Genetic identification of cell types underlying mammalian phenotypes

Name: Lusheng Li, Applied Genomics
Supervisor: Dr Nathan Skene, Department of Brain Sciences, Imperial College London

(Note: This article only contains introduction and results part)

Abstract

Rare diseases are one of the biggest challenges in medicine due to limited understanding and lack an effective treatment for most rare diseases. The phenotypic abnormalities is one of the diagnostic criteria and the mammalian phenotypes could be a powerful tool to support disease research. Most phenotypes are clinical features or symptoms of rare diseases and in most cases, cell types relevant to those rare diseases will be currently unknown. The identification of cell types is essential to improve understanding, diagnosis, and treatment of diseases. Here, we identified the associations between cell types and phenotypes using the Tabula Muris scRNA-seq dataset and the genotype-to-phenotype annotations from Mouse Genome Database (MGD) and Human Phenotype Ontology (HPO). The association results showed that the expected cell types were associated with the main MPO and HPO branches and low ontology level phenotypes were associated with novel cell types. The connections between phenotypes and diseases were constructed to explore the cell-phenotype-disease associations. We investigated the cell types and phenotypes that connected to progressive myoclonus epilepsy (DOID:891) and Wolfram syndrome. In addition, the cross-species phenotypes were compared and analysed to support translational research. To better maximise the value of these association results, the web-based application (http://www.cell-phenotype.com/) was developed for clinicians and researchers to retrieve and explore the cell type-phenotype associations, cell type-phenotype-disease associations, and the cross-species phenotypes. We expected the association results to improve the understanding of diseases and develop novel therapies targeting rare disease symptoms.

Introduction

Despite the name, it is estimated that there are over 6000 rare diseases. Based on analysis of Orphanet (Pavan *et al.*, 2017), a database for providing information on rare diseases, a conservative estimate for the population prevalence of rare diseases is 3.5-5.9%, which equates to 263–446 million persons affected globally (Nguengang Wakap *et al.*, 2020). More than 90% of rare diseases still lack an effective treatment due to high development costs and low potential for return-on-investment (Kaufmann, Pariser and Austin, 2018). According to OMIM (Amberger *et al.*, 2015), a comprehensive, authoritative database provides the information about human genes and genetic phenotypes related to all known mendelian disorders, most phenotypes are associated with 1~3 genes and 6163 phenotypes and 4308 genes are included in the single gene disorders and traits category (https://www.omim.org/statistics/geneMap), which indicates most rare disease are caused by genetic conditions. Accurate diagnosis of rare disease is one of the key challenges due to

complex genetic conditions, pathogenic mechanisms and the lack of relevant professional knowledge of healthcare professionals. Many patients have been misdiagnosed or undiagnosed, leading to delays in treatment and exacerbation of their condition. The phenotypic abnormalities is one of the diagnostic criteria (Slavotinek, 2002; Birgfeld *et al.*, 2011) and the phenotypic abnormalities have been established to diagnose the diseases. The Orphanet and human phenotype ontology (HPO) are typically leveraged to retrieve information about symptoms and gene variants of rare diseases. However, in some cases the HPO does not have enough information to support the practical clinical application or research. As a classical model organism, the mice have been widely and intensively investigated, and numerous genotype-to-phenotype annotations have been discovered and accumulated.

The Mouse Genome Database (MGD, http://www.informatics.jax.org/) is the community model organism database for the laboratory mouse, providing comprehensive information about mouse gene function, genotype-to-phenotype annotations, and mouse models of human disease (Bult *et al.*, 2019). MGD integrates mouse phenotype annotations using the Mammalian Phenotype Ontology (MPO) and the mouse phenotypes are associated with genotypes. The genotype-to-phenotype annotations were collected and integrated from published phenotyping studies of transgenic mice. The mammalian phenotypes could be a powerful tool to complement HPO and play its unique roles. The corresponding mammalian phenotypes of human disease can provide the information on human orthologue genes that are associated with disease based on phenotypic similarity analysis, and the genetic information could be a reference or guild for disease genetic diagnosis and further analysis when the pathogenic gene is currently unknown. In addition, the rare genetic variant on the genome of a patient can be used to build the mouse model to investigate the mechanism and symptoms of disease using gene-editing techniques. The observed and recorded mammalian phenotypes can be used for supporting the disease research.

The identification of cell types is essential to improve understanding, diagnosis, and treatment of diseases (Mathys *et al.*, 2019; Bryois *et al.*, 2020), which will enable precision therapies for the specific cell types associated with specific diseases. The cell types associated with disease could be the potential markers for the diagnosis, progression and prognosis of disease based on the cell states and gene specific expression. Single-cell sequencing techniques provide information about fundamental features of cell types, including molecular profiles, cell states, state transitions, and cell−cell interactions (Panina *et al.*, 2020). Many of the phenotypes are symptoms of rare disease syndromes and in most cases, cell types relevant to those rare diseases will be currently unknown.

Here, we used the Tabula Muris, a single cell transcriptomic dataset containing nearly 100,000 cells from 20 mouse organs (Tabula Muris Consortium *et al.*, 2018) and the genotype-to-phenotype annotations from MGD to identify the association between cell types and phenotypes using Expression Weighted Celltype Enrichment (EWCE) (Skene and Grant, 2016; Tabula Muris Consortium *et al.*, 2018). EWCE can be applied to map the genes associated with each MPO phenotype onto cell types in which the genes are enriched for expression specificity. In order to compare the difference between MPO and human phenotype ontology (HPO), the genotype-to-phenotype annotations from HPO were used for enrichment analysis with 38 broad cell types from Tabula Muris. We found that the expected cell types were associated with the main MPO and HPO branches, and low ontology level

phenotypes tended to have higher specific expression in cells and low ontology level phenotypes were associated with expected and novel cell types. The cross-species phenotypes between MPO and HPO were analysed for supporting translational study. Because most phenotypes are the clinical features and symptoms of disease, the connection between mammalian phenotypes and human diseases comes from HMDC and the connection between human phenotypes and human diseases comes from MONDO database were used to explore the cell type-phenotype-disease association. To better maximise the value of these association results, the web-based application (http://www.cell-phenotype.com/) was developed to retrieve and explore the cell type-phenotype associations, cell type-phenotype-disease associations, and the cross-species phenotypes. We expected the association results to be of interest to clinicians, researchers and pharmaceutical companies looking to develop novel therapies targeting rare disease symptoms.

Analysis workflow

Figure 1 illustrates the analysis workflow in our study. The first is dataset preparation - downloading data from online public resources and processing it to make it compatible for use with EWCE. The CTD file (TabulaMuris_n.rds) was generated using the EWCE::generate_celltype_data function, using the Tabula Muris scRNA-seq dataset. The genotype-to-phenotype annotations were derived from MGD and HPO. Next, the EWCE::bootstrap_enrichment_test function was performed to generate the cell type-phenotype associations. To get credible and statistically significant association results, bootstrapping was employed by randomly generating 10,000 gene lists. The code for the enrichment test is available at (https://github.com/LushengLi9909/Cell-mpheno/blob/main/R/mpo_gen_results_level1.R). Due to the enrichment analysis with 100000 repetitions performed on several thousand phenotypes, significant computational resources were required. The analysis was performed on the Imperial College High Performance computing (HPC) server with 8 core CPU and 96 GB random-access memory (RAM). The cell type-phenotype association results were integrated and merged into a .rds file for subsequent analysis and interpretation. The cross-species phenotypes that derived from uPheno were used to compare the difference of cell types that were associated with mammalian phenotypes and human phenotypes. Because most phenotypes are features of human diseases and research of mammalian phenotype in the human disease, the mammalian phenotypes that were significantly associated with cell types were connected to human disease based on the connection between mammalian phenotypes and human diseases from HMDC. MONDO provided the connection between human phenotypes and diseases. To better maximise the value of these association results, the shiny web-based application was developed to retrieve and explore the cell type-phenotype association, cell type-phenotype-disease association, and cross-species phenotypes.

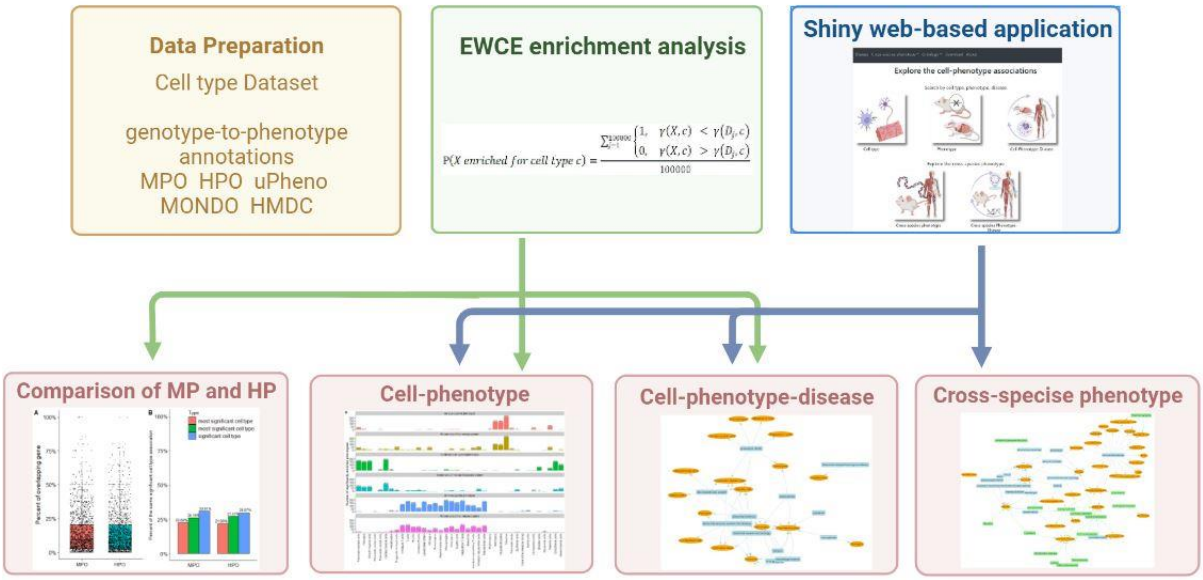


Figure 1. Workflow. The analysis workflow includes prerequisite data preparation, EWCE enrichment analysis, the visualisation and interpretation of cell type-phenotype association and cell type-phenotype-disease association, the comparison of MPO and HPO, and shiny web-based application development.

Results

In this study, 8,353 MPO phenotypes and 6,419 HPO phenotypes were used for EWCE enrichment analysis with the Tabula Muris scRNA-seq dataset clustered and annotated into 38 broad cell types. The enrichment analysis results revealed 14,787 significant cell type-phenotype associations using mammalian phenotypes and 8,677 significant cell type-phenotype associations using human phenotypes ($q < 0.05$ and fold change > 1). The number of significantly enriched phenotypes for each cell type can be seen in Figure 2B-C. Overall, the mammalian phenotypes have more significant associations than human phenotypes, which suggests that mammalian phenotypes could provide more comprehensive information for application in research. Pancreatic stellate cells have the most phenotype associations in MPO and HPO:807 and 711, respectively. We would expect cell types which are associated with more phenotypes to represent good targets for cell-targeted therapy development. The dendrogram in Figure 2A represents the relationship between cell types, and was used to group the 38 broad cell types into categories. In addition, some cell types come from multiple organs, such as Mesenchymal cells, B cells, T cells, Macrophages, Leukocytes, Endothelial cells, Endothelial cells and Myeloid cells. This indicates that these cell types may play a role in multiple organs and tissues. Most cell types that are related to multiple organs are immune cells, which reflects the universality and diversity of the immune system.

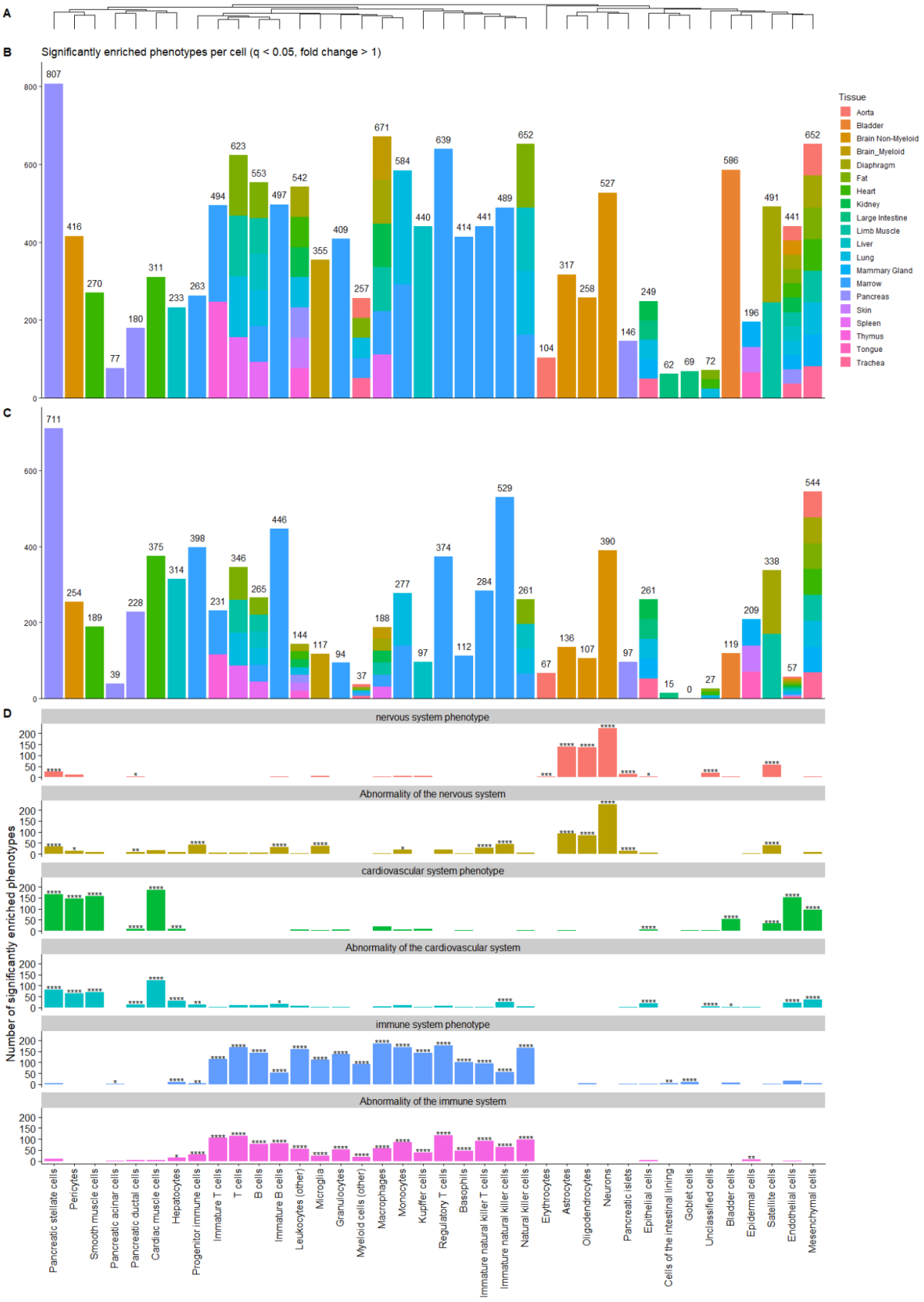


Figure 2. (A) The dendrogram of 38 cell types from the Tabula Muris. (B) The number of significantly enriched phenotypes for each cell type from MPO (q < 0.05, fold change > 1). The legend only indicates the organ type from which the cells are derived, not the proportion

of the cell type. (C) The number of significantly enriched phenotypes for each cell type from HPO ($q < 0.05$, fold change > 1). (D) The number of significantly enriched phenotypes in three main branches of MPO ($q < 0.05$, fold change > 1). It shows the significant associations between the main branches and expected cell types. The hypergeometric test results are indicated with asterisk (* $q < 0.05$, ** $q < 0.005$, *** $q < 0.0005$ and **** $q < 0.00005$).

Expected cell types are associated with the main MPO and HPO branches

The mammalian phenotype (MP:0000001) has 29 direct descendant phenotype terms in the MPO. The main MPO branches and the number of descendant phenotypes for each branch can be seen in Supplementary Figure 1A. We chose nervous system phenotype, cardiovascular system phenotype and immune system phenotype from main MPO branches, because these three branches have many descendant phenotypes, many of which are associated with cell types and these three branches would be expected to associate with expected cell types (Neurons, Cardiac muscle cells, and immune cells, respectively). Correspondingly, the Abnormality of the nervous system, Abnormality of the cardiovascular system and Abnormality of the immune system were chosen from main HPO branches. Figure 2D shows that the phenotypes from these main branches are more significantly enriched in expected cell types. The hypergeometric test was used to calculate the probability of whether the number of significantly enriched phenotypes for each cell type is significant in the main branch. The Benjamini-Hochberg method was used to correct multiple hypothesis testing. The significant results suggest that the main branch is significantly associated with these cell types. There are many significantly enriched phenotypes that are associated with neuron cells and glial cells in the nervous system phenotype branch, such as Neurons, Astrocytes, Oligodendrocytes in the central nervous system and Satellite cells in the peripheral nervous system. More phenotypes were associated with the cells in the central nervous system (CNS), which indicates that the core function of CNS in the nervous system (Brodal, 2004) and the dysfunction of CNS results in many neurological diseases. Interestingly, Microglia was associated with more phenotypes in the immune system phenotype branch rather than nervous system phenotype branch, which indicates the important role of Microglia in maintaining CNS homeostasis as the resident macrophage cells in CNS (Perry and Teeling, 2013; Filiano, Gadani and Kipnis, 2015). Previous studies (Perry, Nicoll and Holmes, 2010; Salter and Stevens, 2017; Wolf, Boddeke and Kettenmann, 2017) suggested that the Microglia plays a crucial role in neurodegenerative diseases and CNS pathology and disorders, which could be a potential marker for diagnosis, progression and prognosis of diseases and novel opportunities for treatment of diseases by targeting microglia. Cardiac muscle cells and Smooth muscle cells were associated with the most phenotypes in the cardiovascular system phenotype branch, that is consistent with the current understanding. There are many phenotypes that are associated with Pericytes, which could be that Pericytes play a crucial role in the physiological and pathological processes of cardiovascular disease (Su *et al.*, 2021). Surprisingly, many phenotypes were associated with Pancreatic stellate cells. It is not clear why they are associated and how Pancreatic stellate cells works to contribute the cardiovascular system phenotypes, as the current studies mainly focus on the roles of Pancreatic stellate cells on pancreatic disorders and pancreatic cancer (Apte, Pirola and Wilson, 2012; Ferdek and Jakubowska, 2017). Fewer phenotypes from HPO were associated with Endothelial cells within the cardiovascular system phenotype branch compared to the phenotypes from MPO. In the

immune system phenotype branch, there are many significantly enriched phenotypes that were associated with immune cells. There are many cell types of immune cells, which helps to study the interacting role of immune cells in the underlying mechanism of diseases. These association results demonstrate significant association between expected cell types and phenotypes.

To further demonstrate the association results between expected cell types and phenotypes, we calculated the proportion of significantly enriched phenotypes from main MPO and HPO branches that are associated with the expected cell type with different significance thresholds. The Neurons, Cardiac muscle cells, and Natural killer cells were chosen as expected cell types. Figure 3 shows that the proportion of significantly enriched phenotypes that were associated with the expected cell types from the expected main MPO and HPO branch increases, as the significance threshold becomes more stringent. This suggests that the significant association results between expected cell types and phenotypes are reliable for supporting scientific research and application.

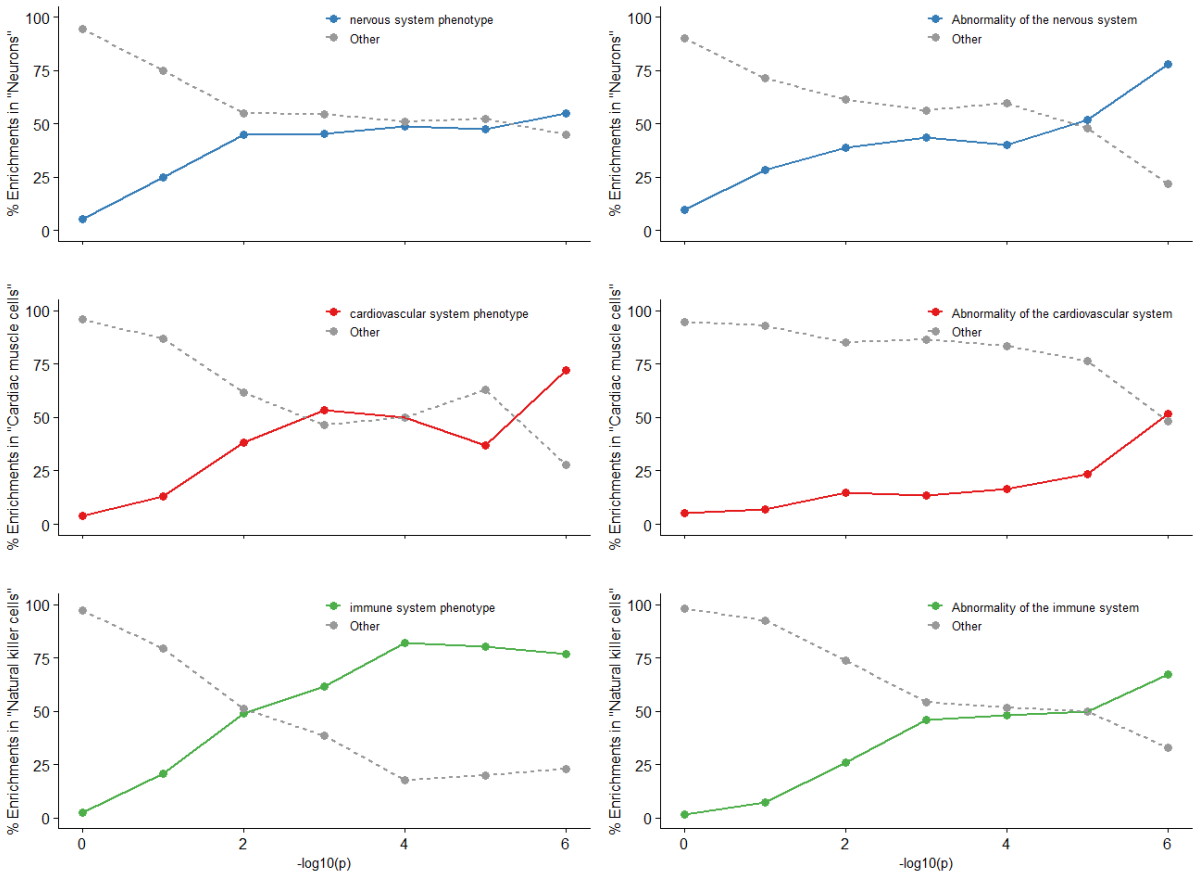


Figure 3. The proportion of significantly enriched phenotypes from main MPO and HPO branches with different significance thresholds. For expected cell type, as the significance threshold becomes more stringent, the proportion of significantly enriched phenotypes within the main MPO branch increases.

Low ontology level phenotypes tend to have higher specific expression in cells

We hypothesised that the low ontology level phenotypes in the hierarchy of the MPO and HPO have higher specific expression and should be associated with fewer cell types. The

ontology level is the number of generations of phenotype's descendants, which represents the position of phenotype term in the hierarchy ontology data. The fold change was used to quantify the specific expression. It describes how much specific expression of gene lists that are associated with phenotype changes in a cell type, compared to averaged specific expression of random gene lists that are generated from background gene lists. Figure 4E.F shows that the low ontology level phenotypes are associated with shorter gene lists in general, compared to high ontology phenotypes are associated with longer gene lists. It suggests that the low ontology level phenotypes are the more specific phenotypes and the fewer gene variants result in more specific phenotype change. On the other hand, it indicates that the low ontology phenotypes can better describe the symptoms and clinical features of disease. The fewer associated cell types (Figure 4AB) of low ontology phenotypes enable research to focus on fewer useful cell types. In addition, the lengths of gene lists associated with HPO phenotypes are overall shorter than those associated with MPO phenotypes, which could be that a large number of phenotyping studies over a long period of time have discovered numerous genotype-to-phenotype annotations. Figure 4 suggests that low ontology level phenotypes have higher fold change in specific expression and fewer associated cell types. In a word, the significant association at low ontology level phenotypes that are the more specific phenotypes tend to have higher specific expression of gene lists in the associated cells.

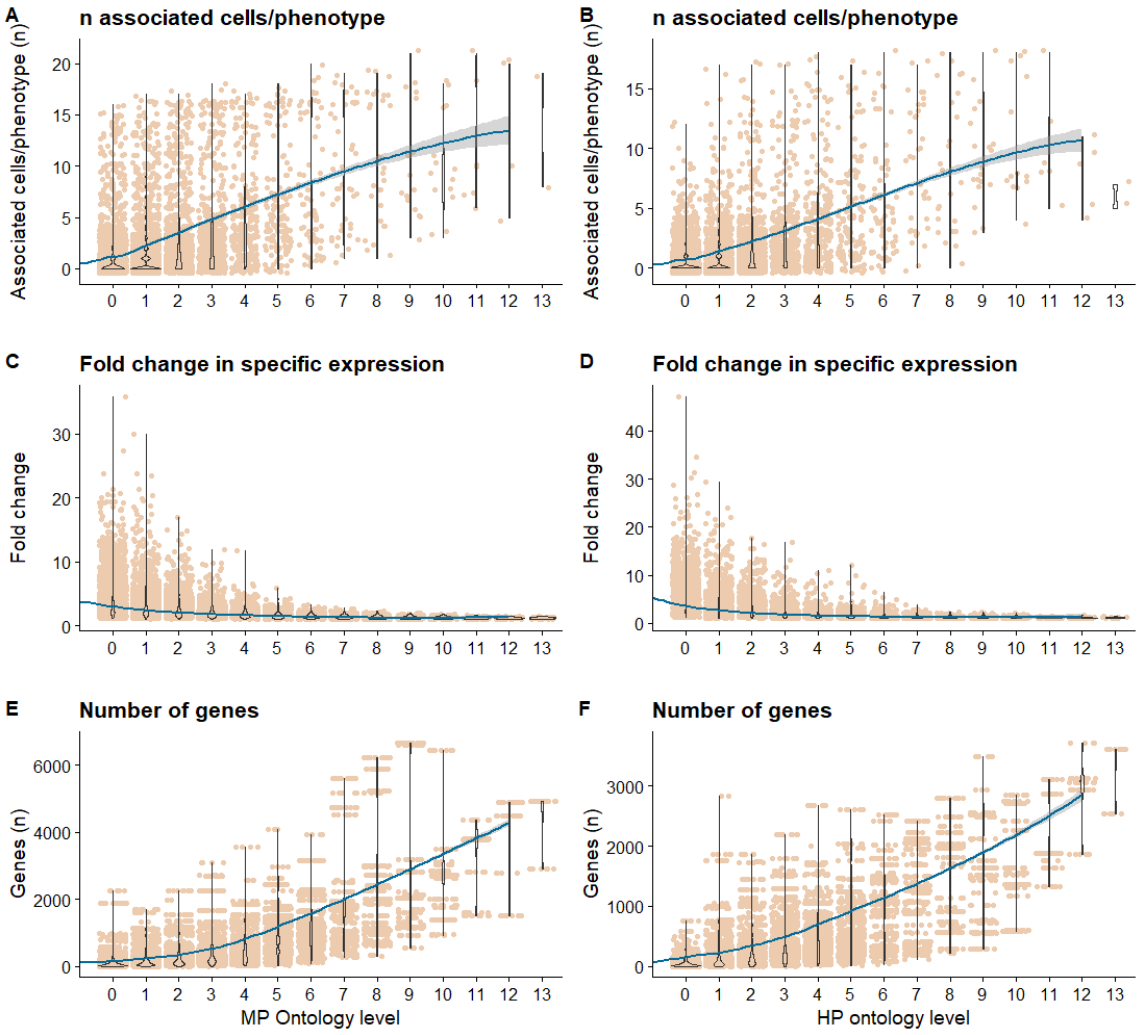


Figure 4. (A) (B) The relationship between ontology level and associated cells. Low ontology level phenotypes tend to have fewer associated cells. (C)(D) The relationship between ontology level and fold change in specific expressions. Low ontology level terms tend to have higher fold change in specific expression. (E)(F) The relationship between ontology level and number of genes. Low ontology level phenotypes are associated with shorter gene lists compared to the high ontology level phenotypes.

Low ontology level phenotypes are associated with expected and novel cell types

Due to low ontology level phenotypes that tend to have higher specific expression of genes, the phenotypes are more likely to be associated with expected and novel cell types. The abnormal consumption behavior (MP:0002069) from behavior/neurological phenotype (MP:0005386) main branch was chosen as an example. The definition of abnormal consumption behavior is altered ability or inability to eat or drink, or unusual choice or avoidance of foods or drink. It contains 51 descendant phenotypes from ontology level 0 to 3. The number of significant enrichments associated with cell types can be seen in Figure 5A. And Figure 6 shows the descendant phenotypes within abnormal consumption behavior that are associated with the cell types. The arrows between phenotype terms present the hierarchy in the ontology data. The majority of phenotypes are associated with Pancreatic islets and Neurons cell types, which could be that Pancreatic islets and Neurons have tight inter-regulation and influence (Thorens, 2014). The other cell types, such as Satellite cells, Pericytes, Smooth muscle cells, Macrophages, Erythrocytes, Pancreatic acinar cells and Epithelial cells are found to be associated with phenotypes.

The descendant phenotypes were analysed to explore the novel cell types. Figure 4B shows the fold change of the descendant phenotypes that are associated with the Satellite cells, Pericytes and Epithelial cells. The abnormal eating behavior, abnormal food intake, absent gastric milk in neonates and abnormal suckling behavior are significantly associated with Satellite cells. And the abnormal eating behavior has the most significant association with Satellite cells ($q = 0.008613333$). Satellite cells play an important role in growth, maintenance, and regeneration of skeletal muscle (Yin, Price and Rudnicki, 2013; Snijders *et al.*, 2015). How the Satellite cells work to regulate and influence these phenotypes are currently unknown and deserve further exploration. These phenotypes are relevant to metabolic derangements, which could impact the myogenesis and skeletal muscle regeneration (Joseph and Doles, 2021).

In addition, Pericytes are significantly associated with aphagia, absent gastric milk in neonates, abnormal food intake and abnormal eating behavior. And the aphagia has the most significant association with Pericytes (0.00019). Pericytes are part of the neurovascular unit (NVU), that consists of neurons and the cerebral vasculature to achieve the energy demands of the brain by cellular interaction. Pericytes enable cellular communication in neurovascular unit (NVU) to play a crucial role in maintaining brain functions, such as cerebral blood flow (CBF), vascular development and maintenance, and the formation of the blood-brain barrier (BBB) (Brown *et al.*, 2019). The review (Cheng *et al.*, 2018) summarised how the Pericytes malfunction contributes to numerous CNS disorders. Based on these phenotypes also associated with the neuron cells, such as Neurons and Astrocytes, Pericytes may play an important role in abnormal consumption behavior.

Figure 5. (A) The number of significant enrichments associated with cell types. (B) The associations between cell types and phenotypes of the descendants of abnormal

consumption behavior. The q value significance are indicated with asterisk (* q < 0.05, ** q < 0.005, *** q < 0.0005 and **** q < 0.00005).

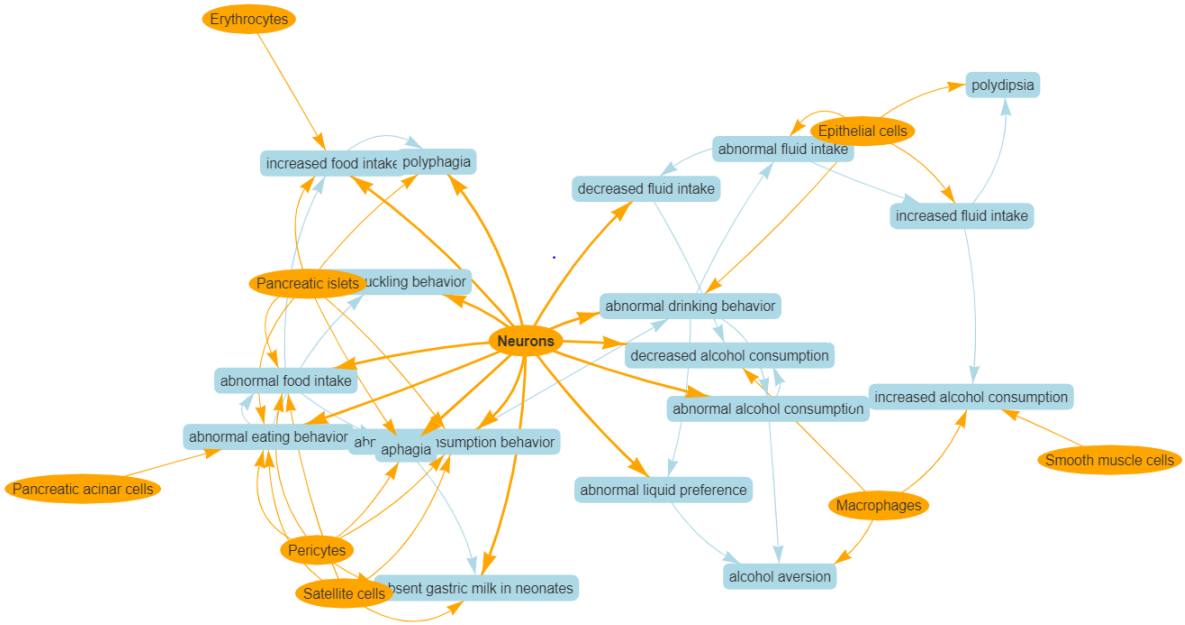
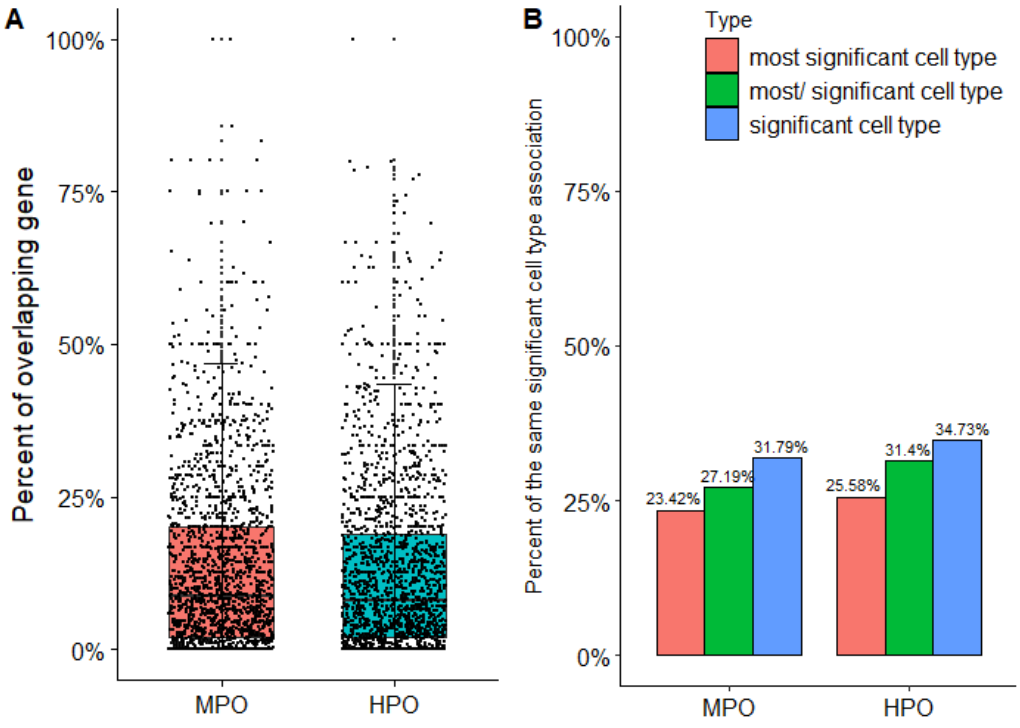


Figure 6. The association between cell types and descendant phenotypes within abnormal consumption behavior (MP:0002069). The arrows between phenotype terms indicate the hierarchy in the ontology data.

Comparison of cross-species phenotypes between the MPO and HPO

The cross-species phenotypes are phenotypes that have the phenotypic orthologue relation and SSSOM mapping relation between MPO and HPO terms. The phenotypic orthologue relation originated from unified phenotype ontology (uPheno) and the SSSOM mapping relation was derived from Monarch. There are 3208 pair phenotypes that have phenotypic orthologue and SSSOM mapping relation. We investigated the overlapping genes of the cross-species phenotypes to support the translational study. The percentage of overlapping genes of cross-species phenotypes can be seen in Figure 7A. For 50% cross-species phenotypes between upper quartile and lower quartile, the percentages of overlapping genes are concentrated in the 2%~20% range. The cross-species phenotypes (increased circulating ammonia level (MP:0005309) and Hyperammonemia (HP:0001987)) are shown as examples in Figure 7C. The phenotype-gene network shows the gene lists and common genes of the cross-species phenotypes. For better comparison, the gene symbols were denoted in corresponding human gene symbols. In addition, we calculated the percentage of the same significant cell type association in the cross-species phenotypes that were associated with cell types (Figure 6B). For MPO terms, the 23.42% phenotypes have the same most-significant cell type association as HPO terms. And 27.19% phenotypes have a significant cell type association (q<=0.05) as the cross-species phenotype's most significant cell type. For the same significant cell type association (q<=0.05), the percentage is 31.79%. For HPO terms, the 25.58% phenotypes have the same most-significant cell type association as HPO terms. And 31.4% phenotypes have a significant cell type association

(q<=0.05) as the cross-species phenotype's most significant cell type. For the same significant cell type association (q<=0.05), the percentage is 34.73%. These percentages of the same significant cell type association may be explained by the percentage of overlapping genes and the gene list of cross-species phenotypes. The gene list that is associated with mammalian phenotypes is generally longer than the gene list that is associated with human phenotypes (Figure 4EF). These data demonstrate that the mammalian phenotypes would be beneficial to translational research to some extent. The cell types associated with mammalian phenotypes could provide new insights for understanding of human phenotypes and diseases. The significant cell type association (q<=0.05) of cross-species phenotypes (abnormal liver physiology (MP:0000609), Elevated hepatic transaminase (HP:0002910) and Abnormal liver physiology (HP:0031865)) can be seen in Figure 8A. To make it easier to explore the cross-species phenotypes, a shiny web-based application (https://cell-phenotype.shinyapps.io/pheno-gene_network/) was developed, that enables users view the phenotype-gene network and cell-phenotype network of cross-species phenotypes.



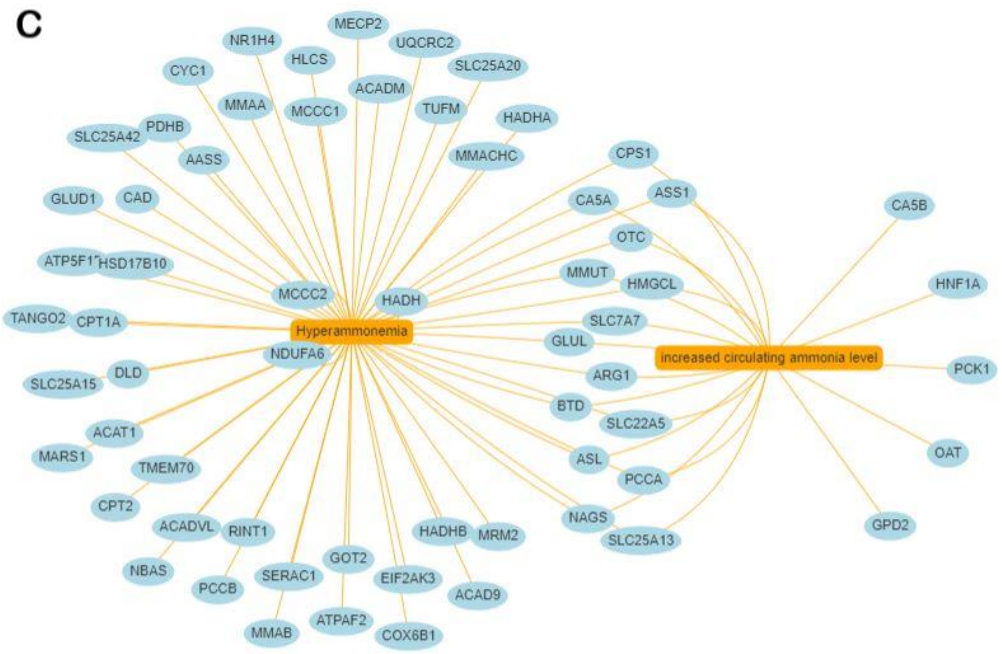


Figure 7. (A) The percentage of overlapping genes of cross-species phenotypes. (B) The percentage of the same most-significant cell type association in the cross-species phenotypes. (C) The gene lists and common genes of the cross-species phenotypes. The increased circulating ammonia level (MP:0005309) and Hyperammonemia (HP:0001987) are shown as examples.

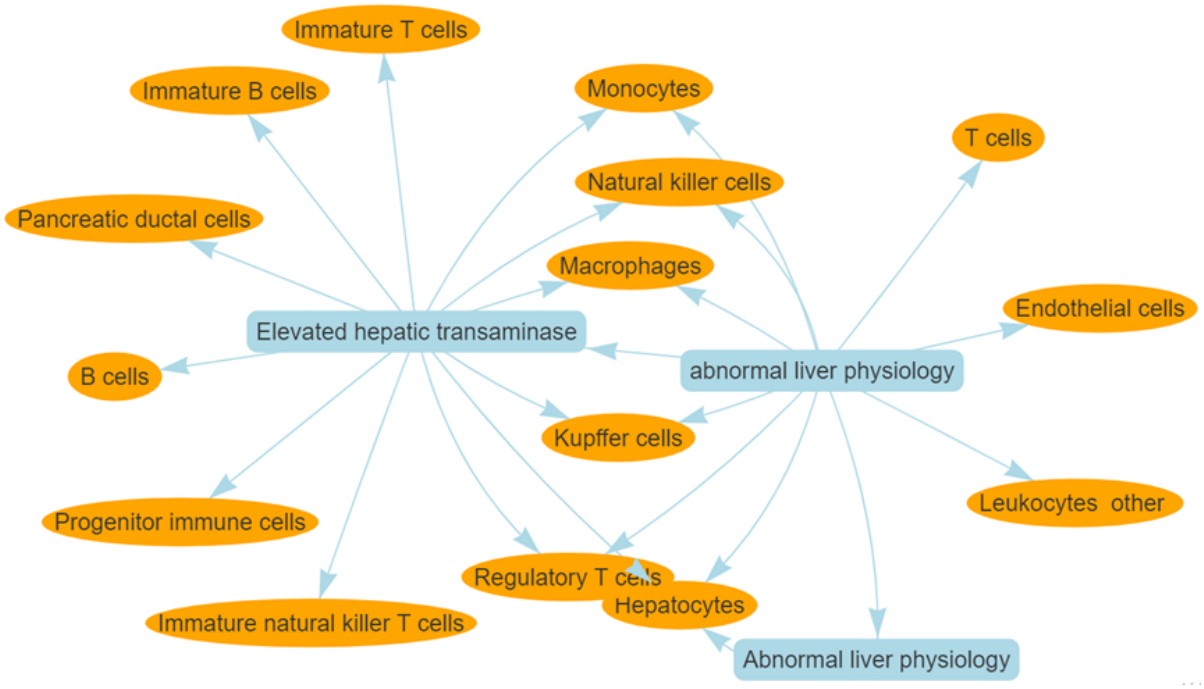


Figure 8. (A) The significant cell type association (q<=0.05) of cross-species phenotypes. The abnormal liver physiology (MP:0000609), Elevated hepatic transaminase (HP:0002910) and Abnormal liver physiology (HP:0031865) are shown as examples.

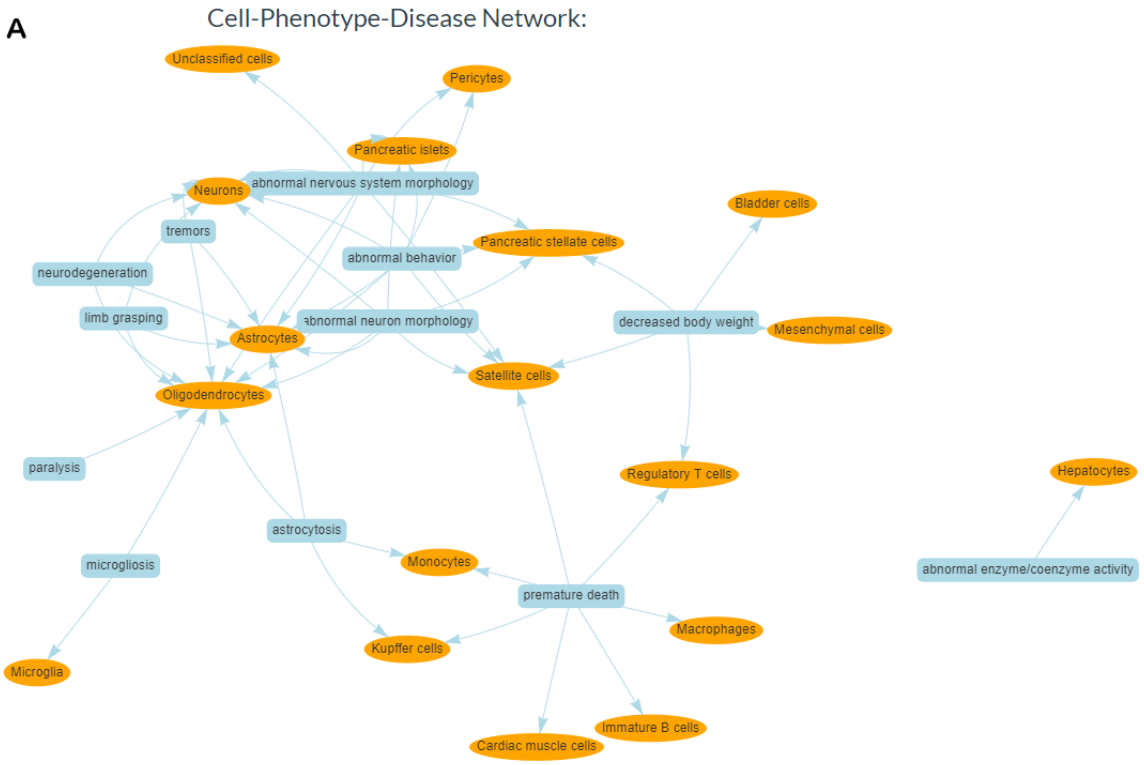
Connection from cell types to human diseases

Due to most phenotypes being the clinical features and symptoms of diseases, the connection between mammalian phenotypes and human diseases from Human - Mouse: Disease Connection (HMDC) was used to further explore the roles of mammalian phenotypes in human disease research. The aim of Human - Mouse: Disease Connection (HMDC) is to discover the candidate genes and investigate the phenotypic similarity between mouse models and human patients by identifying the published and potential mouse models of human disease. The 2915 mammalian phenotypes that significantly associated with cell types (q <= 0.05) were mapped to the 1222 human disease ontology (DO) and Online Mendelian Inheritance in Man (OMIM). We can explore the connections between mammalian phenotypes that are significantly associated with cell types and human diseases. Here, we chose progressive myoclonus epilepsy (DOID:891) and Wolfram syndrome as examples.

Progressive myoclonus epilepsy

The progressive myoclonus epilepsy (DOID:891) is a group of rare diseases with myoclonic seizures and progressive neurodegeneration (Girard *et al.*, 2013). The common feature of these diseases contains focal occipital seizures presenting and fragmentary, symmetric, or generalised myoclonus (Jansen and Andermann, no date). The progressive myoclonus epilepsy (PME) patients have some symptoms, such as cognitive decline, dysarthria and ataxia, emotional disturbance and confusion (Jansen and Andermann, no date). The phenotypes that were connected with PME are shown in Figure 9A and most of these phenotypes have the same or similar symptoms with PME. This indicates that the mammalian phenotypes can be the translationally support for the human diseases. The Figure 9A also shows that the cell types are associated with the phenotypes, which helps to facilitate the understanding of diseases. Oligodendrocytes, Astrocytes, Neurons and Satellite cells have the most associated phenotypes (Figure), which is consistent with our understanding of the role of these cells in neurological diseases. A study (Rubio-Villena *et al.*, 2018) suggested that the Astrocytes may play a crucial role in the pathophysiology of Lafora disease (a type of Progressive myoclonus epilepsy), as the accumulation of glycogen inclusions in these Astrocytes may affect normal function and lead to possible neuronal dysfunction. Other studies (Takao *et al.*, 2000) have reported that S52R Neuroserpin mutation causes neuroserpin increase in the perikarya and cell processes of the neurons and is associated with hereditary progressive myoclonus epilepsy. The study also observed a cluster of satellite cells in the place of a degenerated neuron. These research suggest that Astrocytes, Neurons and Satellite cells may play an important role in PME. Although Oligodendrocytes have the most associated phenotypes that are associated with PME, there are no current studies describing the association between the Oligodendrocytes and PME. This also suggests that Oligodendrocytes may be a good research direction and target for PME. Pancreatic stellate cells and Pancreatic islets were found to associate with phenotypes that were connected to PME (Figure). Although there are no current studies that describe the association between the Pancreatic stellate cells, Pancreatic islets and PME, the mammalian phenotypes may provide some evidence for the understanding of PME. Pancreatic stellate cells and Pancreatic islets were associated with abnormal behavior (MP:0004924), abnormal nervous system morphology (MP:0003632), abnormal neuron morphology (MP:0002882), decreased body weight (MP:0001262) and these phenotypes

are the same or similar symptoms with PME. The review (Omary *et al.*, 2007) illustrated that the Pancreatic stellate cells express the intermediate filament proteins desmin and glial fibrillary acidic protein (GFAP), neural markers for astrocytes. A study (Demir *et al.*, 2017) reported that Pancreatic ductal adenocarcinoma cancer (PDAC) cells have an ability to invade nerves via pronounced crosstalk between nerves and cancer cells, which facilitate migration of glia cells to cancerous cells to relieve pain by downregulation of pain-associated targets in Schwann cells and suppression of central glia. The other review (Thorens, 2014) illustrated that the nervous system and Pancreatic islet have tight inter-regulation and influence, which impact endocrine pancreas development and function. From the above study, Pancreatic stellate cells and Pancreatic islets may provide new insights for the understanding of the progressive myoclonus epilepsy. PME contains many different types and different types have different pathogeny. With the rapid development of high-throughput sequencing technology, molecular genetic testing (Katsanis and Katsanis, 2013) is a powerful and effective tool to diagnose PME. Furthermore, the identification of cell types that are associated with specific types of disease may be beneficial to accurately diagnose the types on the early stage and personalised precision treatment.



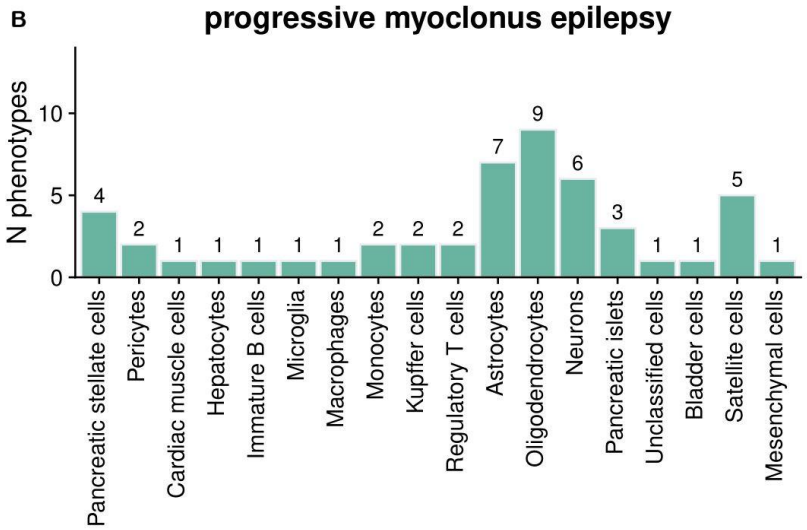


Figure 9. (A) The cell type-phenotype network of progressive myoclonus epilepsy (DOID:891). (B) The number of phenotypes connected to progressive myoclonus epilepsy were associated with cell types.

Wolfram syndrome

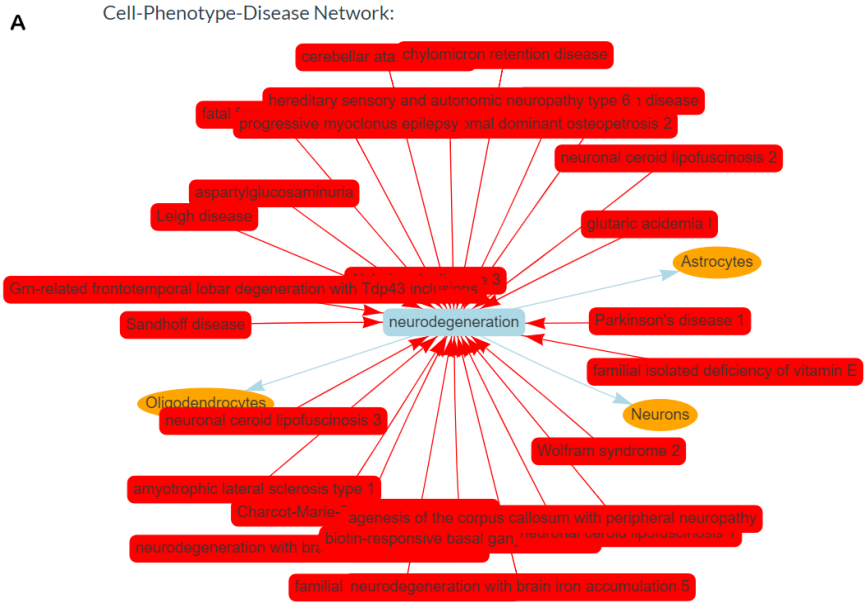
The research and exploration pattern of progressive myoclonus epilepsy is from disease to phenotype to cell type. On the other hand, we can start with phenotype and then see which diseases are connected with it and which cell types are associated with it. For example, the neurodegeneration (MP:0002229) that was significantly associated with Neurons, Astrocytes and Oligodendrocytes is mapped to many neurological diseases (Figure 10A). This is consistent with our current understanding, that neurodegeneration is the common symptom of neurological diseases. The significant celltype-phenotype association results indicate that the Neurons, Astrocytes and Oligodendrocytes may play an important role in these neurological diseases and these cell types could be the focus of the relevant research. In addition to neurological diseases, there are other types of diseases that are connected with neurodegeneration, and Wolfram syndrome 2 (DOID:0110630) is one of them. Wolfram syndrome is typically associated with diabetes insipidus, childhood-onset insulin-dependent diabetes mellitus, optic atrophy, and deafness. Wolfram syndrome 2 that is associated with neurodegeneration arouses interests to further explore.

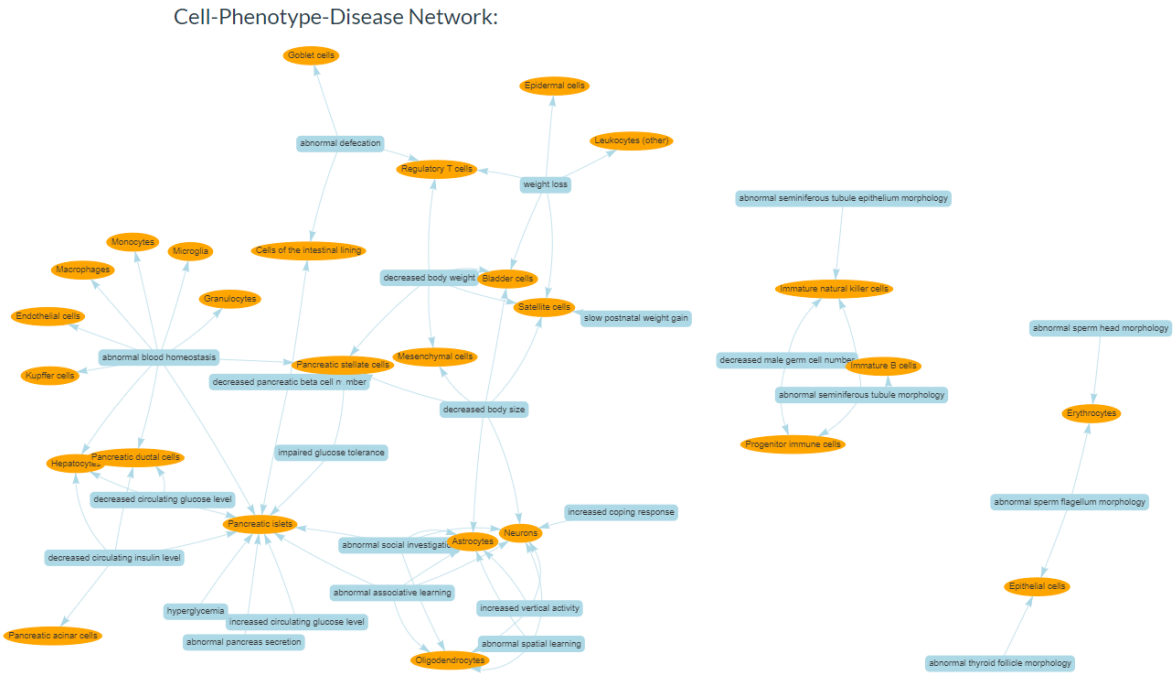
Wolfram syndrome (WS) is a rare genetic and endoplasmic reticulum (ER) disease. There are two types of WS, Wolfram syndrome 1 (WS1, DOID:0110629) and Wolfram syndrome2 (WS2, DOID:0110630). Wolfram syndrome2 is rarer than Wolfram syndrome 1.The mutation of WFS1 gene causes the majority of symptoms of WS1 and the mutation of CISD2 gene causes the symptoms of WS2. The primary symptoms of WS are the diabetes insipidus, childhood-onset insulin-dependent diabetes mellitus, optic atrophy, deafness, and neurological signs (Pallotta *et al.*, 2019). A review of the Wolfram syndrome described that 62% of WS patients develop psychiatric illness and neurological complications, and that the most common complication is cerebellar ataxia (Pallotta *et al.*, 2019). The majority of the deaths were caused by central apnea due to the brain stem atrophy (Urano, 2016). It is credible that the neurodegeneration is associated with Wolfram syndrome 2.

Figure 10B and Figure 11A show that the phenotypes that are separately connected to Wolfram syndrome 1 and Wolfram syndrome 2 are associated with the cell types. And most of the phenotypes are the symptoms of Wolfram syndrome. The phenotypes that are related to diabetes are abnormal glucose homeostasis, abnormal blood homeostasis, abnormal pancreas secretion, abnormal mitochondrial morphology, decreased body size, decreased body weight, weight loss, decreased hair follicle number, decreased insulin secretion, decreased circulating glucose level, decreased circulating insulin level, decreased pancreatic beta cell number, increased circulating glucose level, impaired glucose tolerance and hyperglycemia (Barrett, Bundey and Macleod, 1995; Urano, 2016; Pallotta *et al.*, 2019). The phenotypes associated with the nerve are abnormal optic nerve morphology, abnormal respiratory electron transport chain, abnormal respiratory function, axon degeneration, decreased fluid intake, neurodegeneration, premature aging, premature death, retina neovascularization, abnormal associative learning, abnormal social investigation, abnormal social investigation, increased coping response and increased vertical activity (Barrett, Bundey and Macleod, 1995; Urano, 2016). This suggested that these mammalian phenotypes may provide us new perspectives to understand and study diseases. Pancreatic stellate cells and Pancreatic islets have the most associated significant phenotypes (Figure 10C, Figure 11B), which is consistent with our understanding on the important roles of these cell types in the Wolfram syndrome. Many studies proved that Pancreatic stellate cells are highly associated with the process of diabetes (Del Guerra *et al.*, 2005; Kaddis *et al.*, 2009; Yang *et al.*, 2020; Perera *et al.*, 2021; Zhang *et al.*, 2022). It is explainable and understandable that many phenotypes are associated with Neurons, Oligodendrocytes and Astrocytes, as Wolfram syndrome are typically associated with abnormal nervous systems.The main role of Satellite cells is the growth, maintenance and regeneration of skeletal muscle (Yin, Price and Rudnicki, 2013; Snijders *et al.*, 2015) and Satellite cells were associated with phenotypes that were related to muscle, weight and respiratory function. The common urological manifestations of Wolfram syndrome are structural and functional urinary tract abnormalities (Urano, 2016; Pallotta *et al.*, 2019) and Bladder cells may be deeply associated with these manifestations. Diabetes patients are at higher risk of urinary tract infection (UTI) due to various immune system abnormalities, poor metabolic control and autonomic neuropathy (Nitzan *et al.*, 2015). It is possible that Pancreatic stellate cells, Pancreatic islets and Bladder cells involved the progress of Wolfram syndrome via interactions.

In addition, it is surprising that Cardiac muscle cells and Smooth muscle cells were significantly associated with 7 phenotypes and only Wolfram syndrome 2 were connected with these phenotypes. These phenotypes are abnormal mitochondrial morphology, abnormal muscle fiber morphology, abnormal myocardium layer morphology, abnormal respiratory electron transport chain, muscle degeneration, decreased subcutaneous adipose tissue amount and premature death. One of the reasons could be that the Wfs1 gene is highly expressed in the heart tissue (Inoue *et al.*, 1998; Fonseca *et al.*, 2010). WFS1 gene encodes a protein wolframin that is localized in the membrane of endoplasmic reticulum (ER) (Inoue *et al.*, 1998), and one of the functions of endoplasmic reticulum (ER) is responsible for calcium homeostasis in the cytoplasm. A study (Cagalinec *et al.*, 2019) characterized that invalidation of WFS1 causes elevated contraction and prolonged calcium transients and affects release-dependent inactivation of calcium channels in isolated ventricular myocytes. This may explain why Cardiac muscle cells were associated with these phenotypes and suggest the important roles of Cardiac muscle cells in the Wolfram

syndrome. As the WFS1 gene only accounts for Wolfram syndrome 1, these results may suggest that the analysis of subtypes of diseases could facilitate the comprehensive understanding of diseases. And different cell types and numbers of associated phenotypes were shown in different subtypes of diseases.The identification of cell types would help to understand how the different cells interact with each other to contribute the process of disease and different cell types of subtype of diseases may beneficial to diagnose the specific subtype and corresponding precision treatments on specific cell types.





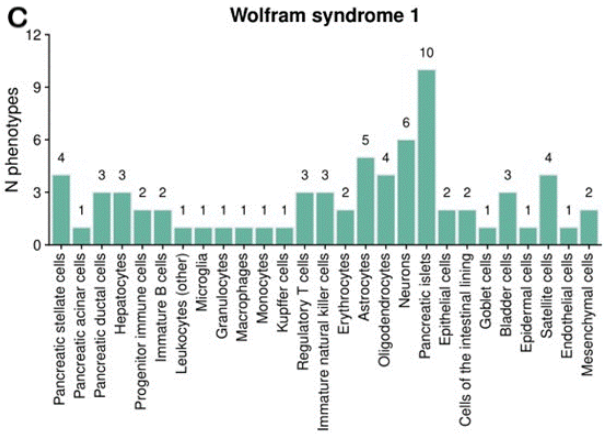
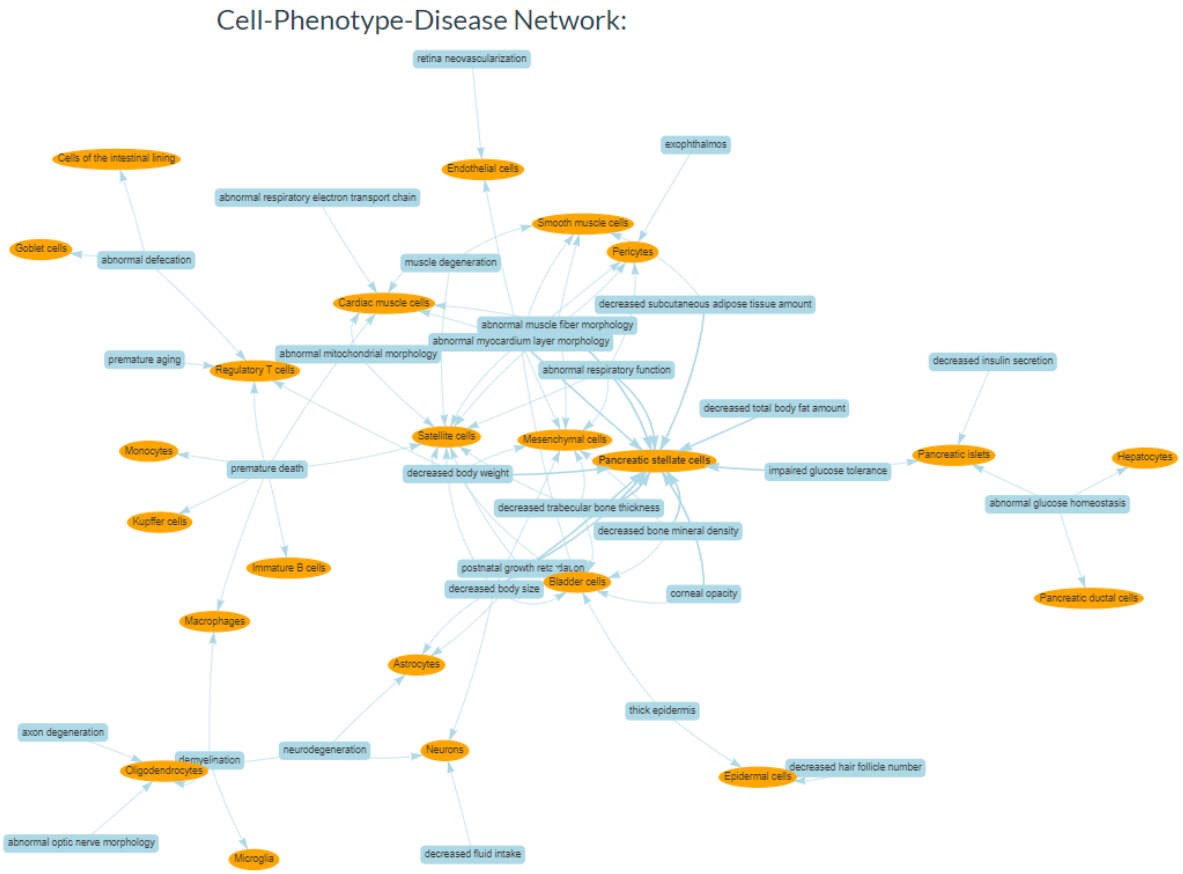


Figure 10. (A) the cell-phenotype-disease network shows the diseases connected to neurodegeneration (MP:0002229) that were significantly associated with the cell types. (B) The cell type-phenotype network of Wolfram syndrome 1 (DOID:0110629). (C) The number of phenotypes connected to Wolfram syndrome 1 were associated with cell types.



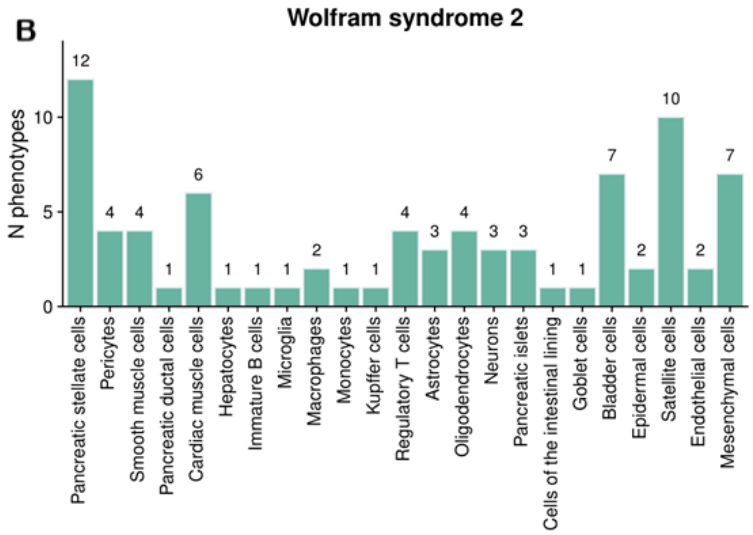


Figure 11. (A) The cell type-phenotype network of Wolfram syndrome 2 (DOID:0110630). (C) The number of phenotypes connected to Wolfram syndrome 2 were associated with cell types.

Comparison of phenotypes associated with disease between MPO and HPO

To demonstrate that mammalian phenotypes can translationally support the human disease study, we summarised the percentage of the same significant cell type that were connected to human disease between mammalian phenotypes and human phenotypes. The connection between mammalian phenotypes and human diseases comes from HMDC and the connection between human phenotypes and human diseases comes from MONDO database. The 1017 human diseases that have the same disease name were used to summarise. Figure 12A shows that most diseases are connected to the same cell type, that is because most diseases were connected to many phenotypes. For 50% mammalian phenotypes between upper quartile and lower quartile, the percentages of the same significant cell type are concentrated between 40.91% and 84.38%. For 50% human phenotypes between upper quartile and lower quartile, the percentages of the same significant cell type are concentrated between 33.33% and 81.48%. This suggests that cell types associated with mammalian phenotypes and human phenotypes within human diseases are highly overlapping. Furthermore, we summarised the percentage of cross-species phenotypes that were connected to human diseases. Nearly half of diseases were connected to cross-species phenotypes (Figure B). Most of the number of cross-species phenotypes are 1, 2, 3 (Figure B) and the percentage of cross-species phenotypes in mammalian phenotypes and human phenotypes are concentrated below 20% (Figure C). This summarised data demonstrated that using mammalian phenotypes would be a reasonable, understandable, and useful tool for human disease research.

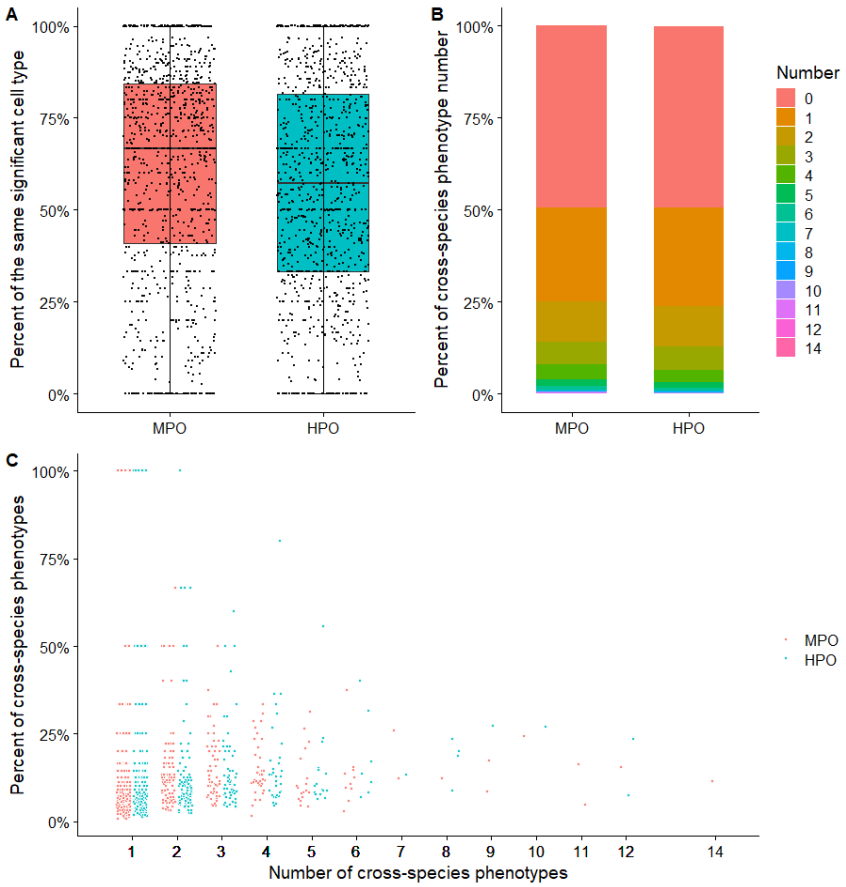


Figure 12. (A) The percentage of the same significant cell type connected to disease. (B) The percentage of cross-species phenotype number in the phenotypes connected to disease. (C) The percentage of cross-species phenotypes for different numbers of cross-species phenotypes.

Shiny web-based application

A shiny web-based application was developed for searching and retrieving the association results. The researchers and clinicians can conveniently and interactively retrieve the association results relevant to their field of study and interest without the requirement of significant computational resources and bioinformatic skills. The web-based application was created by Shiny that can build interactive web apps from R and deployed on the shiny server (http://www.cell-phenotype.com/). It contains 5 main functions to explore the cell-phenotype association results, including searching for cell types, phenotypes, diseases and exploring cross-species phenotypes. The home page (Figure 13A) was created by HTML and CSS to link the web-based applications of different function and data sources, and code repositories. The code for building the home page is available at (https://github.com/LushengLi9909/Home_page). The minimalist user interface (UIs) facilitates user-friendly operation and enhances the user experience. The 5 main functions were developed through separate shiny applications to reduce the time cost and improve the task efficiency.

Function 1: Cell type application

The cell type application (Figure 13B) can search for the phenotypes that are associated with a particular cell type by selecting the cell type, significance threshold and fold change of expression. The phenotypes that have the hierarchical relationship on the ontology data were linked to the network by ggnetwork (Version 0.5.10) (Briatte, 2020), network (Version 1.17.2) (Butts, 2008) and sna (Version 2.6) R packages. The interactive plot was achieved by ggplotly R package (Sievert, 2020) to convert the ggplot2 to plotly. The hover box can show the detailed information of phenotypes. Simultaneously, the reactive data frame will display the detailed information about phenotypes, descriptions, fold change and statistical values. As the network images are too large and complex to be displayed on the web interface, the higher resolution and more ordered network images can be downloaded by adjusting the width and height parameters. The cell type application was deployed on shiny server (https://cell-phenotype.shinyapps.io/cell-search/) and the code of cell type application available at (https://github.com/LushengLi9909/Cell-search).

Function 2: Phenotype application

The phenotype application (Figure 14A) can search for the cell types that are associated with phenotypes by entering the keywords of phenotypes, significance threshold, fold change of expression and standard deviations from mean. The input can be the keywords of phenotype or a phenotype term. If users enter a phenotype term, the Cell-Phenotype network will show the descendant phenotypes (Figure 6), which helps to look at the cell-phenotype association within a phenotype in more detail. If users enter the keywords of phenotype, the regular expression would be created to search the relevant phenotypes. For example, the Cell-Phenotype network (Figure 14B) shows that the cell types are associated with these phenotypes related to anxiety and gut. The histogram that shows the number of phenotypes associated with the cell types was created by the ggplot2 (Wickham, 2016) R package and the interactive plot was achieved by ggplotly (Sievert, 2020) R package to convert the ggplot2 to plotly. If every cell type only has one significant phenotype association, the histogram will show the Fold Change of specific expression of phenotype. The Cell-Phenotype network was created by visNetwork (version 2.1.0)(Almende, Thieurmel and Robert, no date). The visNetwork is an efficient and easy-to-use R package for network visualisation. The phenotype application was deployed on shiny server (https://cell-phenotype.shinyapps.io/phenotype_search/) and the code of cell type application available at (https://github.com/LushengLi9909/Phenotype-search).

Function 3: Phenotype-Disease application

The Phenotype-Disease application (Figure 14 A)can explore the cell-phenotype-disease association. The connection between phenotypes and diseases comes from the Human - Mouse: Disease Connection (HMDC). And the cell types that were significantly associated with the phenotypes were mapped to the human diseases. The users can view which cell types were associated with the phenotype and which DO diseases were connected to the phenotype by entering a mammalian phenotype (Figure 14B). The users also can view which phenotypes were connected to DO diseases or OMIM diseases and which cell types were significantly associated with these phenotypes by entering Disease Ontology term or OMIM number. The different choices enable users to study the cell-phenotype-disease association from different patterns. The histogram would be created to show the number of phenotypes that were associated with the cell types (Figure 14A). The search box widget

can provide the contextual suggestions based on the Elasticsearch fuzzy search as users type, which provides a good user experience and ensures the accuracy of input terms. The interactive table would display the details of cell types, phenotypes and human diseases. The phenotype application was deployed on shiny server (https://cell-phenotype.shinyapps.io/cell-disease/) and the code of cell type application available at (https://github.com/LushengLi9909/Cell-Disease).

Function 4: Cross-species phenotype application

The cross-species phenotype interface enables users to view the phenotype-gene network that shows gene lists of cross-species phenotypes and common genes (Figure 7C) by entering phenotype. When users type phenotype, the search box widget would provide the phenotype choice suggestions based on the Elasticsearch fuzzy search, that allows users to enter phenotype terms more easily and ensures accuracy. And only phenotypes that have the phenotypic orthologue relation are on the choice list. It also can identify the phenotype whether it is the cross-species phenotype between mammalian and human phenotypes. If the cross-species phenotypes are significantly associated with cell types, the Cell-Phenotype network(Figure 8) and data frame would be displayed. The users can view the significant association between cross-species phenotypes and cell types, and relevant detailed information, such as significance value, fold change and standard deviation from mean. The cross-species phenotype application was deployed on shiny server (https://cell-phenotype.shinyapps.io/pheno-gene_network/) and the code of cell type application available at (https://github.com/LushengLi9909/Pheno-gene-network).

Function 5: Cross-species phenotype-Disease application

The Cross-species phenotype-Disease application enables users to retrieve and explore which mammalian phenotypes and human phenotypes were connected to the human diseases and which cell types were significantly associated with these phenotypes (Figure 15). Only diseases connected with both mammalian phenotypes and human phenotypes are on the choice list of diseases search box. Simultaneously, the histogram would be created to show the number of mammalian and human phenotypes that were significantly associated with the cell types ($q <= 0.05$). The Cell-Phenotype-Disease Network and histogram allow users to compare differences between MPO and HPO more easily and intuitively. The Cross-species phenotype-Disease application was deployed on shiny server (https://cell-phenotype.shinyapps.io/cross-species-disease/) and the code of cell type application available at (https://github.com/LushengLi9909/Cross-species-Disease).

Figure 13. (A) Screenshot of the home page. (B) Cell type application for searching the phenotypes that are associated with a particular cell type.

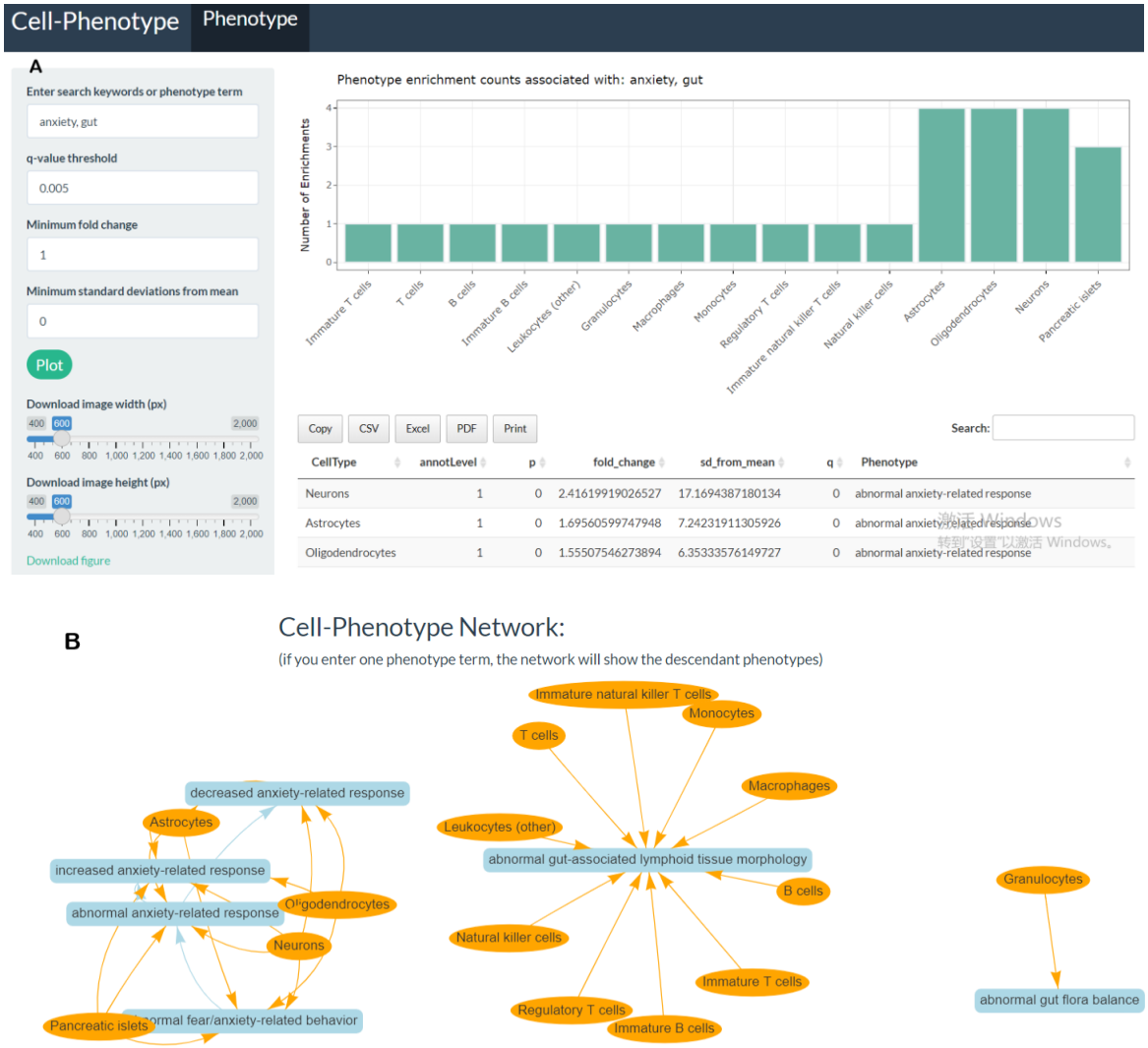


Figure 14. (A) Phenotype application for searching the cell types that are associated with phenotypes. (B) the Cell-Phenotype network shows that the cell types are associated with these phenotypes related to anxiety and gut.

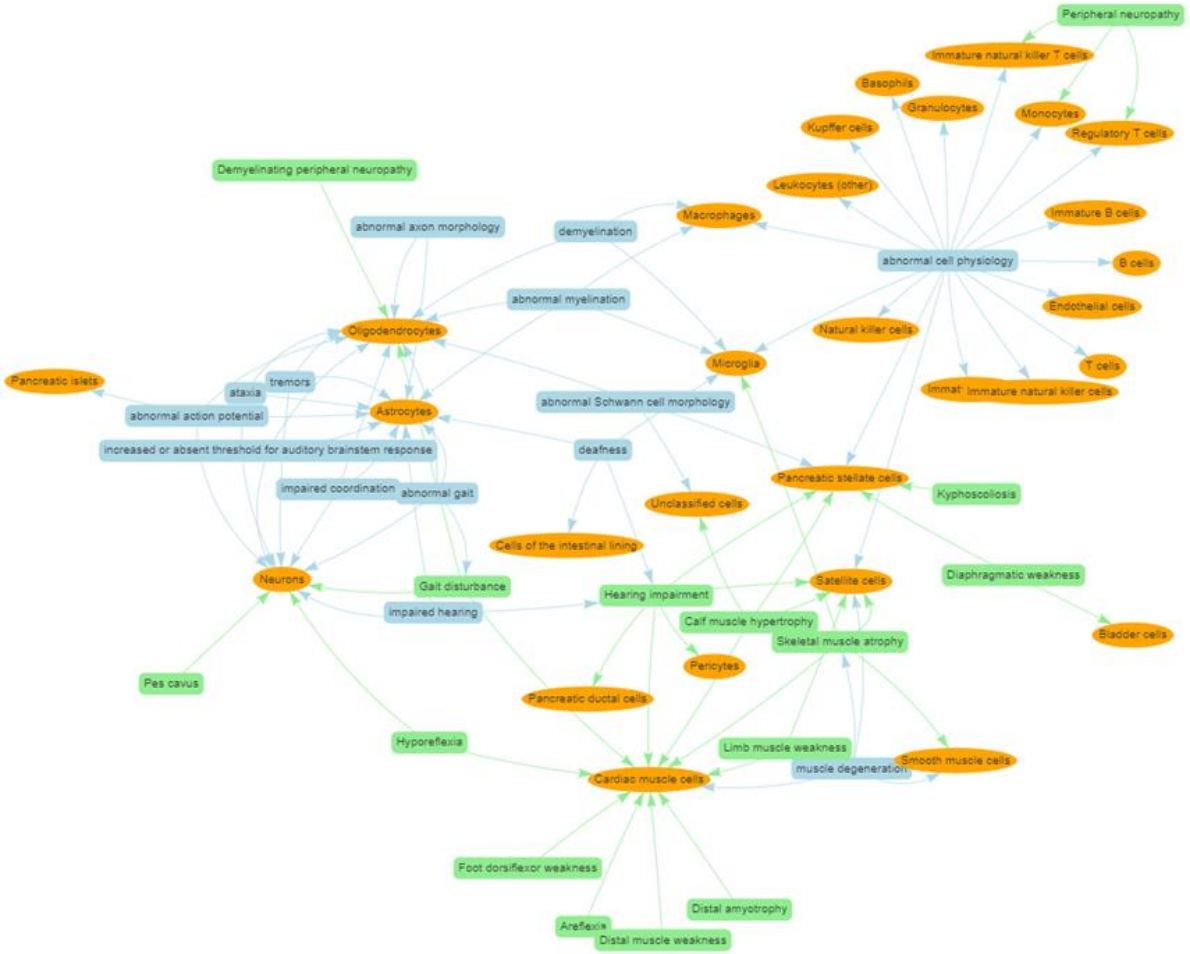


Figure 15. The cell-phenotype-disease network. The light blue boxes are denoted as MP, the light green boxes are denoted as HP.

Supplementary Files

Supplementary Figures

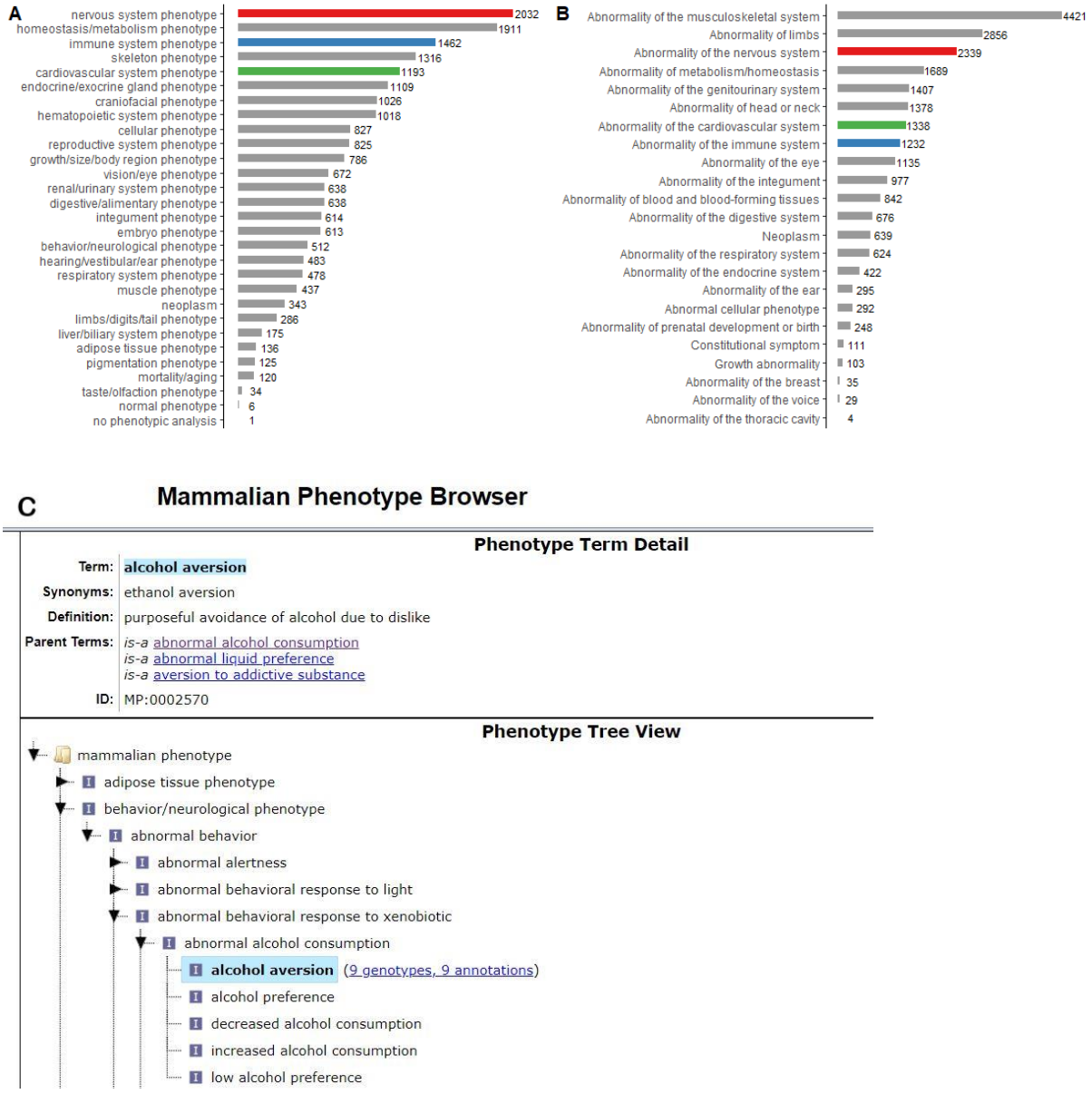


Figure 1. (A) the main branches of MPO. It shows the child terms of the mammalian phenotype (MP:0000001) and the number of descendants of main branches. (B) the main branches of HPO. It shows the child terms of the Phenotypic abnormality (HP:0000118) and the number of descendants of main branches.

References

‘Alliance of Genome Resources Portal: unified model organism research platform’ (2020) *Nucleic acids research*, 48(D1), pp. D650–D658.

Almende, Thieurmel and Robert (no date) ‘visNetwork: Network Visualization using’vis. js' Library’, *R package version* [Preprint].

Amberger, J.S. *et al.* (2015) ‘OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders’, *Nucleic Acids Research*, pp. D789–D798. doi:10.1093/nar/gku1205.

Andersson, A. *et al.* (2020) ‘Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography’, *Communications biology*, 3(1), p. 565.

Apte, M.V., Pirola, R.C. and Wilson, J.S. (2012) ‘Pancreatic stellate cells: a starring role in normal and diseased pancreas’, *Frontiers in physiology*, 3, p. 344.

Barrett, T.G., Bundey, S.E. and Macleod, A.F. (1995) ‘Neurodegeneration and diabetes: UK nationwide study of Wolfram (DIDMOAD) syndrome’, *The Lancet*, 346(8988), pp. 1458–1463.

Benjamini, Y. and 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)*, pp. 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Birgfeld, C.B. *et al.* (2011) ‘A phenotypic assessment tool for craniofacial microsomia’, *Plastic and reconstructive surgery*, 127(1), pp. 313–320.

Briatte, F. (2020) ‘ggnetwork: geometries to plot networks with “ggplot2”. R package version 0.5. 8’.

Brodal, P. (2004) *The Central Nervous System: Structure and Function*. Oxford University Press, USA.

Brown, L.S. *et al.* (2019) ‘Pericytes and Neurovascular Function in the Healthy and Diseased Brain’, *Frontiers in cellular neuroscience*, 13, p. 282.

Bryois, J. *et al.* (2020) ‘Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson’s disease’, *Nature genetics*, 52(5), pp. 482–493.

Bult, C.J. *et al.* (2019) ‘Mouse Genome Database (MGD) 2019’, *Nucleic acids research*, 47(D1), pp. D801–D806.

Butts, C.T. (2008) ‘**network**: A Package for Managing Relational Data in*R*’, *Journal of Statistical Software*. doi:10.18637/jss.v024.i02.

Cagalinec, M. *et al.* (2019) ‘Calcium Signaling and Contractility in Cardiac Myocyte of Wolframin Deficient Rats’, *Frontiers in physiology*, 10, p. 172.

Cao, J. *et al.* (2020) ‘A human cell atlas of fetal gene expression’, *Science*, 370(6518). doi:10.1126/science.aba7721.

Cao, Z.-J. *et al.* (2020) ‘Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST’, *Nature communications*, 11(1), p. 3458.

Cheng, J. *et al.* (2018) ‘Targeting pericytes for therapeutic approaches to neurological disorders’, *Acta neuropathologica*, 136(4), pp. 507–523.

Del Guerra, S. *et al.* (2005) ‘Functional and molecular defects of pancreatic islets in human type 2 diabetes’, *Diabetes*, 54(3), pp. 727–735.

Demir, I.E. *et al.* (2017) ‘Early pancreatic cancer lesions suppress pain through CXCL12-mediated chemoattraction of Schwann cells’, *Proceedings of the National Academy of Sciences of the United States of America*, 114(1), pp. E85–E94.

Domcke, S. *et al.* (2020) ‘A human cell atlas of fetal chromatin accessibility’, *Science*, 370(6518). doi:10.1126/science.aba7612.

Dong, J. *et al.* (2022) ‘Integrating single-cell datasets with ambiguous batch information by incorporating molecular network features’, *Briefings in bioinformatics*, 23(1). doi:10.1093/bib/bbab366.

Eraslan, G. *et al.* (2022) ‘Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function’, *Science*, 376(6594), p. eabl4290.

Ferdek, P.E. and Jakubowska, M.A. (2017) ‘Biology of pancreatic stellate cells—more than just pancreatic cancer’, *Pflügers Archiv - European Journal of Physiology*, 469(9), pp. 1039–1050.

Filiano, A.J., Gadani, S.P. and Kipnis, J. (2015) ‘Interactions of innate and adaptive immunity in brain development and function’, *Brain research*, 1617, pp. 18–27.

Fonseca, S.G. *et al.* (2010) ‘Wolfram syndrome 1 gene negatively regulates ER stress signaling in rodent and human cells’, *The Journal of clinical investigation*, 120(3), pp. 744–755.

Gentleman, R.C. *et al.* (2004) ‘Bioconductor: open software development for computational biology and bioinformatics’, *Genome biology*, 5(10), p. R80.

Ghoussaini, M. *et al.* (2021) ‘Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics’, *Nucleic acids research*, 49(D1), pp. D1311–D1320.

Girard, J.-M. *et al.* (2013) ‘Progressive myoclonus epilepsy’, *Handbook of clinical neurology*, 113, pp. 1731–1736.

Inoue, H. *et al.* (1998) ‘A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome)’, *Nature Genetics*, pp. 143–148. doi:10.1038/2441.

Jagadeesh, K.A. *et al.* (no date) ‘Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics’. doi:10.1101/2021.03.19.436212.

Jansen and Andermann (no date) ‘Progressive myoclonus epilepsy, Lafora type’, *GeneReviews®[Internet]* [Preprint]. Available at: https://www.ncbi.nlm.nih.gov/sites/books/NBK1389/.

Joseph, J. and Doles, J.D. (2021) ‘Disease-associated metabolic alterations that impact satellite cells and muscle regeneration: perspectives and therapeutic outlook’, *Nutrition & metabolism*, 18(1), p. 33.

Kaddis, J.S. *et al.* (2009) ‘Human pancreatic islets and diabetes research’, *JAMA: the journal of the American Medical Association*, 301(15), pp. 1580–1587.

Kashima, Y. *et al.* (2020) ‘Single-cell sequencing techniques from individual to multiomics analyses’, *Experimental & molecular medicine*, 52(9), pp. 1419–1427.

Katsanis, S.H. and Katsanis, N. (2013) ‘Molecular genetic testing and the future of clinical genomics’, *Nature reviews. Genetics*, 14(6), pp. 415–426.

Kaufmann, P., Pariser, A.R. and Austin, C. (2018) ‘From scientific discovery to treatments for rare diseases – the view from the National Center for Advancing Translational Sciences – Office of Rare Diseases Research’, *Orphanet Journal of Rare Diseases*. doi:10.1186/s13023-018-0936-x.

Korsunsky, I. *et al.* (2019) ‘Fast, sensitive and accurate integration of single-cell data with Harmony’, *Nature methods*, 16(12), pp. 1289–1296.

Lee, J., Hyeon, D.Y. and Hwang, D. (2020) ‘Single-cell multiomics: technologies and data analysis methods’, *Experimental & molecular medicine*, 52(9), pp. 1428–1442.

Matentzoglu, N. *et al.* (2022) ‘A Simple Standard for Sharing Ontological Mappings (SSSOM)’, *Database: the journal of biological databases and curation*, 2022. doi:10.1093/database/baac035.

Mathys, H. *et al.* (2019) ‘Single-cell transcriptomic analysis of Alzheimer’s disease’, *Nature*, 570(7761), pp. 332–337.

Mungall, C.J. *et al.* (2010) ‘Integrating phenotype ontologies across multiple species’, *Genome biology*, 11(1), p. R2.

Nguengang Wakap, S. *et al.* (2020) ‘Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database’, *European journal of human genetics: EJHG*, 28(2), pp. 165–173.

Nitzan, O. *et al.* (2015) ‘Urinary tract infections in patients with type 2 diabetes mellitus: review of prevalence, diagnosis, and management’, *Diabetes, metabolic syndrome and obesity: targets and therapy*, 8, pp. 129–136.

Niwattanakul and Singthongchai (no date) ‘Using of Jaccard coefficient for keywords similarity’, *Proceedings of the Estonian Academy of Sciences. Biology, ecology = Eesti Teaduste Akadeemia toimetised. Bioloogia, okoloogia* [Preprint]. Available at: https://www.researchgate.net/profile/Ekkachai-Naenudorn/publication/317248581_Using_of_Jaccard_Coefficient_for_Keywords_Similarity/links/592e560ba6fdcc89e759c6d0/Using-of-Jaccard-Coefficient-for-Keywords-Similarity.pdf.

Omary, M.B. *et al.* (2007) ‘The pancreatic stellate cell: a star on the rise in pancreatic diseases’, *The Journal of clinical investigation*, 117(1), pp. 50–59.

Pallotta, M.T. *et al.* (2019) ‘Wolfram syndrome, a rare neurodegenerative disease: from pathogenesis to future treatment perspectives’, *Journal of translational medicine*, 17(1), p. 238.

Panina, Y. *et al.* (2020) ‘Human Cell Atlas and cell-type authentication for regenerative medicine’, *Experimental & molecular medicine*, 52(9), pp. 1443–1451.

Pavan, S. *et al.* (2017) ‘Clinical Practice Guidelines for Rare Diseases: The Orphanet Database’, *PloS one*, 12(1), p. e0170365.

Perera, C.J. *et al.* (2021) 'Role of Pancreatic Stellate Cell-Derived Exosomes in Pancreatic Cancer-Related Diabetes: A Novel Hypothesis', *Cancers*, 13(20). doi:10.3390/cancers13205224.

Perry, V.H., Nicoll, J.A.R. and Holmes, C. (2010) 'Microglia in neurodegenerative disease', *Nature reviews. Neurology*, 6(4), pp. 193–201.

Perry, V.H. and Teeling, J. (2013) 'Microglia and macrophages of the central nervous system: the contribution of microglia priming and systemic inflammation to chronic neurodegeneration', *Seminars in immunopathology*, 35(5), pp. 601–612.

Real, R. and Vargas, J.M. (1996) 'The Probabilistic Basis of Jaccard's Index of Similarity', *Systematic biology*, 45(3), pp. 380–385.

Rouhana, J.M. *et al.* (2021) 'ECLIPSER: identifying causal cell types and genes for complex traits through single cell enrichment of e/sQTL-mapped genes in GWAS loci', *bioRxiv*. doi:10.1101/2021.11.24.469720.

Rubio-Villena, C. *et al.* (2018) 'Astrocytes: new players in progressive myoclonus epilepsy of Lafora type', *Human molecular genetics*, 27(7), pp. 1290–1300.

Salter, M.W. and Stevens, B. (2017) 'Microglia emerge as central players in brain disease', *Nature medicine*, 23(9), pp. 1018–1027.

Sievert, C. (2020) *Interactive Web-Based Data Visualization with R, plotly, and shiny*. CRC Press.

Skene, N.G. and 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, p. 16.

Slavotinek, A.M. (2002) 'Fraser syndrome and cryptophthalmos: review of the diagnostic criteria and evidence for phenotypic modules in complex malformation syndromes', *Journal of Medical Genetics*, pp. 623–633. doi:10.1136/jmg.39.9.623.

Snijders, T. *et al.* (2015) 'Satellite cells in human skeletal muscle plasticity', *Frontiers in Physiology*. doi:10.3389/fphys.2015.00283.

Su, H. *et al.* (2021) 'Emerging Role of Pericytes and Their Secretome in the Heart', *Cells*, p. 548. doi:10.3390/cells10030548.

Tabula Muris Consortium *et al.* (2018) 'Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris', *Nature*, 562(7727), pp. 367–372.

Tabula Sapiens Consortium* *et al.* (2022) 'The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans', *Science*, 376(6594), p. eabl4896.

Takao, M. *et al.* (2000) 'Neuroserpin mutation S52R causes neuroserpin accumulation in neurons and is associated with progressive myoclonus epilepsy', *Journal of neuropathology and experimental neurology*, 59(12), pp. 1070–1086.

Thorens, B. (2014) 'Neural regulation of pancreatic islet cell mass and function', *Diabetes, obesity & metabolism*, 16 Suppl 1, pp. 87–95.

Urano, F. (2016) 'Wolfram Syndrome: Diagnosis, Management, and Treatment', *Current diabetes reports*, 16(1), p. 6.

Vasilevsky, N.A. *et al.* (2022) ‘Mondo: Unifying diseases for the world, by the world’, *bioRxiv*. doi:10.1101/2022.04.13.22273750.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wolf, S.A., Boddeke, H.W.G.M. and Kettenmann, H. (2017) ‘Microglia in Physiology and Disease’, *Annual review of physiology*, 79, pp. 619–643.

Yang, Y. *et al.* (2020) ‘Pancreatic stellate cells in the islets as a novel target to preserve the pancreatic β-cell mass and function’, *Journal of diabetes investigation*, 11(2), pp. 268–280.

Yin, H., Price, F. and Rudnicki, M.A. (2013) ‘Satellite Cells and the Muscle Stem Cell Niche’, *Physiological Reviews*, pp. 23–67. doi:10.1152/physrev.00043.2011.

Zhang, J. *et al.* (2022) ‘Glutathione prevents high glucose-induced pancreatic fibrosis by suppressing pancreatic stellate cell activation via the ROS/TGFβ/SMAD pathway’, *Cell death & disease*, 13(5), p. 440.