will be useful for similar studies in the future, as well as for follow up RD studies when more data becomes available. A future goal for this project is to share the code in open-source software repositories. This would make this kind of analysis accessible to the wider research community. A start was made by submitting additional features to the EWCE Bioconductor package, and storing the main RD EWCE script on the UKDRI Neurogenomics lab GitHub repository, but much more remains to be done.

Although individual RDs have a low prevalence, when taken as a whole, they impact the lives of a significant number of people worldwide. The increasing accessibility of genomic and phenotypic data, high throughput analytic techniques, and improved ways to organise and disseminate information have made this project possible. It is hoped that the results presented here can help to overcome some of the problems of resource allocation and information scarcity that have hindered RD research in the past.

Acknowledgments

This work would not have been possible without the continued guidance and support from Dr Nathan Skeen, Brian M. Shlider, Alan Murphy and the rest of the UKDRI Neurogenomics lab from the Imperial college department of Brain Sciences. The generosity with lab resources and their personal time and energy made the project both enjoyable and an exceptional learning experience.

References

Anaconda software distribution, 2020. Anaconda Documentation.

Bell, E.J., Decker, P.A., Tsai, M.Y., Pankow, J.S., Hanson, N.Q., Wassel, C.L., Larson, N.B., Cohoon, K.P., Budoff, M.J., Polak, J.F., Stein, J.H., Bielinski, S.J., 2018. Hepatocyte growth factor is associated with progression of atherosclerosis: The Multi-Ethnic Study of Atherosclerosis (MESA). Atherosclerosis 272, 162–167. https://doi.org/10.1016/j.atherosclerosis.2018.03.040

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Briatte, F., 2020. Ggnetwork: Geometries to plot networks with 'ggplot2'.

Bulaklak, K., Gersbach, C.A., 2020. The once and future gene therapy. Nature Communications 11, 5820. https://doi.org/10.1038/s41467-020-19505-2

Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., Spielmann, M., Palis, J., Doherty, D., Steemers, F.J., Glass, I.A., Trapnell, C., Shendure, J., 2020. A human cell atlas of fetal gene expression. Science 370, eaba7721. https://doi.org/10.1126/science.aba7721

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B., 2021. Shiny: Web application framework for r.

Conrad, A.L., Canady, J., Richman, L., Nopoulos, P., 2008. Incidence of Neurological Soft Signs in Children with Isolated Cleft of the Lip or Palate. Perceptual and Motor Skills 106, 197–206. https://doi.org/10.2466/pms.106.1.197-206

Eliasson, R., Mossberg, B., Camner, P., Afzelius, B.A., 1977. The Immotile-Cilia Syndrome: A Congenital Ciliary Abnormality as an Etiologic Factor in Chronic Airway Infections and Male Sterility. New England Journal of Medicine 297, 1–6. https://doi.org/10.1056/NEJM197707072970101

Fijen, C.A.P., Kuijper, E.J., Tjia, H.G., Daha, M.R., Dankert, J., 1994. Complement Deficiency Predisposes for Meningitis Due to Nongroupable Meningococci and Neisseria-Related Bacteria. Clinical Infectious Diseases 18, 780–784. https://doi.org/10.1093/clinids/18.5.780

Frith, C., 1996. Brain mechanisms for 'having a theory of mind'. Journal of Psychopharmacology 10, 9–15. https://doi.org/10.1177/026988119601000103

Gardner, W.J., 1981. Overdistention of the neural tube causes congenital heart disease. Medical Hypotheses 7, 411–420. https://doi.org/10.1016/0306-9877(81)90028-1

Greene, D., Richardson, S., Turro, E., 2017. ontologyX: A suite of R packages for working with ontological data. Bioinformatics btw763. https://doi.org/10.1093/bioinformatics/btw763

Haendel, M.A., Chute, C.G., Robinson, P.N., 2018. Classification, ontology, and precision medicine. New England Journal of Medicine 379, 1452–1462. https://doi.org/10.1056/NEJMra1615014

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., Callahan, T.J., Chute, C.G.,

Est, J.L., Galer, P.D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., Harris, N.L., Hartnett, M.J., Hastreiter, M., Hauck, F., He, Y., Jeske, T., Kearney, H., Kindle, G., Klein, C., Knoflach, K., Krause, R., Lagorce, D., McMurry, J.A., Miller, J.A., Munoz-Torres, M.C., Peters, R.L., Rapp, C.K., Rath, A.M., Rind, S.A., Rosenberg, A.Z., Segal, M.M., Seidel, M.G., Smedley, D., Talmy, T., Thomas, Y., Wiafe, S.A., Xian, J., Yüksel, Z., Helbig, I., Mungall, C.J., Haendel, M.A., Robinson, P.N., 2020. The Human Phenotype Ontology in 2021. Nucleic Acids Research 49, D1207–D1217. https://doi.org/10.1093/nar/gkaa1043

Kuhn, A., Thu, D., Waldvogel, H.J., Faull, R.L.M., Luthi-Carter, R., 2011. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nature Methods 8, 945–947. https://doi.org/10.1038/nmeth.1710

Lewis, L.A., Ram, S., 2020. Complement interactions with the pathogenic Neisseriae: Clinical features, deficiency states, and evasion mechanisms. FEBS Letters 594, 2670–2694. https://doi.org/10.1002/1873-3468.13760

Mossberg, B., Camner, P., Afzelius, B., 1983. The immotile-cilia syndrome compared to other obstructive lung diseases: A clue to their pathogenesis. European journal of respiratory diseases. Supplement 127, 129—136.

Mueller, A.A., Sader, R., Honigmann, K., Zeilhofer, H.-F., Schwenzer-Zimmerer, K., 2007. Central nervous malformations in presence of clefts reflect developmental interplay. International Journal of Oral and Maxillofacial Surgery 36, 289–295. https://doi.org/10.1016/j.ijom.2006.10.018

Nguengang Wakap, S., Lambert, D.M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., Rath, A., 2020. Estimating cumulative point prevalence of rare diseases: Analysis of the Orphanet database. European Journal of Human Genetics 28, 165–173. https://doi.org/10.1038/s41431-019-0508-0

Peltonen, L., Perola, M., Naukkarinen, J., Palotie, A., 2006. Lessons from studying monogenic disease for common disease. Human Molecular Genetics 15, R67–R74. https://doi.org/10.1093/hmg/ddl060

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., Ayme, S., 2012. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. Human Mutation 33, 803–808. https://doi.org/10.1002/humu.22078

Rueda, D.A., Aballay, L., Orbea, L., Carrozza, D.A., Finocchietto, P., Hernandez, S.B., Volpacchio, M.M., Fonzo, H. di, 2017. Fitz-Hugh-Curtis Syndrome Caused by Gonococcal Infection in a Patient with Systemic Lupus Erythematous: A Case Report and Literature Review. American Journal of Case Reports 18, 1396–1400. https://doi.org/10.12659/AJCR.906393

Skene, N., 2021. EWCE: Expression weighted celltype enrichment.

Skene, N.G., Grant, S.G.N., 2016. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. Frontiers in Neuroscience 10. https://doi.org/10.3389/fnins.2016.00016

The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, Principal investigators, 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562, 367–372. https://doi.org/10.1038/s41586-018-0590-4

Wickham, H., 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Xu, Y., Li, Y., Jadhav, K., Pan, X., Zhu, Y., Hu, S., Chen, S., Chen, L., Tang, Y., Wang, H.H., Yang, L., Wang, D.Q.-H., Yin, L., Zhang, Y., 2021. Hepatocyte ATF3 protects against atherosclerosis by regulating HDL and bile acid metabolism. Nature Metabolism 3, 59–74. https://doi.org/10.1038/s42255-020-00331-1

Zwaigenbaum, L., Bauman, M.L., Choueiri, R., Kasari, C., Carter, A., Granpeesheh, D., Mailloux, Z., Smith Roley, S., Wagner, S., Fein, D., Pierce, K., Buie, T., Davis, P.A., Newschaffer, C., Robins, D., Wetherby, A., Stone, W.L., Yirmiya, N., Estes, A., Hansen, R.L., McPartland, J.C., Natowicz, M.R., 2015. Early Intervention for Children With Autism Spectrum Disorder Under 3 Years of Age: Recommendations for Practice and Research. PEDIATRICS 136, S60–S81. https://doi.org/10.1542/peds.2014-3667E