

Integrative Analysis of GWAS Risk Genes in the Human Protein Atlas

Zhuoran Cai

2026-02-13

This report is generated automatically from the GWAS and Human Protein Atlas (HPA) analysis pipeline.

It is meant as: - a **reference** for what the code can do, and
- a **summary** of the main results for a given disease or trait.

The report assumes that:

- risk genes were retrieved from the **GWAS Catalog**, expression and cluster information were retrieved from **HPA** (can be done through the searching script), and
- enrichment and specificity analyses were run using the provided scripts.

Full datasets can be found in the exported file for further investigation.

0. Input and Search Terms

This section lists the key inputs used to generate this report: EFO terms used for GWAS risk gene retrieval; Date of the GWAS search; Human Protein Atlas (HPA) version used for expression data.

```
## # A tibble: 7 x 3
##   efo_id      trait      uri
##   <chr>      <chr>      <chr>
## 1 EFO_0006514 Alzheimer's disease biomarker measurement http://www.ebi.ac.uk~
## 2 EFO_0006801 Alzheimer's disease neuropathologic change http://www.ebi.ac.uk~
## 3 MONDO_0004975 Alzheimer disease http://purl.obolibrary~
## 4 EFO_1001870 late-onset Alzheimer's disease http://www.ebi.ac.uk~
## 5 EFO_0009268 family history of Alzheimer's disease http://www.ebi.ac.uk~
## 6 OBA_2001000 age of onset of Alzheimer disease http://purl.obolibrary~
## 7 EFO_0022957 early-onset Alzheimers disease http://www.ebi.ac.uk~

## [1] "2026-02-13 18:38:39 CET"

## [1] "https://v24.proteinatlas.org/api/search_download.php"
```

1. Summary of statistic

This section gives a quick overview of how many risk genes were identified, how many were found in HPA, and how many could be mapped onto the brain expression UMAP.

```
## Total genes in gene list: 530
## Risk genes found in HPA: 476
## Risk genes that can be mapped in brain UMAP: 453
```

2. What clusters do risk gene enrich?

2a. Perform Over-representation analysis (ORA) using fisher's exact test to see enriched clusters.

-An over-representation analysis is used to determine which clusters are significantly enriched. Then, fold enrichment score for each cluster is calculated. **Fold Enrichment** measures how much the observed overlap exceeds random expectation. It is calculated as:

$$\text{Fold Enrichment} = \frac{\text{Observed Overlap}}{\text{Expected Overlap}}$$

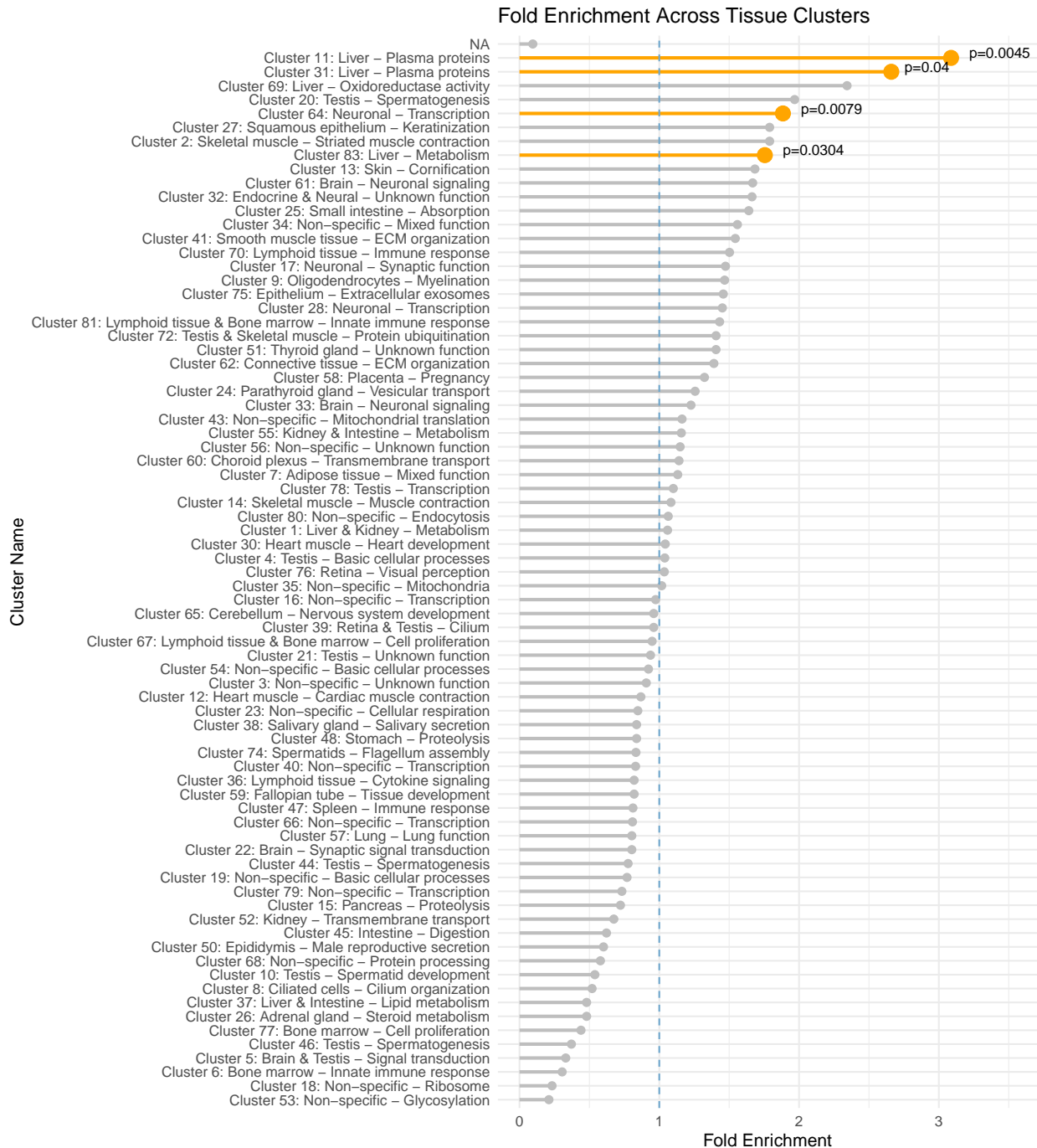
Where:

$$\text{Expected Overlap} = \frac{\text{Risk Genes} \times \text{Cluster Size}}{\text{Total Genes}}$$

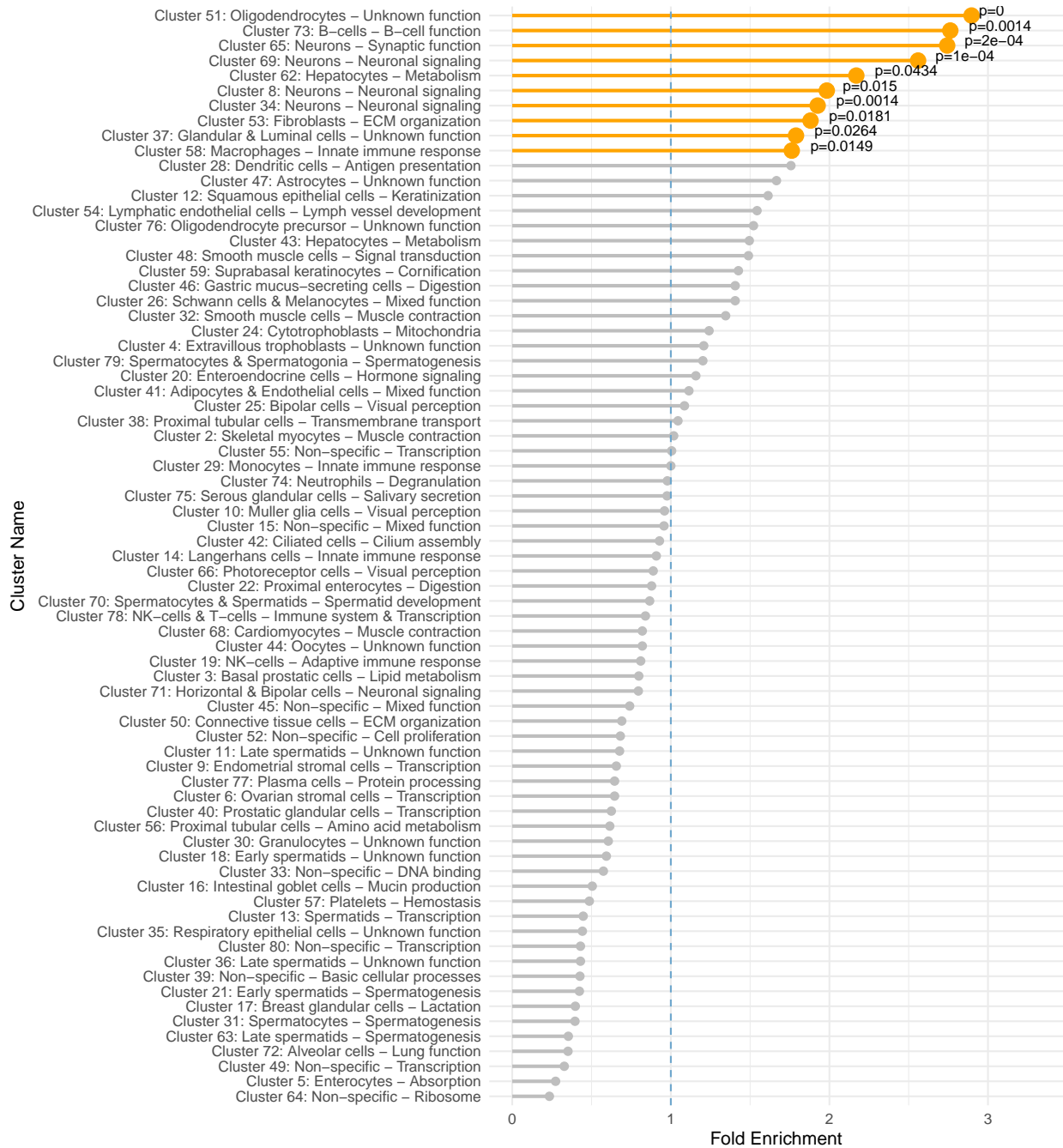
- **Fold Enrichment** > 1: Indicates enrichment.
- **Fold Enrichment** = 1: Indicates no enrichment.
- **Fold Enrichment** < 1: Indicates depletion.

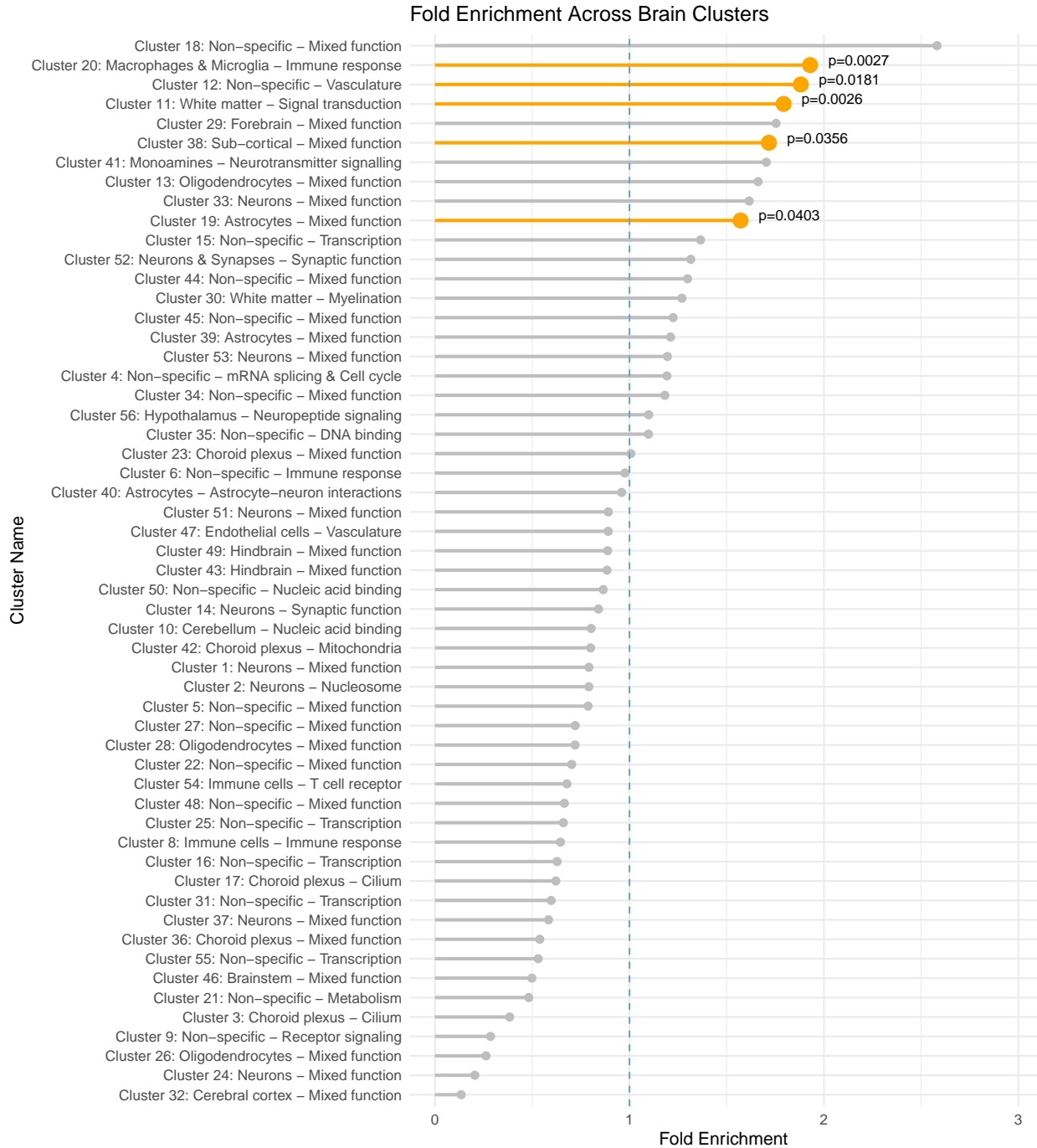
In the lollipop plots below, enriched ($p < 0.05$) clusters are highlighted in orange. x-axis represents fold enrichment.

- In the lollipop plots, each dot is a cluster. The x-axis shows fold enrichment and the y-axis shows the cluster label. Orange dots indicate significantly enriched clusters ($p < 0.05$). Note that since this pipeline is initially designed for HPA brain section/brain related diseases, genes not found in the HPA brain section database are excluded from further analysis.



Fold Enrichment Across Single Cell Clusters





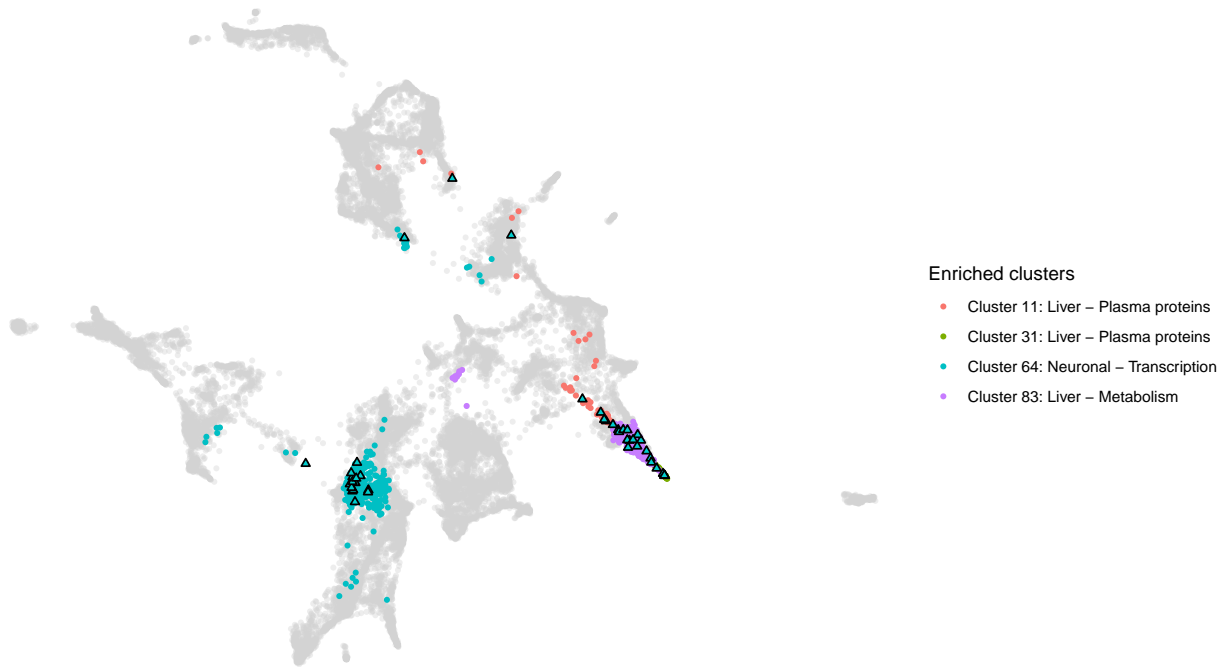
2b. Visualization of enriched clusters

The UMAP plots below show:

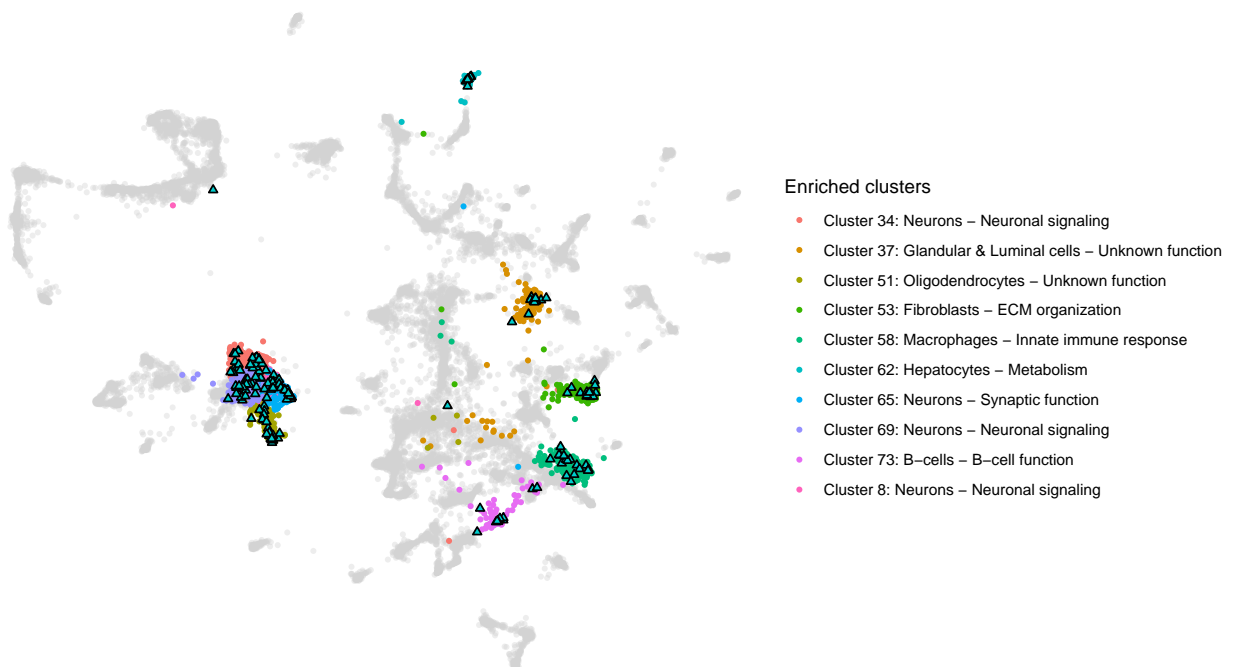
- All genes (grey)
- Genes in enriched clusters (colored)
- Risk genes within enriched clusters (cyan triangles)

These plots highlight where risk genes concentrate in transcriptomic space.

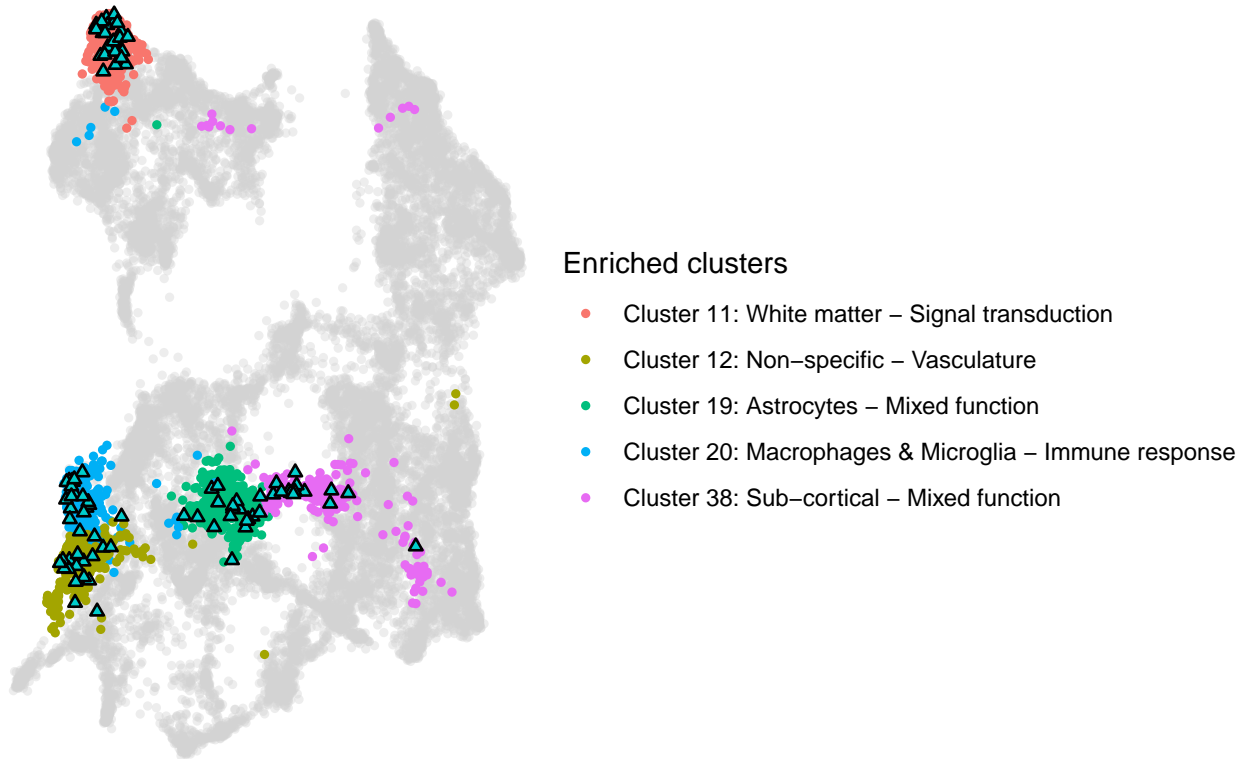
UMAP Plot (Tissue v24.1): Enriched clusters with risk genes



UMAP Plot (Single-cell v24.1): Enriched clusters with risk genes



UMAP Plot (Brain v24.1): Enriched clusters with risk genes



3. Specificity of Risk Genes

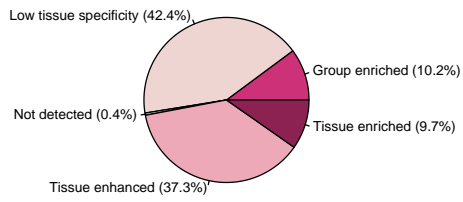
HPA assigns genes to specificity categories that describe how selectively they are expressed across tissues or cell types:

- **Tissue/Cell-type Enriched** genes shows markedly higher expression in one tissue or cell type compared to all others
- **Group enriched genes** displayed elevated expression in a small group of related tissues or cell types
- **Tissue/Cell-typeEnhanced** genes showed increased expression compared to the average across all tissues or cell types
- Genes that did not meet these thresholds were considered to have low specificity or broadexpression. We examine how risk genes are distributed across these categories in tissues, single-cell types, and brain datasets.

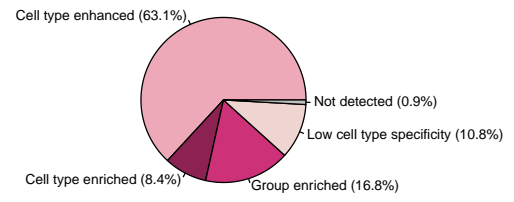
3a. Specificity component of risk genes

These plots show how many risk genes fall into each specificity class in different expression contexts.

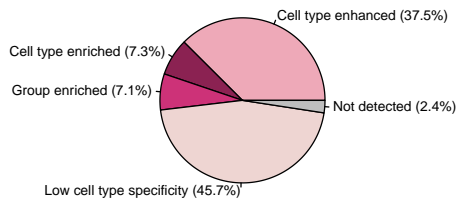
RNA.tissue.specifcicity



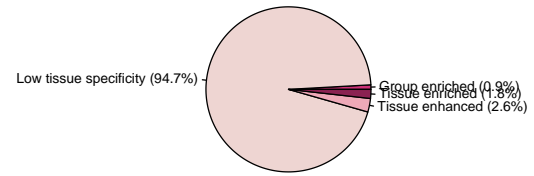
RNA.single.cell.type.specifcicity



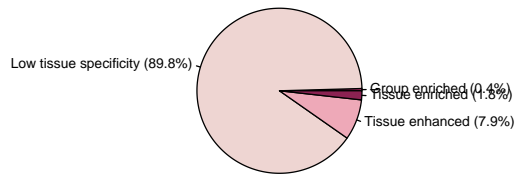
RNA.single.nuclei.brain.specifcicity



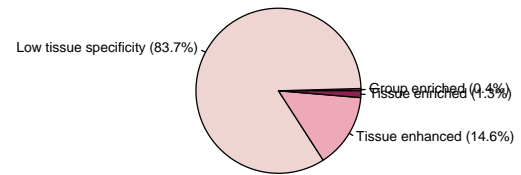
Human.brain.regional.RNA.Specifcicity



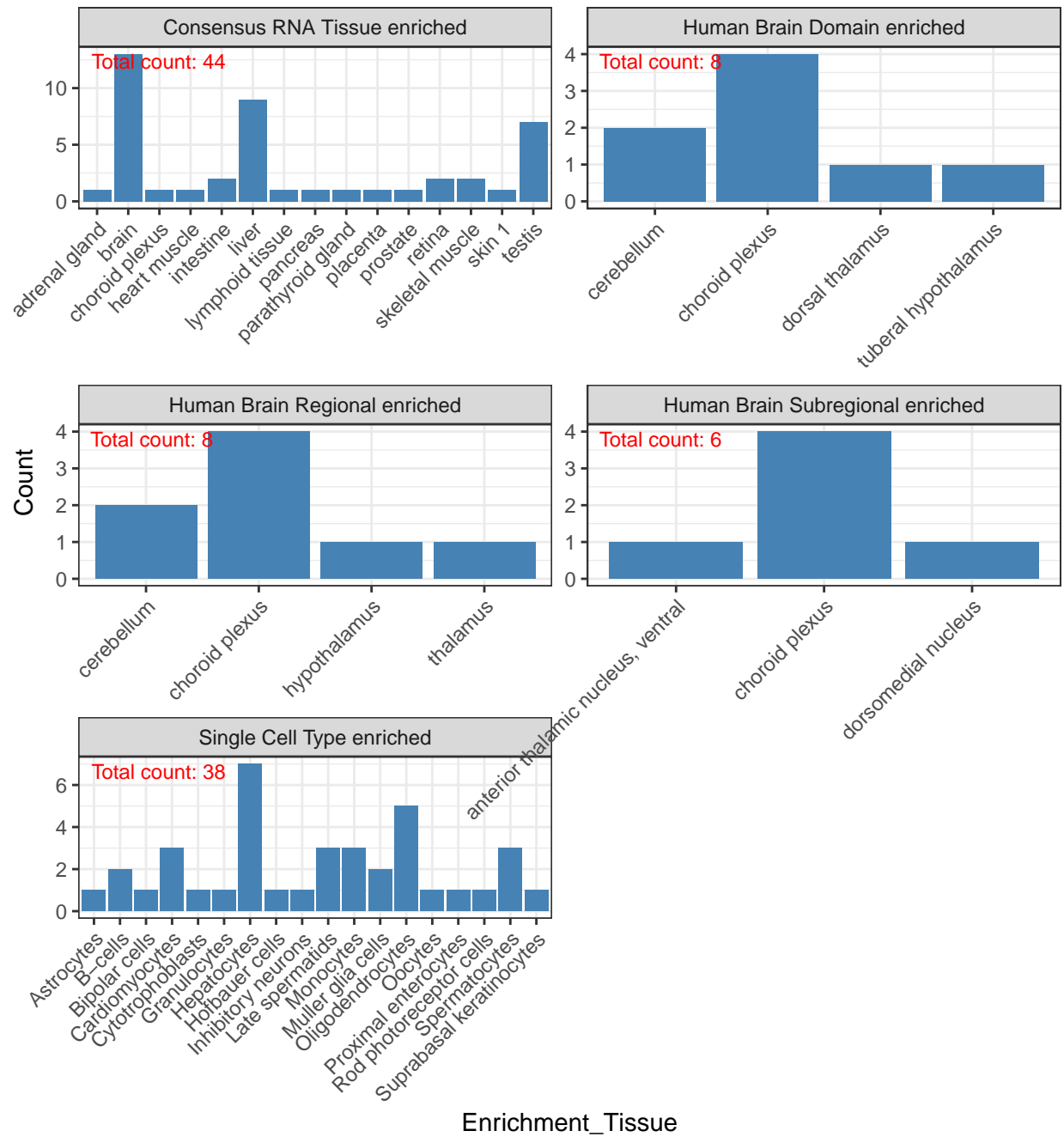
Human.brain.domain.RNA.Specifcicity



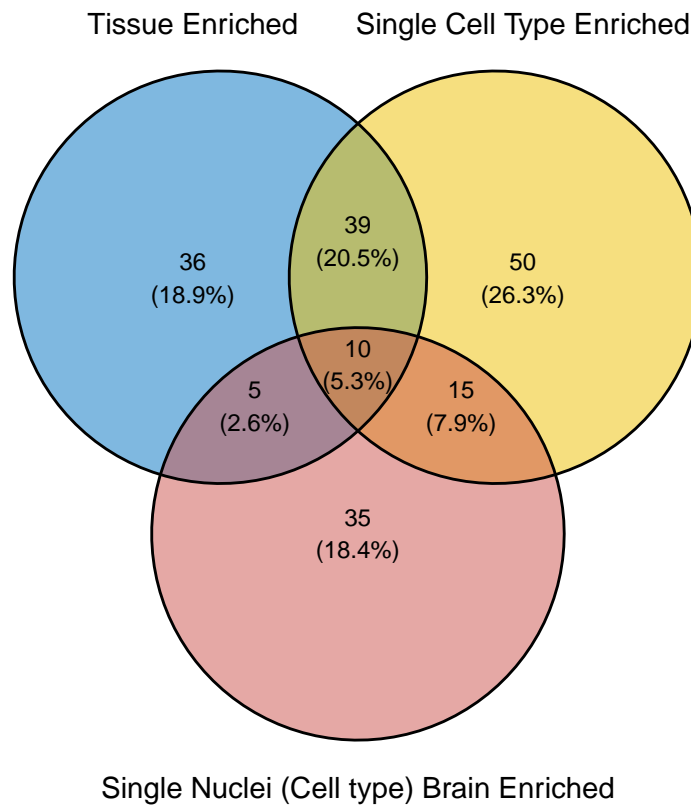
Human.brain.subregional.RNA.Specifcicity



3b. Detailed component of feature enriched genes (group enriched excluded)



Venn diagram of enriched genes (feature enriched+group enriched)



Venn diagram of enriched genes (feature enriched+group enriched)

