Expression Weighted Cell Type Enrichment as a Tool for Identifying Cell Types Underlying Rare Disease Phenotypes

Jai Chapman [1], Brian M Schilder [1], Nathan G Skene [1,2]
[1] UK Dementia Research Institute, Imperial College London, London W1T 7NF
[2] Lead Contact

CID: 01560783 Word count: 4990

This research report is submitted in partial fulfilment of the requirements for the degree of BSc in Medical Biosciences.

COVID-19 Disruption Form 2020-21

This form should be completed by supervisors and inserted into the report after the cover page.

The purpose of the form is to highlight the extent of disruption due to the COVID-19 pandemic on the LABP, LITP or WKBP project. Please state whether the project has encountered problems beyond your control and give a brief description. Please choose the extent of disruption to the project for each category and overall:

- Minimal (at least two thirds carried out as expected)
- Moderate (at least one third carried out as expected)
- Severe (less than one third carried out as expected)

Changes to project plan Please use this section to state any modifications/changes to the original project plan that were required as a result of COVID-19 restrictions. Specify the dates that the project was impacted and the overall extent of the impact.	☒ Minimal ☐ Moderate ☐ Severe
Project is entirely computational so we were able to conduct remotely. However, working entirely remotely made it harder for Jai to get support from other lab members which will no doubt have impacted his progress given the steep learning curve that was associated with the project.	

Changes to project implementation Please use this section to highlight any issues that arose during the project that prevented its implementation as a result of COVID-19 restrictions. This may include any restrictions to site access or labs or resources, placement activity or student-supervisor contact. Specify the dates that the project was impacted and the overall extent of the impact.	☒ Minimal ☐ Moderate ☐ Severe
Site access was restricted/discouraged for people working on computational projects throughout the whole project	

	Name	Signature	Date
Supervisor	Nathan Skene		20th April 2021
Student	Jai Chapman		20th April 2021

ABSTRACT

The introduction of single cell RNA sequencing has seen vast improvements in the characterisation of individual cell types. Gene enrichment analysis can harness this to statistically determine whether genes associated with disease are highly represented or 'enriched' in given cell type. The Expression Weighted Cell Type Enrichment (EWCE) method was developed in order to facilitate the genetic identification of cell types underlying disease and has been shown to be successful in determining gene enrichment with no prior knowledge of genetic specificity to cells. So far, genetic cell typing has been largely focused on polygenic diseases. Rare diseases are poorly understood and require new approaches such as enrichment analysis to elucidate their causal cell types, as it is difficult to test individual cases. In this study, EWCE was applied to genetic susceptibility data derived from the Human Phenotype Ontology (HPO) to demonstrate a high-throughput method of gene enrichment analysis able to identify cell types underlying rare disease phenotypes. This was achieved by testing for enrichment across all phenotypes within the HPO using single cell mouse transcriptomic data as a reference atlas for gene expression. It was demonstrated that the approach was able to recover accurate, expected cell type enrichments across multiple levels of the HPO, and that the method recovered higher proportions of expected relationships in line with increased stringency. This demonstration opens avenues for further study, and future extensions of this work could apply EWCE to further single cell datasets and eventually propose cell type-specific molecular targets for therapeutics.

Keywords: Cell types, gene enrichment analysis, gene expression, phenotypes, rare diseases, transcriptomics

Abbreviations: Expression Weighted Cell Type Enrichment (EWCE), Human Phenotype Ontology (HPO), Single Cell RNA Sequencing (scRNAseq)

INTRODUCTION

Developments in single cell sequencing technology have made it possible to profile the individual characteristics of distinct cell types with increased accuracy. (1) Single cell RNA sequencing (scRNAseq) quantifies the abundance of messenger RNA (mRNA) transcribed for all genes within a given cell. (2) In understanding this, cell functionality can be ascertained through interpreting protein expression, distinguishing cells phenotypically. As scRNAseq datasets have become more readily available, opportunities have arisen in the domain of identifying the cell types implicated in diseases and disease phenotypes with no currently understood mechanisms or treatments. (3) Whereas genetic variants associated with disease continue to be catalogued via approaches such as genome-wide association studies (GWAS) (4,5) and family pedigree studies (6), there remains a crucial need to draw links between these variants and the underlying cellular mechanisms that ultimately lead to disease.

Gene set enrichment analysis tests whether the genes in list A (e.g., cell type markers) are overrepresented in gene list B (e.g., those associated with a disease or phenotype) relative to what would be expected chance. (7) Through this, cell types potentially causal to disease can be identified by determining whether genes associated with a given disease or phenotype are particularly enriched for a given cell type. In 2016, Skene et al. introduced the Expression Weighted Cell Type Enrichment (EWCE) method, which uses an iterative bootstrapping procedure to statistically test for enrichment of disease-associated genes in cell type-specific gene signatures derived from scRNAseq data. (8) Compared to prior methods, EWCE simplifies the approach of cellular gene enrichment analysis by allowing any input gene list to be computed without the need for prior knowledge regarding the specificity of genes to particular cell types. This means that any list of genes can be tested. The original 2016 publication focused on identifying causal cell types in disorders of the brain such as Alzheimer's and schizophrenia, examples of complex, polygenic diseases.

Rare diseases remain poorly understood for the most part, compared to more common disorders. Complex phenotypic traits are widely understood to be driven by small effects driven by many genetic variants distributed across the genome, that combine to produce an overall large effect (e.g., height, Alzheimer's disease). (9) In contrast, rare diseases tend to be associated with fewer genes and generally involve single mutations (e.g., Huntington's disease). (10,11) There are thousands of rare diseases currently documented, making it impractical to study all of them in depth through traditional approaches. (12,13) Therefore, high-throughput computational methods like gene enrichment analysis are necessary, as well as tools to systematically investigate their underlying biological mechanisms.

In this study, it was hypothesised that by considering rare diseases as clusters of shared phenotypes, EWCE could be used to perform enrichment analysis across multiple phenotypes in order to simultaneously recover cell type enrichments underlying numerous, individually reported disorders. Using EWCE, gene enrichment analysis was carried out using every phenotype term contained within the Human Phenotype Ontology (HPO). (14) The HPO provides an invaluable library of disease phenotypes, including rare diseases, alongside genes with associated variants. In the case of rare diseases, these variants are mainly mendelian in nature. This work contrasts the previous 2016 demonstration of EWCE which focused specifically on gene lists for polygenic disorders primarily gathered from GWAS. (8)

Here, the approach is shown to recover accurate cell-phenotype relationships across a broad range of phenotypes throughout the HPO. The recently published body-wide scRNAseq mouse atlas *Tabula Muris* was utilised as a reference transcriptome dataset for EWCE. (15) *Tabula Muris* is comprised of 120 distinct cell types from 20 tissues, which for this study have been grouped into 38 broader classes of cell types.

The primary aim of this initial study was to demonstrate that the approach could recover expected cell-phenotype relationships. Utilising the structure of the HPO, which categorises phenotypes according to distinct body systems, it was found that this was indeed the case. Thousands of enrichments were found, so for the purposes of this study results are demonstrated in the context of three high-level branches of terms within the HPO representing phenotypes associated with the nervous system, cardiovascular system, and immune system. These branches were chosen as they were expected to directly correlate to a large proportion of cells within the *Tabula Muris* dataset, which is highly representative of immune cells, neurons, glial cells, and cardiac muscle cells. This was confirmed to a high degree of significance. For example, it is shown here that the highest proportion of enrichments for the 'Abnormality of the nervous system' HPO branch was associated with neurons. Following this, various tests are carried out to confirm that increased stringency recovered higher proportions of expected enrichments. Finally, lower level HPO terms representing rare phenotypes are analysed, demonstrating that the approach can recover precise cell-phenotype relationships for rare phenotypes from the most specific levels of the HPO. Though just one transcriptomic reference dataset is utilised here (*Tabula Muris*), this approach may be applied using other single cell libraries as they become available. Overall, this exploratory study provides initial results for a promising method of gene enrichment analysis that in the future may be utilised to elucidate novel cell type enrichments for rare disease phenotypes as more detailed scRNAseq datasets are published.

METHODS AND MATERIALS

All analyses were performed in R (version 4.0.3). (16)

Processing of Single Cell Transcriptome Data

Single cell mouse RNA sequencing data was obtained from *Tabula Muris*. This dataset contains gene expression data derived from fluorescence-activated cell sorting of over 100,000 cells of *Mus musculus*, representing 120 distinct classifications of cells from 20 different tissues. The raw files were first processed into a 'Cell Type Data' (CTD) file, a required format for the EWCE method whereby a matrix is generated representing the mean expression of every gene for each cell type. Additionally, gene-by-cell specificity matrices were generated. The calculation for gene specificity involves dividing the average level of expression for a cell type by the sum of its entire expression across all cell types.

Here, the 120 cell types in the original *Tabula Muris* publication were further grouped into 38 broader classes according to a hierarchal clustering algorithm (*hclust*, R version 4.0.3) ran on respective gene specificity matrices. This grouping involved categorising individual cells, e.g., endothelial cells from various tissues, into broader groups, e.g., 'endothelial cells'. Final cell classifications are visible in **Supplementary Figure 1**. Though this broader grouping reduces the amount of necessary computational power, lessens the multiple-testing burden, and simplifies interpretability of results, the cost of using these grouped cell classes is reduced granularity for particular cell types involved in each phenotype.

Gene List Acquisition from the Human Phenotype Ontology

The HPO provides an extensive library of human disease phenotypes alongside any genes reported to have a variant linked with these phenotypes. Many associations are based on rare disease phenotypes and are mendelian in nature. At present, over 13,000 terms are contained within the ontology. This data is available from the main HPO website in tabular format, which is updated every two months. (17) This study utilised the December 2020 release for all analyses.

Target gene lists were created by iteratively running through each unique term in the HPO and storing their associated gene lists in a 'list' object within R. The HPO exists as a directed acrylic graph, whereby phenotype terms are related through 'is_a' relationships. The resulting structure contains ancestor terms as well as descendant terms for a particular phenotype. For example, the term 'Abnormality of the vasculature' has an ancestor term of 'Abnormality of the cardiovascular system', as well as several descendant terms such as 'Abnormal vascular morphology' and 'Vascular neoplasm'. In this study, gene lists were derived from each term irrespective of ontology depth, but analysis was completed respective to ontology levels.

Application of Expression Weighted Cell Type Enrichment (EWCE)

EWCE exists as a package for the R programming language. (8,16) It tests for enrichment of a target gene list for each cell type within a scRNAseq dataset.

EWCE identifies cell type-specific signatures for gene lists using the aforementioned specificity matrices, in each cell type, through bootstrapping analysis. Bootstrapping was performed by generating lists of randomly selected genes (without replacement) from the background set (all genes identified within *Tabula Muris*), in order to calculate a probability density for an average level of gene expression. Bootstrapping was run for 100,000 iterations to ensure robustness of the results (the default value within EWCE is 100). For each cell type, enrichment for a target gene list is determined through the number of bootstrapped gene lists found to have higher specificity than the target list:

$$p\ (target\ gene\ list\ X\ enriched\ for\ cell\ type\ c) = \frac{\sum_{j=1}^{100000} \begin{cases} 1 & \gamma(X,c) < \gamma(D_j,c) \\ 0 & \gamma(X,c) > \gamma(D_j,c) \end{cases}}{100000}$$

Where:

$$\gamma = Expression$$
$$c = a\ cell\ type$$
$$X = a\ target\ gene\ list$$
$$D_j = a\ bootstrapped\ gene\ list$$

From the bootstrapping results, EWCE computes a table of *p*-values, standard deviation from mean expression, and fold-enrichment (normalised fold-change) for each cell type tested. Fold-enrichment is calculated by dividing the mean expression of a target gene list by the mean expression of bootstrapped gene lists.

Target gene lists were generated in R by iteratively pulling the genes associated with every individual term in the Human Phenotype Ontology (see above). *p*-values were corrected for multiple-testing using the Benjamini-Hochberg method (18) across all cell types and phenotypes.

Statistical Analysis

Statistical analysis was inherent to the EWCE method as its bootstrapping procedure involves the generation of a *p*-value for a cell-phenotype relationship (see above).

Where testing for association of cell types with phenotypes of high level HPO branches, one-tailed hypergeometric tests were carried out using the '*phyper'* function within R to test significance, generating additional *p*-values.

RESULTS

Recovery of Expected Cell Type Enrichments within Broad HPO Branches

EWCE was ran on the reference gene expression atlas obtained from *Tabula Muris*, testing enrichment for every unique term within the HPO. A total of 5,509 significant enrichments ($p < 0.05$) were found across all cell types (**Supplementary Figure 2**). It was first sought to validate the approach by testing whether known relationships between cell types and phenotypes were recovered in the results.

The HPO follows a hierarchal structure of phenotypes, meaning each phenotype has both ancestor and descendant terms. At its highest level, there are 23 broad categories of terms representing phenotypes across a range of body systems and anatomical locations (**Figure 1A**). While *Tabula Muris* does not cover all known cell types, and thus each cell type does not correlate directly to a branch of phenotypes within the HPO, it does sufficiently sample cell types across a wide range of organ systems. In turn, there are several phenotype branches within the HPO that do represent expected phenotypes for various cell types within *Tabula Muris*, offering an opportunity to validate the method by analysing recovery of expected enrichments. These branches are 'Abnormality of the nervous system', 'Abnormality of the cardiovascular system' and 'Abnormality of the immune system'.

The full set of results was analysed to visualise the proportion of significant enrichments ($p < 0.05$) for descendant terms of each of the above HPO branches that mapped to each cell type in the dataset. It was hypothesised that a large proportion would be associated with cells derived from brain, cardiac, and immune tissue, respective to each aforementioned branch. Hypergeometric tests were carried out to determine whether branches were significantly associated with a given cell type, with respect to all of its significant enrichments.

As shown in **Figure 1B**, the cell type found to represent the highest proportion of significant enrichments for descendant terms of 'Abnormality of the nervous system' was neurons (28%, $p < 0.0001$), followed by other cells derived from brain tissue including astrocytes (13%, $p < 0.0001$) and oligodendrocytes (10%, $p < 0.0001$). Enrichments for terms under 'Abnormality of the cardiovascular system' were mainly associated with cardiac muscle cells (17%, $p < 0.0001$) and smooth muscle cells (10%, $p < 0.0001$). A similar proportion to smooth muscle cells was found for pancreatic stellate cells, though no significance was detected (11%, $p = 0.22$). Finally, enrichments for terms under 'Abnormality of the immune system' were overwhelmingly associated with immune cells ($p < 0.0001$).

Thus, for high level HPO annotations, the approach recovered highly significant proportions of expected phenotype-cell relationships, for the branches tested. The remainder of results analysis will also be

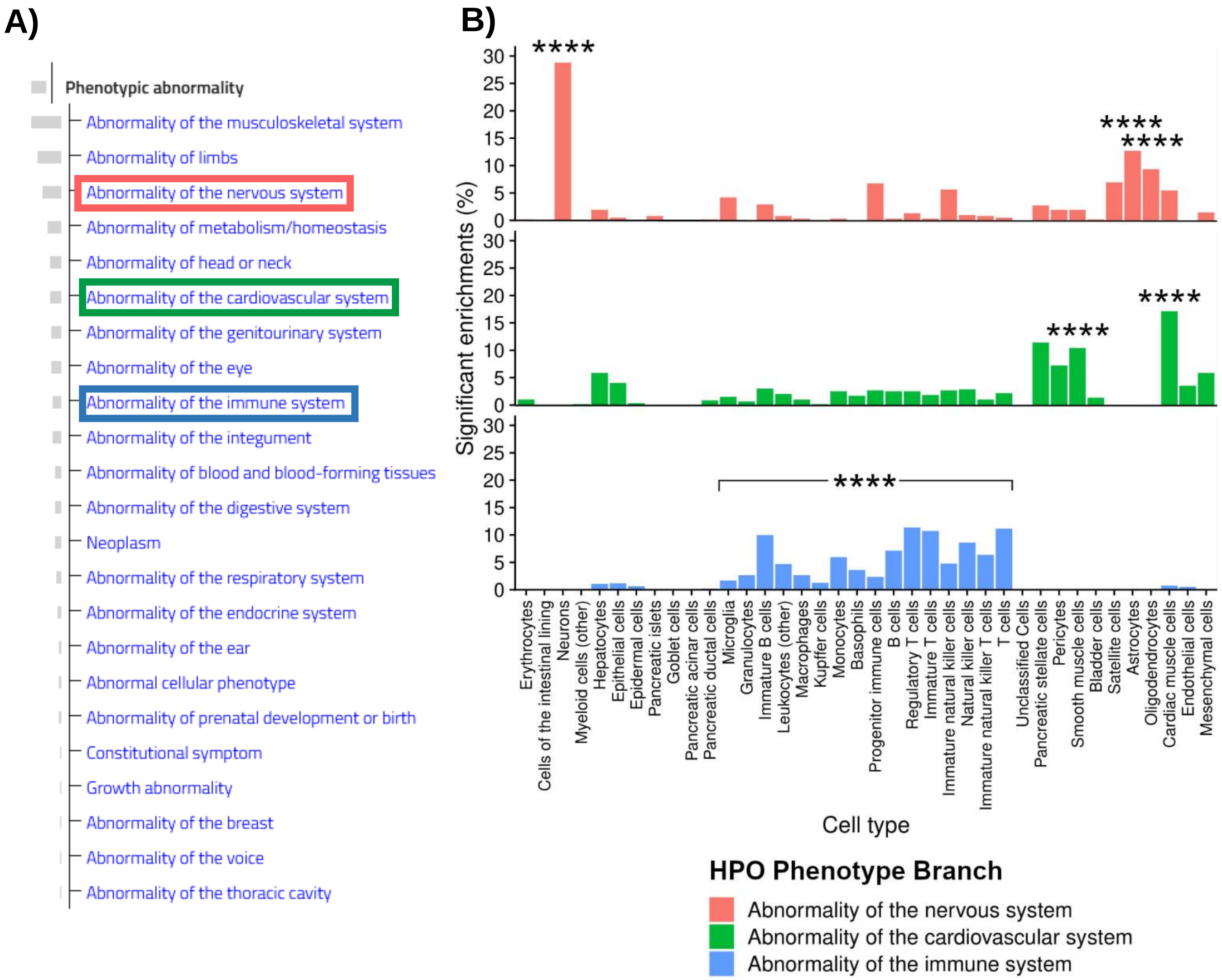


Figure 1 – Significant recovery of expected cell type enrichments according to HPO branches. Recovered cell type enrichments were sorted according to enrichments for phenotypes falling under the three HPO branches: 'Abnormality of the nervous system', 'Abnormality of the cardiovascular system', and 'Abnormality of the immune system'. **A)** The 23 highest level phenotype branches within the HPO are depicted. Branches focused on for this study are highlighted. Branch labels acquired from the HPO. (14) **B)** The percentage of significant enrichments for each of these three branches associated with each cell type is visualised. Total number of enrichments for each tree: nervous system – 622, cardiovascular system – 596, immune system – 942. Significant enrichments are defined by any cell type-phenotype relationship for which: $p < 0.05$ (n = 100,000) and fold-enrichment (fold increase of gene expression) > 1. Significance of HPO branch-cell relationships was calculated through additional hypergeometric tests comparing total branch enrichments across every cell type. Significance is indicated by asterisks (**** $p < 0.0001$).

demonstrated in the context of the three HPO branches utilised here.

Increasing Stringency for Significant Enrichments Recovers Higher Proportions of Expected Phenotype-Cell Associations

To further test that the approach accurately recovered expected cell type enrichments, it was hypothesised that increasing the stringency for defining a significant enrichment would result in a higher proportion of expected enrichments. For example, **Figure 1B** showed that the largest proportion of enrichments for 'Abnormality of the nervous system' terms was found in neurons (28%); by adjusting the cut-off *p*-value for which an enrichment is counted

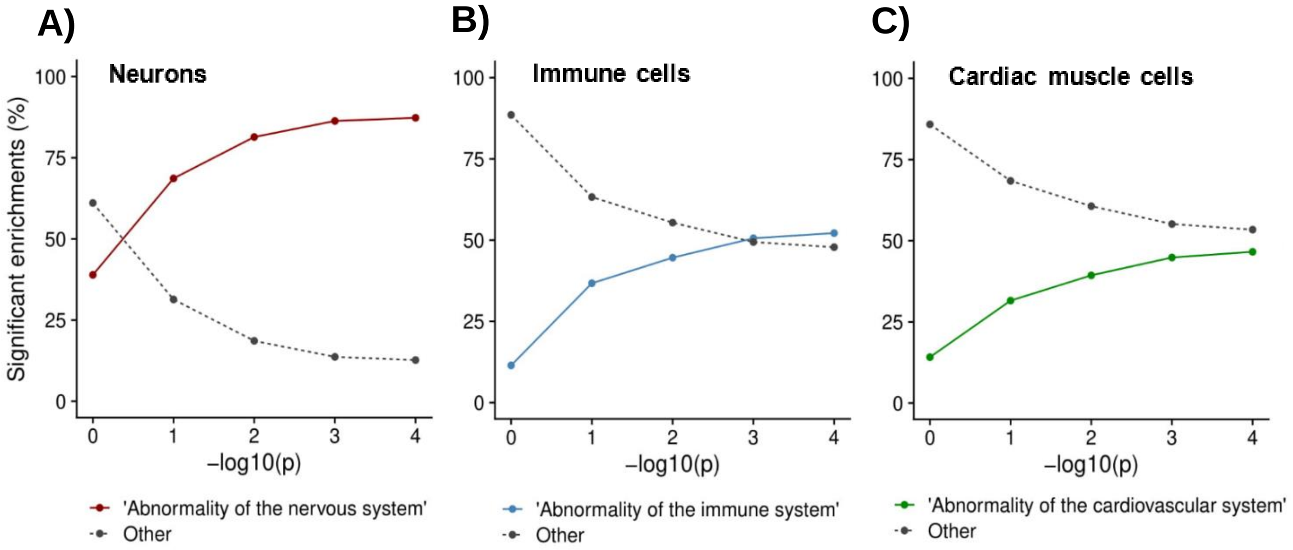


Figure 2 – Increasing stringency for defining a significant enrichment yields a higher proportion of expected cell type enrichments across multiple cell types and HPO term branches. The percentage of significant cell type enrichments is plotted for: **A)** Neurons + 'Abnormality of the nervous system' terms; **B)** Immune cells + 'Abnormality of the immune system' terms; **C)** Cardiac muscle cells + 'Abnormality of the cardiovascular system' terms, for decreasing *p*-value thresholds (n = 100,000). *p*-value threshold is represented logarithmically as -log10(*p*), e.g., $-\log_{10}(2) = 0.01$. In each graph, 'Other' represents any HPO term not falling into the respective branch.

as significant, it was predicted that a higher percentage of expected enrichments would be recovered, i.e., an increased number of significant neuron enrichments would originate in the 'Abnormality of the nervous system' branch.

Upon analysis, it was indeed found that at lower *p*-value cut-offs, the proportion of enrichments that fell into the aforementioned branch rose from ~70% ($p < 0.05$) to ~87.5% ($p < 0.0001$) (**Figure 2A**). This supported the hypothesis for this set of enrichments. The same was repeated for enrichments across all immune cells in the dataset (cell types visible in **Supplementary Figure 1**), analysing enrichment for descendant terms of 'Abnormality of the immune system'. Here, the respective proportion rose from ~40% ($p < 0.05$) to just over 50% ($p < 0.0001$) (**Figure 2B**). Finally, the same test was carried out for 'Abnormality of the cardiovascular system' terms with cardiac muscle cells, finding a rise from ~37.5% ($p < 0.05$) to ~48% ($p < 0.0001$) (**Figure 2C**).

In each case, reducing the cut-off p-value for defining a significant enrichment resulted in increased recovery of expected cell type enrichments. Therefore, it was shown that the approach utilised here tends towards recovery of accurate cell type enrichments in line with increased stringency, a promising result that strengthens confidence in the method.

Cell Type Enrichments are Stronger for More Specific HPO Terms

An additional test for the accuracy of EWCE in the approach was carried out. Generally, it is expected that phenotypes associated with fewer genes would be more specific to individual cell types. Therefore, it was hypothesised that significant enrichments associated with phenotypes found at lower levels of the HPO, which are by definition associated with fewer genes on average, would be stronger and therefore exhibit higher fold-enrichment scores.

Upon analysis, it was found that the mean fold increase of gene expression associated with significant enrichments rose exponentially as

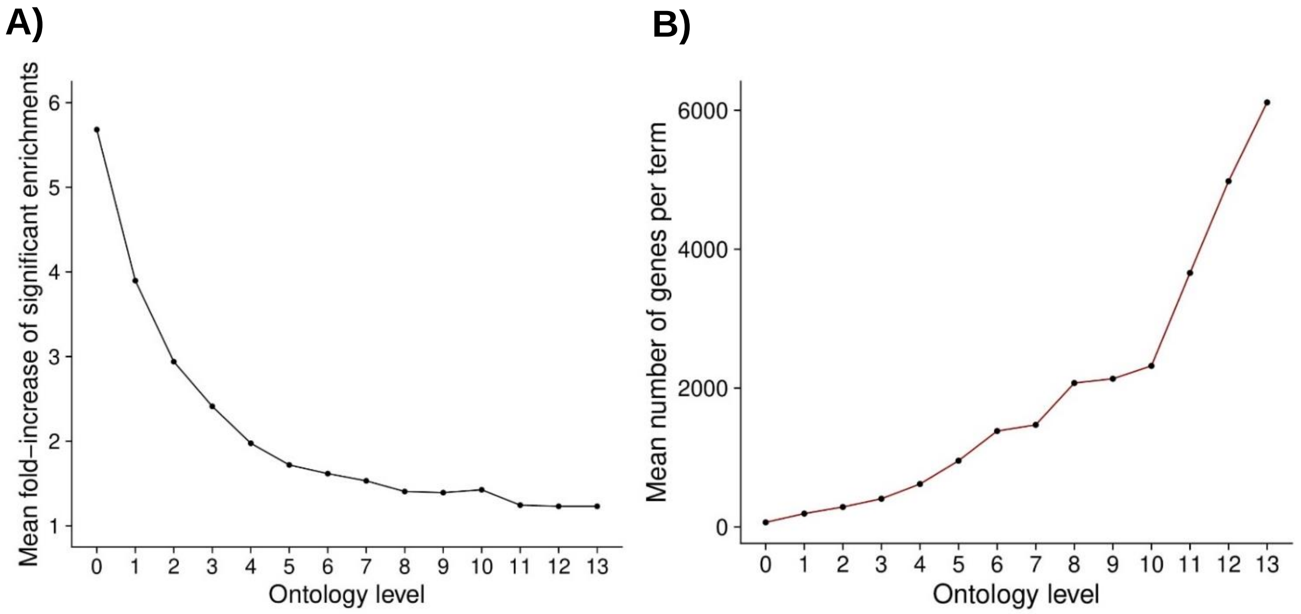


Figure 3 – More specific HPO phenotype terms are associated with stronger enrichments. A) Mean fold-increase of gene expression associated with a significant enrichment is plotted for enrichments at each HPO level. A 'level' of the HPO represents the number of sets of descendant terms falling below a given phenotype, e.g., level 0 represents most specific phenotypes. **B)** Mean number of genes per phenotype term is plotted for respective HPO levels. Significant enrichments are defined by any cell type-phenotype relationship for which: *p* < 0.05 (n = 100,000 and fold-enrichment (fold increase of gene expression) > 1.

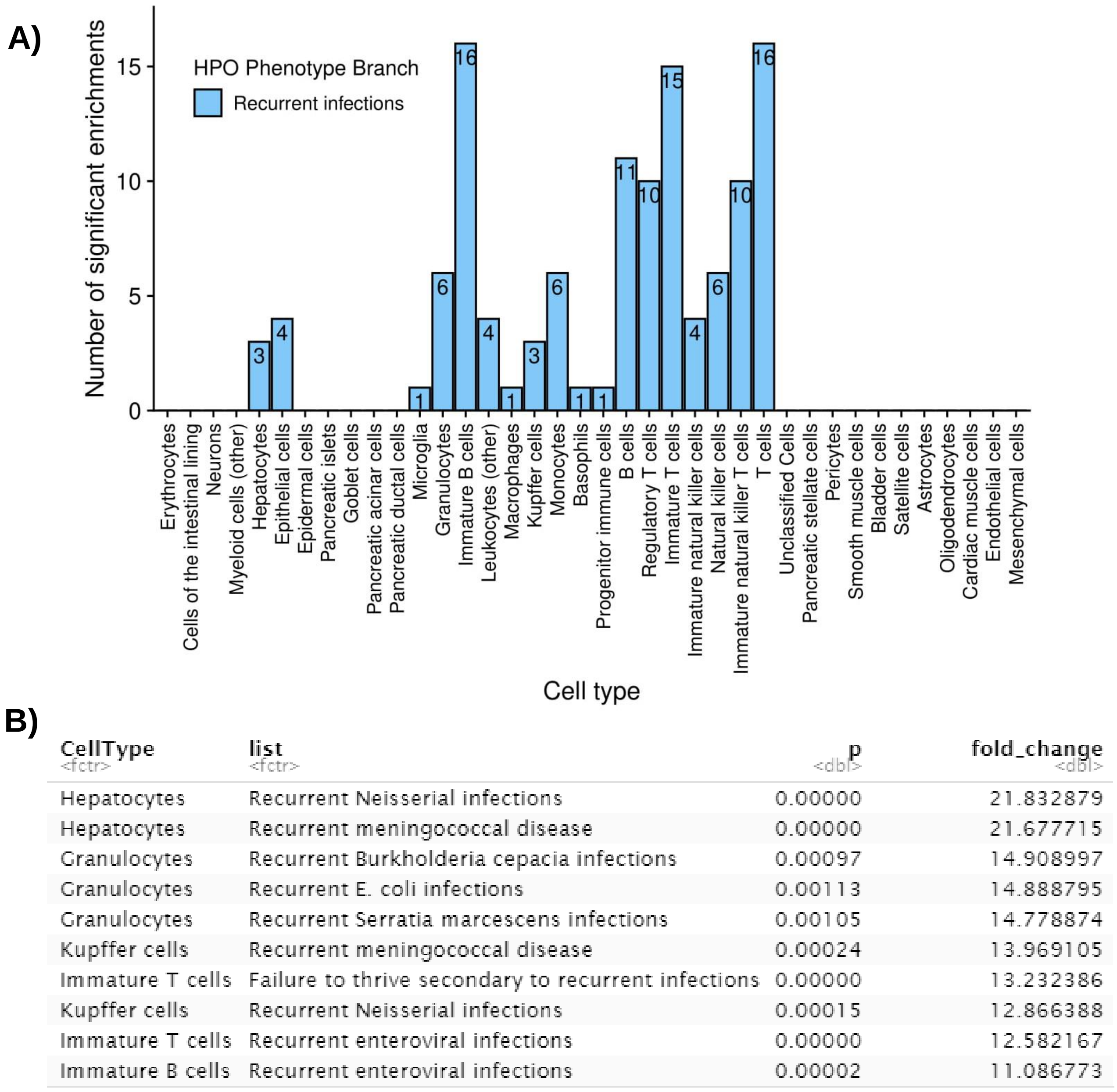
associated HPO terms were found at lower levels of the ontology (**Figure 3A**). The mean fold-enrichment score for broad terms from the top of the ontology was ~1.2, whereas leaf terms, those at the lowest level, had a mean fold-enrichment of ~5.7. **Figure 3B** confirms that terms from lower ontology levels are generally associated with fewer genes. Overall, these results confirm that EWCE was able to recover stronger enrichments for more specific phenotypes, meaning analysis could move on to low level, specific branches of the HPO.

High Granularity Cell Type Enrichment for Rare Disease Phenotypes within Specific HPO Branches

Following confirmational tests that the approach could recover expected cell type enrichments using high level HPO branches, it was sought to demonstrate the same using lower level HPO branches containing more specific terms, representative of rare disease phenotypes.

The main aim of gene enrichment analysis is to identify cell types potentially causal to disease, i.e., novel phenotype-cell associations. To build confidence that the approach utilised here could be utilised in this regard for future studies, it would first be expected for it to be able to accurately distinguish expected enrichments for precise phenotypes with known relationships to cells. Here, this was tested using a low-level HPO branch representing rare phenotypes expected to map to a high proportion of expected cell types, but at a higher granularity than previous examples.

A descendant branch of 'Abnormality of the immune system' found to have several significant enrichments for its own descendant terms was 'Recurrent infections'. This branch encompasses its own 78 unique descendant terms in the HPO, representing a range of phenotypes such as 'Recurrent bacterial infections' and 'Recurrent respiratory infections'. Significant enrichments involving descendant terms of 'Recurrent infections' were plotted across all cell types, with the expectation for the majority to map to immune cells, but an aim of investigating any outliers. As shown in **Figure 4A**, the majority of significant enrichments (*p* < 0.05) were indeed associated with immune cells (111 out of 118). However, 7 enrichments mapped to hepatocytes and epithelial cells. Further analysis showed that these were for the term 'Recurrent Neisserial infections' and its descendant term 'Recurrent meningococcal disease', which showed fold-enrichment



CellType <fctr>	list <fctr>	p <dbl>	fold_change <dbl>
Hepatocytes	Recurrent Neisserial infections	0.00000	21.832879
Hepatocytes	Recurrent meningococcal disease	0.00000	21.677715
Granulocytes	Recurrent Burkholderia cepacia infections	0.00097	14.908997
Granulocytes	Recurrent E. coli infections	0.00113	14.888795
Granulocytes	Recurrent Serratia marcescens infections	0.00105	14.778874
Kupffer cells	Recurrent meningococcal disease	0.00024	13.969105
Immature T cells	Failure to thrive secondary to recurrent infections	0.00000	13.232386
Kupffer cells	Recurrent Neisserial infections	0.00015	12.866388
Immature T cells	Recurrent enteroviral infections	0.00000	12.582167
Immature B cells	Recurrent enteroviral infections	0.00002	11.086773

Figure 4 – Significant enrichments for phenotypes descending from the HPO term 'Recurrent infections' across all cell types. A) The number of significant enrichments associated with HPO terms descending from the 'Recurrent infections' branch is plotted for each cell type. Total enrichments = 118. Significant enrichments are defined by any cell type-phenotype relationship for which: *p* < 0.05 (n = 100,000) and fold-enrichment (fold increase of gene expression) > 1. **B)** The top ten enrichments for descendant terms of 'Recurrent infections' across all cell types are listed in order of decreasing gene expression fold change. Each row represents an enrichment, showing cell type, phenotype (list), *p*-value and fold-enrichment (fold_change).

scores of 21.8 and 21.7, respectively (*p* < 0.0001) (**Figure 4B**).

‘Recurrent Neisserial infections’ is a descendant term of the low-level HPO branch ‘Recurrent gram-negative bacterial infections’, which encompasses 6 phenotypes overall. The aforementioned hepatocyte enrichment was investigated by plotting significant enrichments for each of these phenotypes across all cell types (**Figure 5**). It was found that the ‘Recurrent Neisserial infections’ term was unique in its hepatocyte enrichment (*p* < 0.0001), as all other terms in the branch, where significant, mapped solely to immune cells. Of note, the same term was enriched in Kupffer cells, a resident immune cell within the liver (*p* < 0.0001). Additionally, strong enrichments were found for ‘Recurrent Burkholderia cepacia infections’ and ‘Recurrent E. coli infections’ in granulocytes (*p* < 0.0001).

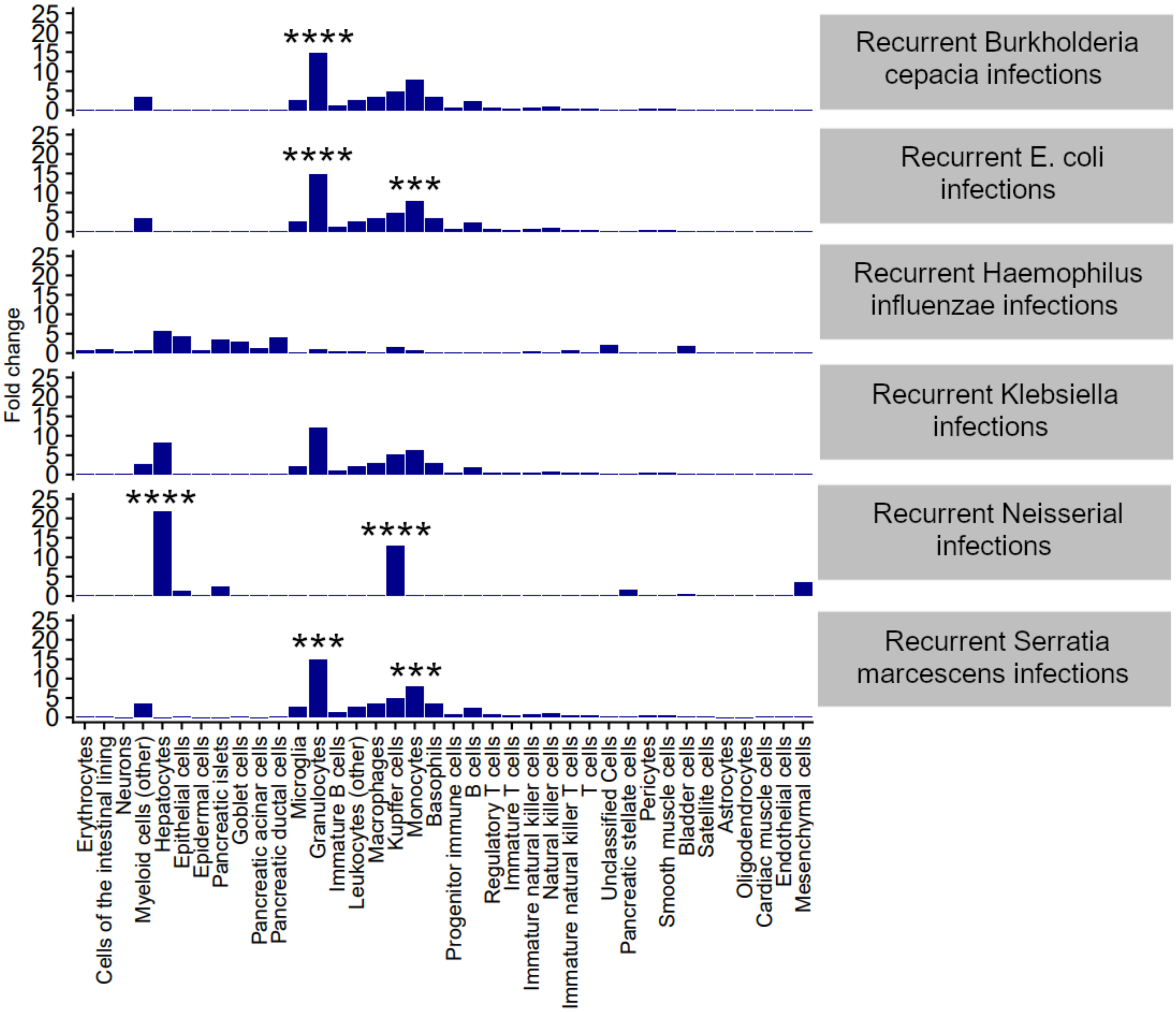


Figure 5 – Enrichments for phenotypes descending from the HPO term ‘Recurrent gram-negative bacterial infections’ across all cell types with significance annotated for individual cell-phenotype relationships. The fold-enrichment (fold increase of gene expression) associated with HPO terms descending from the ‘Recurrent gram-negative bacterial infections’ branch is plotted for each cell type, with each term displayed on the right. ‘Fold change’ indicates fold-increase of gene expression for respective phenotypes. Total significant enrichments = 7. Significance is indicated by asterisks (*** p < 0.001, **** p < 0.0001, n = 100,000).

The 6 terms visualised here share a similar profile of gene annotations within the HPO, with a difference of just one or two genes between terms. Despite this, there was variance in enrichment detection between cell types. Overall, this confirmed that the approach was able to distinguish enrichments for rare phenotypes at a precise level.

DISCUSSION

As single cell sequencing has seen vast improvements over recent years, more opportunities to investigate the precise cell types underlying disease have arisen. (19) The EWCE method was previously demonstrated to be successful in identifying cell types underlying diseases by utilising scRNAseq in combination with genetic susceptibility data, but thus far had only been performed using data for polygenic diseases obtained from GWAS summary statistics. (8) Rare diseases, which are commonly associated with single mutations, remain poorly understood in many cases and have so far been under-represented in gene enrichment studies. Therefore, an opportunity was presented for approaches such as gene enrichment analysis to be extended to also cover rare disease studies. Here, promising results have been generated to suggest that EWCE may be used to elucidate cell types underlying a wide range of rare disease phenotypes in a high-throughput manner. This demonstrates that considering rare diseases as clusters of phenotypes for such studies may provide an opportunity to bypass the need to analyse rare diseases individually, and that focusing on the cells underlying symptoms manifesting throughout multiple disorders may be a viable alternative.

EWCE was ran using scRNAseq data from *Tabula Muris* in combination with gene lists associated with disease phenotypes derived from the HPO, which contains data for thousands of phenotype-gene annotations, including those for rare, mendelian diseases. Enrichment was iteratively tested across thousands all constituent phenotypes. Overall, 5,509 significant cell type enrichments were detected (**Supplementary Figure 2**), and it was demonstrated that the EWCE method was able to accurately recover expected cell type enrichments across a range of HPO levels, including at its lowest level of terms where as few as one gene is associated with a phenotype. (14)

It was first sought to obtain confirmation that expected phenotype-cell relationships were recovered. This was demonstrated using three representative branches of HPO terms: ‘Abnormality of the nervous system’, ‘Abnormality of the immune system’ and ‘Abnormality of the cardiovascular system’. It was hypothesised that the majority of enrichments for each of these HPO categories would map to cells of the brain, immune system, and cardiac tissue, respectively. This was shown to indeed be the case, and each expected relationship was found to be represented at a very high degree of significance ($p < 0.0001$). (**Figure 1B**).

Next, several tests were devised to further investigate the accuracy of the approach. It was hypothesised that increasing the stringency for defining significant enrichments would yield higher proportions of expected enrichments. This was found to be true for both neurons and immune cells, which saw increased proportions of enrichments for nervous and immune system phenotypes, respectively, as the *p*-value threshold for significance was reduced (**Figure 2**). Following this, it was confirmed that lower level HPO terms, which represent more specific phenotypes associated with fewer genes, had stronger enrichments recovered through EWCE (**Figure 3**). This result was of particular importance for this preliminary study, as a desired outcome of accurate enrichment analysis is for short, specific gene lists to map to distinct cell types strongly. This is especially true when considered in the context of single cell sequencing technology improvements, which over the coming years will continue to distinguish cell types at the molecular level. (20)

Confirmation was next required that the approach could recover accurate enrichments for precise rare disease phenotypes. Here, this was demonstrated by analysing enrichments for terms at low, specific levels of the HPO. The 'Recurrent infections' branch of terms was chosen as it is a constituent of 'Abnormality of the immune system', and enrichments for its 78 descendant terms were visualised across all cell types in the dataset. Whilst the majority were expectedly associated with immune cells, several were flagged for hepatocytes and epithelial cells, including a very strong enrichment for 'Recurrent Neisserial infections' in hepatocytes ($p < 0.0001$, **Figure 4**). Investigating this further, it was found that this enrichment was unique for this branch of terms ('Recurrent gram-negative bacterial infections') (**Figure 5**). At first, this appeared as a potentially novel phenotype-cell relationship, however, published literature has demonstrated that Neisserial infections, which primarily involve meningococcal bacteria, are generally a result of complement deficiencies. (21) In line with this, complement proteins are known to be mainly synthesised in the liver. (22) Therefore, this rather demonstrates a very strong example of an accurate enrichment recovery for a specific rare phenotype, supporting the primary goal of this study. Of note, Kupffer cells, which are resident immune cells in the liver, were also enriched for this term; and the next descendant term 'Recurrent meningococcal disease' was also highly significantly enriched in hepatocytes ($p < 0.0001$, **Figure 4B**) (23). Additionally, the same set of results highlighted very strong enrichments in granulocytes for both 'Recurrent E. coli infections' and 'Recurrent Burkholderia cepacia infections' ($p < 0.0001$). These phenotypes have been demonstrated to be associated with granulomatous disease, again confirming the recovery of accurate, precise cell type enrichments for rare phenotypes. (24)

Since its introduction, gene set enrichment analysis has served as an important tool to aid GWAS by reducing the prevalence of false positive results for disease-associated SNPs. (25) The baseline approach of integrating gene set enrichment analysis with GWAS works off the assumption that SNPs associated with a particular disease phenotype inherently act through common pathways. This is similar to the approach of this study, whereby individual genetic variants have been used to generate target gene lists for enrichment analysis in EWCE, with the presumption that variants may act by common cell types. The hypothesis set in this study was that considering diseases as clusters of phenotypes could be a valid approach to elucidating cell type enrichments, and these preliminary results have suggested that this is indeed the case. Overall, this presents a method whereby further study and therapeutics can take a phenotypic approach to studying causal cell types in rare diseases.

Although the results of this study suggest that the approach utilised here is valid, as a high proportion of expected enrichments were recovered, enrichment analysis methods such as EWCE introduce limitations and biases that can be easy to overlook. For example, Simillion et al. described a 'sample source bias' inherent to gene set enrichment analysis, particularly when enrichment scores are based on limited background gene sets. (26) Though EWCE differs from traditional gene set enrichment analysis, a similar bias may be present in this study as the background gene set utilised was based on just one dataset (*Tabula Muris*). Single cell RNA sequencing improves upon traditional RNA sequencing by allowing accurate gene expression measurements with fewer cells, but technical limitations can still result in under-representation of genes in many datasets. (27) In addition, cell types expressing a larger proportion of the genome may be misunderstood as highly enriched for certain phenotypes. An example of this may be seen in this study, where pancreatic stellate cells were found to have the most enrichments out of all results (**Supplementary Figure 2**). Further analysis of this cell type found no trend in enrichments, and in fact 0 enrichments were detected for terms falling under 'Abnormality of the pancreas' (**Supplementary Figure 3**). Additionally, despite representing a high proportion of cardiovascular phenotype enrichments, pancreatic stellate cells were not significantly linked to this HPO branch (**Figure 1B**). Though it is possible that this cell type represents an interesting target for further study, it may be the case that increased weighting has been added to its enrichments as a result of sample bias, resulting in a high number of false positives.

An additional limitation of this study is the use of scRNAseq data obtained from mice. Though mice are commonly used in human disease research due to genetic similarity to humans, studies aiming to utilise transcriptomic data to signpost potential disease treatments, such as gene enrichment studies, may benefit more so from the usage of human data. (28) This is especially true in the case of developing therapeutics. For example, a 2017 review comparing human and mouse transcriptomics highlighted that fewer than 8% of animal models for cancer treatments translate successfully to human applications, on average. (29) Additionally, out of the 18995 genes that have orthologs between humans and mice, just 16470 are 1:1 orthologs. (30)

Overall, this study has generated preliminary results to suggest that EWCE may be applied to rare disease-gene associations to elucidate potentially causal cell types. It has been demonstrated that the approach utilised here can accurately recover expected cell type enrichments, utilising gene lists derived from the HPO that vary greatly in length. Immediate future work will involve working through the 5,509 significant enrichments recovered to identify potentially novel cell-phenotype relationships for further study. Though this was not explored here, the scope of accurate results identified implies that any results not fitting into current literature may have a realistic precedent for being targets of interest. In the long run, increasing resolution and granularity in scRNAseq libraries will allow more precise identification of casual cell types in disease, and the EWCE method may prove to be an invaluable tool in this regard. The reference gene expression atlas utilised in this study (*Tabula Muris*) represents just 120 cell types, which here were further grouped to make 38 cell classifications. This decision was based on the early nature of the study but resulted in reduced granularity of cell type enrichments. Further usage of EWCE would benefit from the use of transcriptomic datasets representing broader classifications of cell types that may have been missing here, e.g., osteoblasts. This data will rapidly become available through the development of the Human Cell Atlas. (31) In the far future, the goal of such analyses would be to signpost potential therapeutic targets in genetic disease.

CONCLUSION

In summary, promising results have been generated to suggest that EWCE may be successfully applied to the identification of causal cell types in rare disease phenotypes, supporting our initial hypothesis. Through integrating scRNAseq with genetic susceptibility data derived from the HPO, it has been demonstrated that EWCE can accurately recover cell type enrichments across a range of rare phenotypes. Going forward, it is envisioned that in the future, rare disease patients may have causal mutations identified early in life, and that approaches such as EWCE will be used to predict the phenotypes they may present based on cellular expression.

ETHICAL CONSIDERATIONS

There were no ethical considerations of note for this project. HPO genetic susceptibility data is anonymously collated.

ACKNOWLEDGMENTS

This research was completed under the Neurogenomics research group at the UK Dementia Research Institute, Imperial College London. I thank Dr Nathan G. Skene and Brian M. Schilder for their support during this project.

References

(1) Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. *Nature reviews. Nephrology.* 2018; 14 (8): 479-492. Available from: doi: 10.1038/s41581-018-0021-7 Available from: https://www.ncbi.nlm.nih.gov/pubmed/29789704 .

(2) Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome biology.* 2017; 18 (1): 83. Available from: doi: 10.1186/s13059-017-1215-1 Available from: https://www.ncbi.nlm.nih.gov/pubmed/28476144 .

(3) Nomura S. Single-cell genomics to understand disease pathogenesis. *Journal of human genetics.* 2021; 66 (1): 75-84. Available from: doi: 10.1038/s10038-020-00844-3 Available from: https://www.ncbi.nlm.nih.gov/pubmed/32951011 .

(4) Mills MC, Rahal C. A scientometric review of genome-wide association studies. *Communications biology.* 2019; 2 (1): 9. Available from: doi: 10.1038/s42003-018-0261-x Available from: https://www.ncbi.nlm.nih.gov/pubmed/30623105 .

(5) Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, et al. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Scientific reports.* 2019; 9 (1): 16844. Available from: doi: 10.1038/s41598-019-53111-7 Available from: https://www.ncbi.nlm.nih.gov/pubmed/31727947 .

(6) Borecki IB, Province MA. Genetic and Genomic Discovery Using Family Studies. *Circulation.* 2008; 118 (10): 1057-1063. Available from: doi: 10.1161/CIRCULATIONAHA.107.714592 Available from: http://circ.ahajournals.org/cgi/content/extract/118/10/1057 .

(7) Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences - PNAS.* 2005; 102 (43): 15545-15550. Available from: doi: 10.1073/pnas.0506580102 Available from: https://www.jstor.org/stable/4143472 .

(8) Skene NG, Grant SGN. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Frontiers in neuroscience.* 2016; 10 16. Available from: doi: 10.3389/fnins.2016.00016 Available from: https://www.ncbi.nlm.nih.gov/pubmed/26858593 .

(9) Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell (Cambridge).* 2017; 169 (7): 1177-1186. Available from: doi: 10.1016/j.cell.2017.05.038 Available from: http://dx.doi.org/10.1016/j.cell.2017.05.038 .

(10) Rahit, K M Tahsin Hassan, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. *Genes.* 2020; 11 (3): 239. Available from: doi: 10.3390/genes11030239 Available from: https://www.ncbi.nlm.nih.gov/pubmed/32106447 .

(11) Bhattacharyya NP. Huntington’s disease: a monogenic disorder with cellular and biochemical complexities. *The FEBS journal.* 2008; 275 (17): 4251. Available from: doi: 10.1111/j.1742-4658.2008.06560.x Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-4658.2008.06560.x .

(12) Augustine EF, Adams HR, Mink JW. Clinical Trials in Rare Disease. *Journal of child neurology.* 2013; 28 (9): 1142-1150. Available from: doi: 10.1177/0883073813495959 Available from: https://journals.sagepub.com/doi/full/10.1177/0883073813495959 .

(13) Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? *Nature reviews. Drug discovery.* 2020; 19 (2): 77-78. Available from: doi: 10.1038/d41573-019-00180-y Available from: https://www.ncbi.nlm.nih.gov/pubmed/32020066 .

(14) Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. *Nucleic acids research.* 2021; 49 (D1): D1207-D1217. Available from: doi: 10.1093/nar/gkaa1043 Available from: https://www.ncbi.nlm.nih.gov/pubmed/33264411 .

(15) Schaum N, Neff NF, May AP, Quake SR, Darmanis S, Batson J, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature (London).* 2018; 562 (7727): 367-372. Available from: doi: 10.1038/s41586-018-0590-4 Available from: https://www.ncbi.nlm.nih.gov/pubmed/30283141 .

(16) R Core Team (2013). *R: A language and environment for statistical computing.* (4.0.3) R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/: 2020.

(17) Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. *Human Phenotype Ontology 'phenotype_to_genes.txt'*
. Available from: https://hpo.jax.org/app/download/annotation [Accessed December 2020].

(18) Yoav Benjamini, Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, Methodological.* 1995; 57 (1): 289-300. Available from: doi: 10.1111/j.2517-6161.1995.tb02031.x Available from: https://www.jstor.org/stable/2346101 .

(19) Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. *Cell & bioscience.* 2019; 9 (1): 53. Available from: doi: 10.1186/s13578-019-0314-y Available from: https://www.ncbi.nlm.nih.gov/pubmed/31391919 .

(20) Yuan F, Pan X, Zeng T, Zhang Y, Chen L, Gan Z, et al. Identifying Cell-Type Specific Genes and Expression Rules Based on Single-Cell Transcriptomic Atlas Data. *Frontiers in bioengineering and biotechnology.* 2020; 8 350. Available from: doi: 10.3389/fbioe.2020.00350 Available from: https://www.ncbi.nlm.nih.gov/pubmed/32411685 .

(21) Mollah F, Tam S. *Complement Deficiency.:* StatPearls; 2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557581/

(22) Merle NS, Church SE, Fremeaux-Bacchi V, Roumenina LT. Complement System Part I - Molecular Mechanisms of Activation and Regulation. *Frontiers in immunology.* 2015; 6 262. Available from: doi: 10.3389/fimmu.2015.00262/full Available from: https://www.ncbi.nlm.nih.gov/pubmed/26082779 .

(23) Dixon LJ, Barnes M, Tang H, Pritchard MT, Nagy LE. *Kupffer Cells in the Liver.* Hoboken, NJ, USA: John Wiley & Sons, Inc; 2013.

(24) Roos D. Chronic granulomatous disease. *British medical bulletin.* 2016; 118 (1): 50-63. Available from: doi: 10.1093/bmb/ldw009 Available from: https://www.narcis.nl/publication/RecordID/oai:pure.amc.nl:publications%2F9816d535-1896-48e7-bfbd-d112729005fe .

(25) Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics.* 2008; 24 (23): 2784-2785. Available from: doi: 10.1093/bioinformatics/btn516 Available from: https://www.ncbi.nlm.nih.gov/pubmed/18854360 .

(26) Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC bioinformatics.* 2017; 18 (1): 151. Available from: doi: 10.1186/s12859-017-1571-6 Available from: https://www.ncbi.nlm.nih.gov/pubmed/28259142 .

(27) Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics (Oxford, England).* 2018; 19 (4): 562-578. Available from: doi: 10.1093/biostatistics/kxx053 Available from: https://www.ncbi.nlm.nih.gov/pubmed/29121214 .

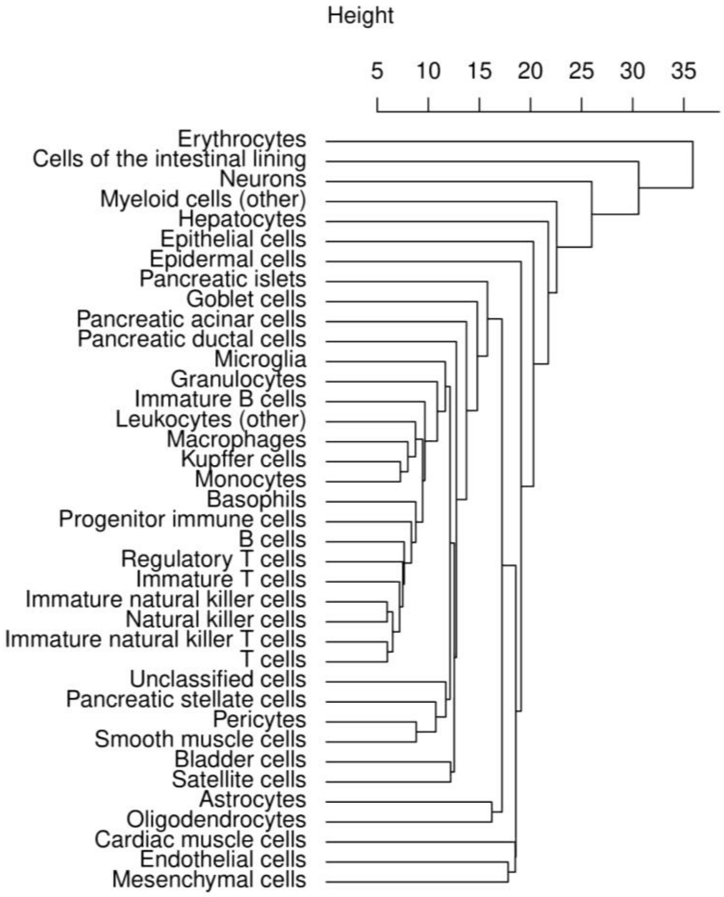
(28) Bryda EC. The Mighty Mouse: the impact of rodents on advances in biomedical research. *Missouri medicine.* 2013; 110 (3): 207-211. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23829104 .

(29) Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. *Nature reviews. Genetics.* 2017; 18 (7): 425-440. Available from: doi: 10.1038/nrg.2017.19 Available from: https://www.ncbi.nlm.nih.gov/pubmed/28479595 .

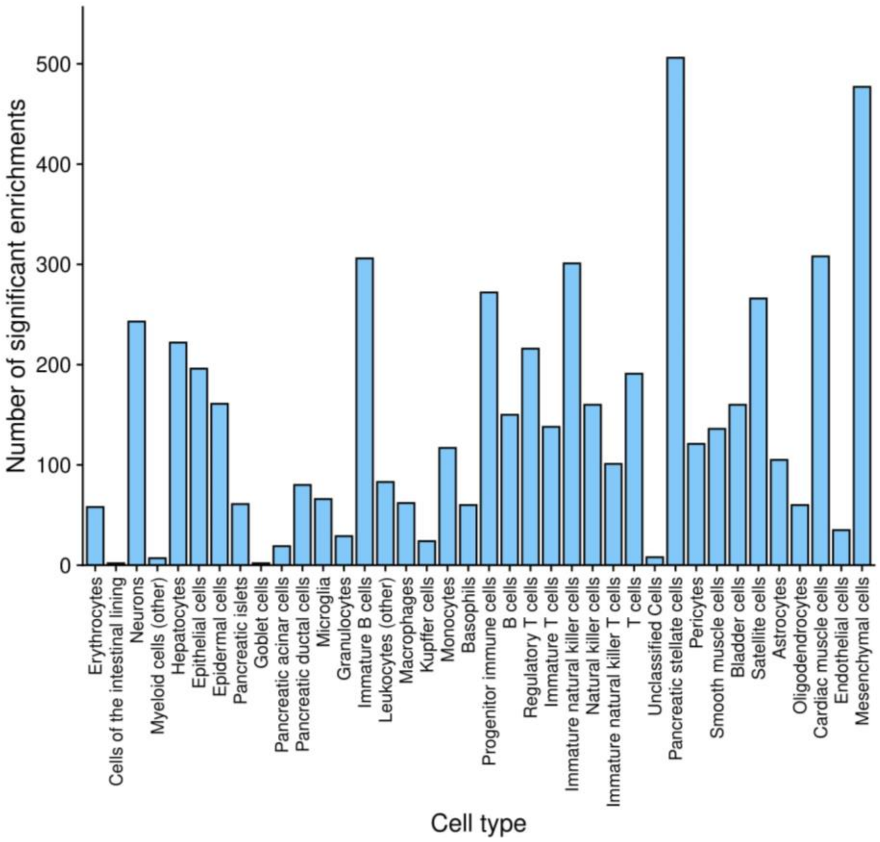
(30) Skene N. *One2One: an R package for recovering 1:1 orthologs based on MGI homology data*
. Available from: https://github.com/NathanSkene/One2One.

(31) Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. *eLife.* 2017; 6 Available from: doi: 10.7554/eLife.27041 Available from: https://search.proquest.com/docview/1992867111 .

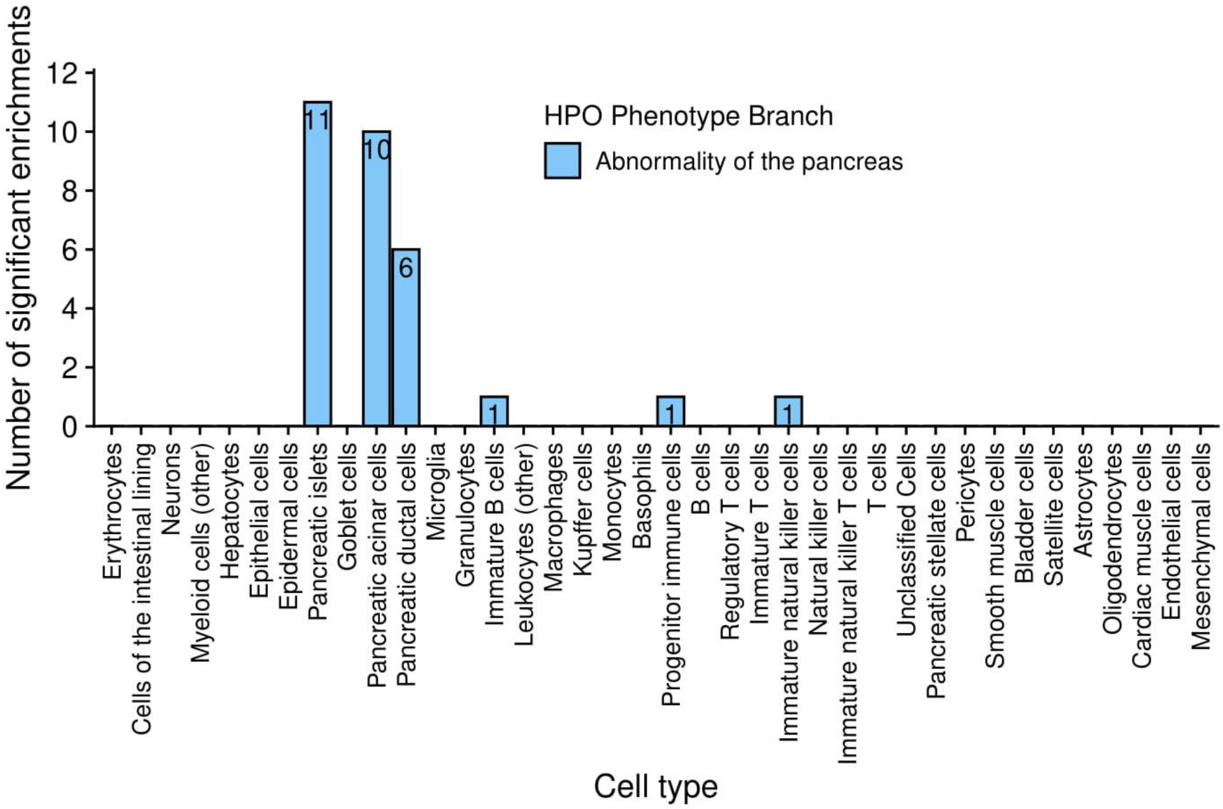
SUPPLEMENTARY FIGURES



Supplementary Figure 1 – Dendrogram of final cell type classifications. A hierarchal clustering algorithm (*hclust*, R version 4.0.3) was ran utilising relative gene specificity matrices for each cell type. Cells are arranged according to results of the algorithm, where closer cells have more similar gene specificity profiles. Height comparatively indicates similarity of cell types according to divergence.



Supplementary Figure 2 – Total numbers of significant enrichments across all cell types. The number of significant cell type enrichments is plotted for each cell type in the dataset. Total significant enrichments = 5,509. Significant enrichments are defined by any cell type-phenotype relationship for which: $p < 0.05$ and fold-enrichment (fold change of gene expression) > 1.



Supplementary Figure 3 – Enrichments for HPO terms under ‘Abnormality of the pancreas’ are not associated with pancreatic stellate cells. The number of significant enrichments associated with HPO terms descending from the ‘Abnormality of the pancreas’ branch is plotted for each cell type. Total enrichments = 118. Significant enrichments are defined by any cell type-phenotype relationship for which: $p < 0.05$ fold-enrichment (fold increase of gene expression) > 1.