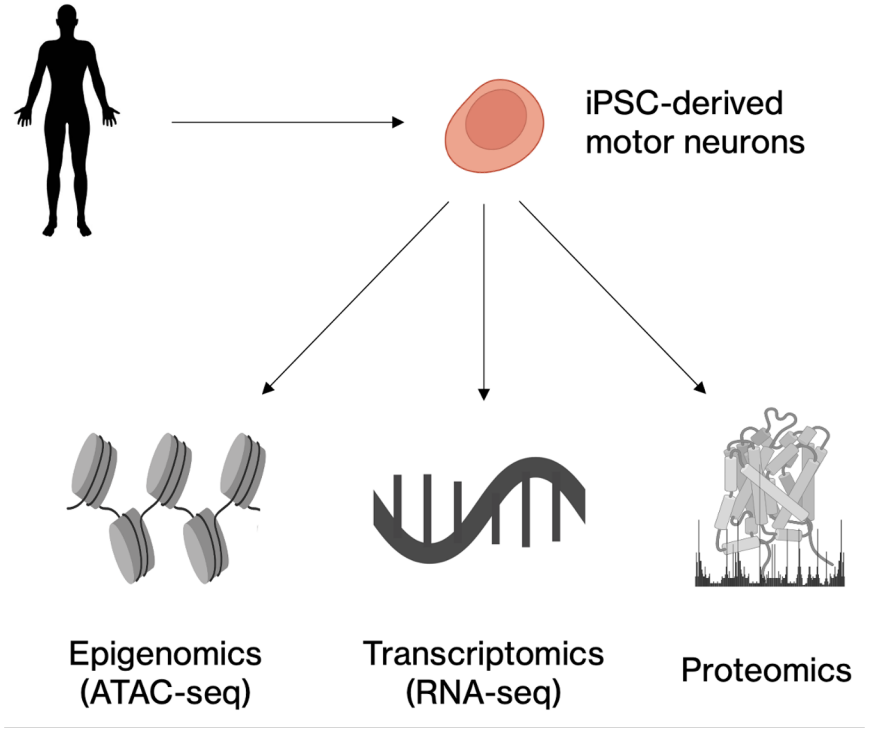
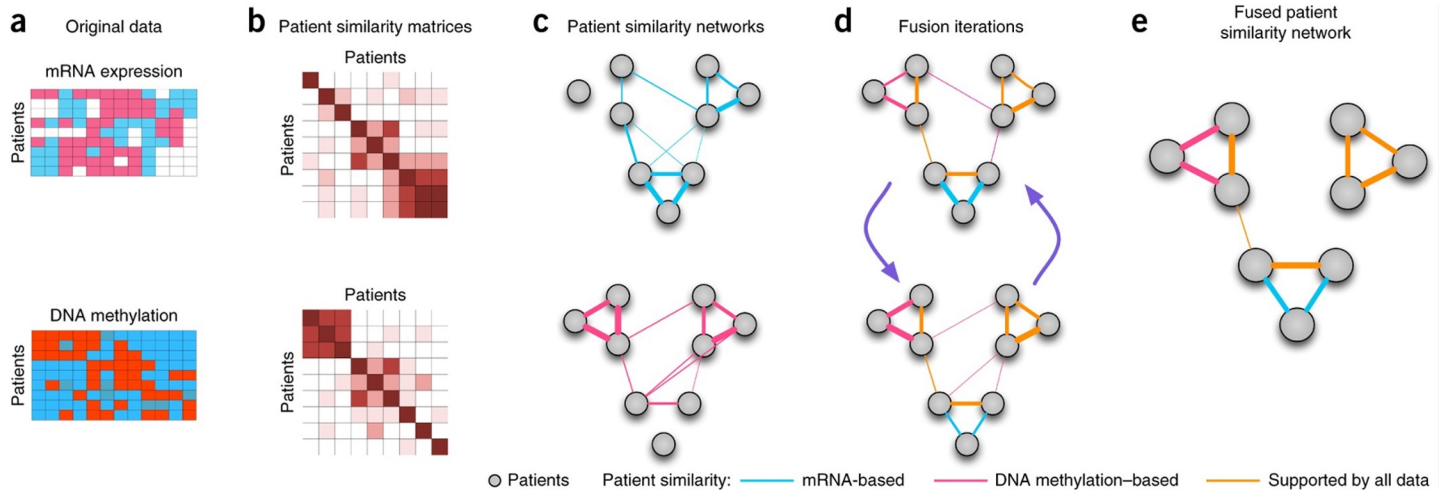
ALS Subtype Discovery Through Multi-Omic Integration and Clustering

Bridget Li, Velina Kozareva, Dr. Ernest Fraenkel

Data source: Answer ALS



Similarity Network Fusion (SNF)



Iteratively update status matrices:

$$\mathbf{P}_{t+1}^{(1)} = \mathbf{S}^{(1)} \times \mathbf{P}_t^{(2)} \times (\mathbf{S}^{(1)})^T$$

$$\mathbf{P}_{t+1}^{(2)} = \mathbf{S}^{(2)} \times \mathbf{P}_t^{(1)} \times (\mathbf{S}^{(2)})^T$$

After t steps, overall status matrix:

$$\mathbf{P}^{(c)} = \frac{\mathbf{P}_t^{(1)} + \mathbf{P}_t^{(2)}}{2}.$$

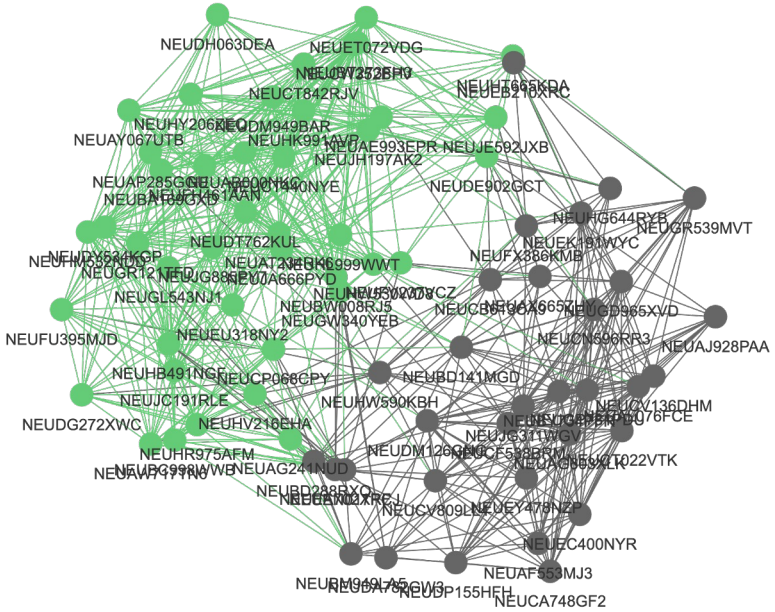
Using SNF to detect ALS subtypes

- Apply similarity network fusion
  - Integrate proteomics, transcriptomics, and epigenomics datasets
  - Incorporate patient data, such as sex, age at onset, site of onset, etc. from metadata
- Find clusters of patients that represent subtypes of ALS
- Identify molecular signatures that differentiate clusters

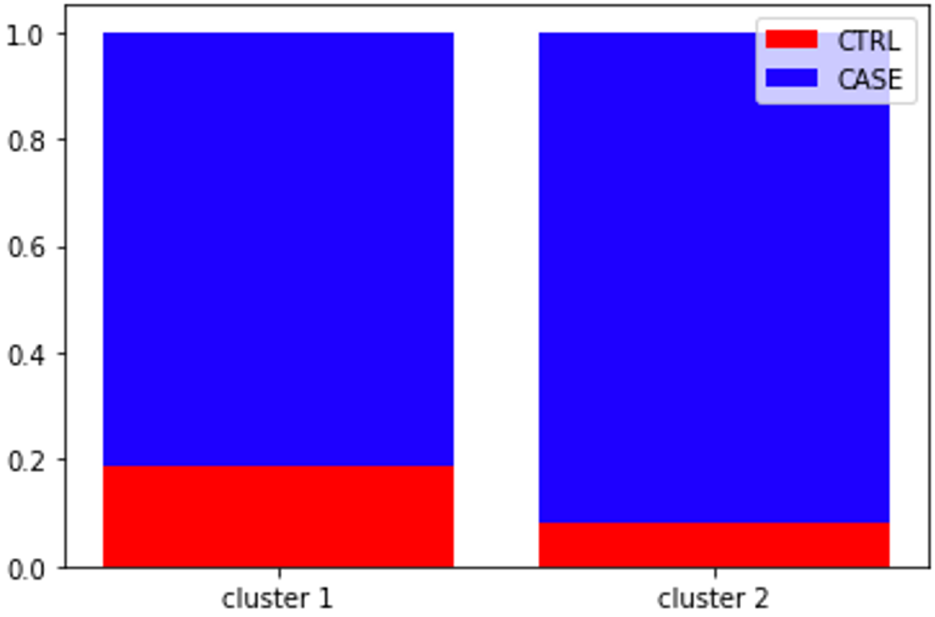
SNF: Iteration #1

SNF network visualization

- 196 patients
- 2 clusters
  - Cluster 1 size: 95
  - Cluster 2 size: 101
- NMI (case/control): 0.0248
- Silhouette score = 0.234

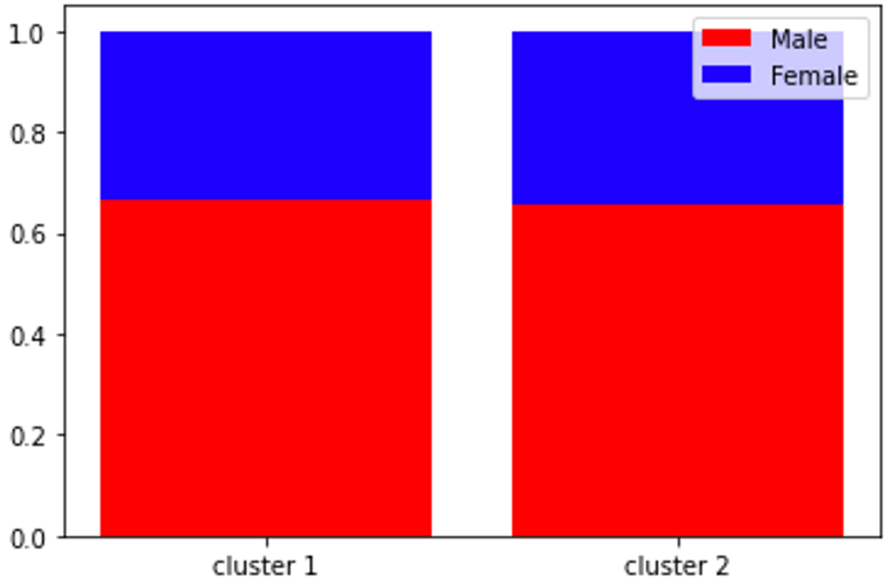


Slight cluster separation based on case/control



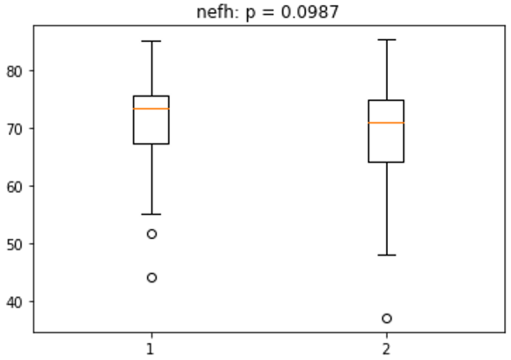
No cluster separation based on sex

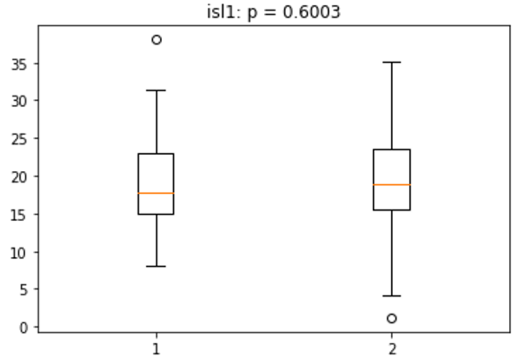
- Males are more likely to have ALS
- Answer ALS has an even stronger male case imbalance



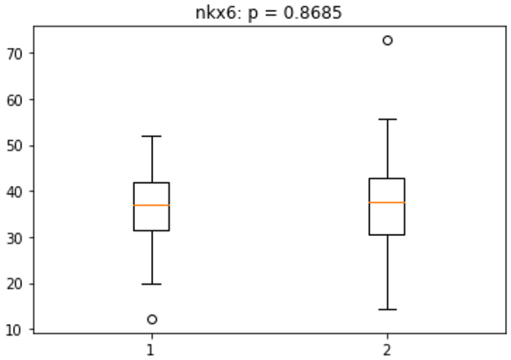
Clusters are correlated with covariates s100b and tuj1

- Staining markers that indicate cell type

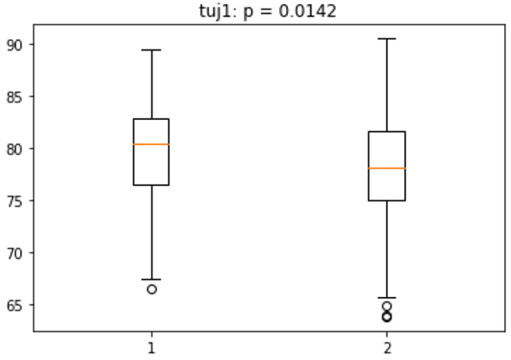




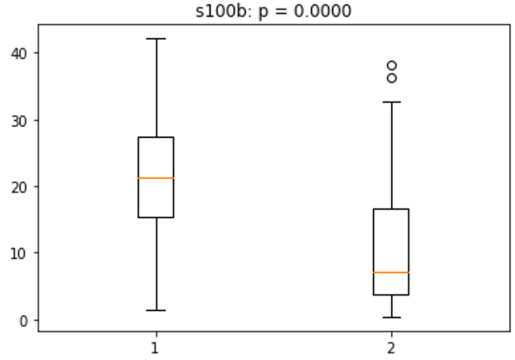
Motor neuron



Ventral progenitor



General neuron



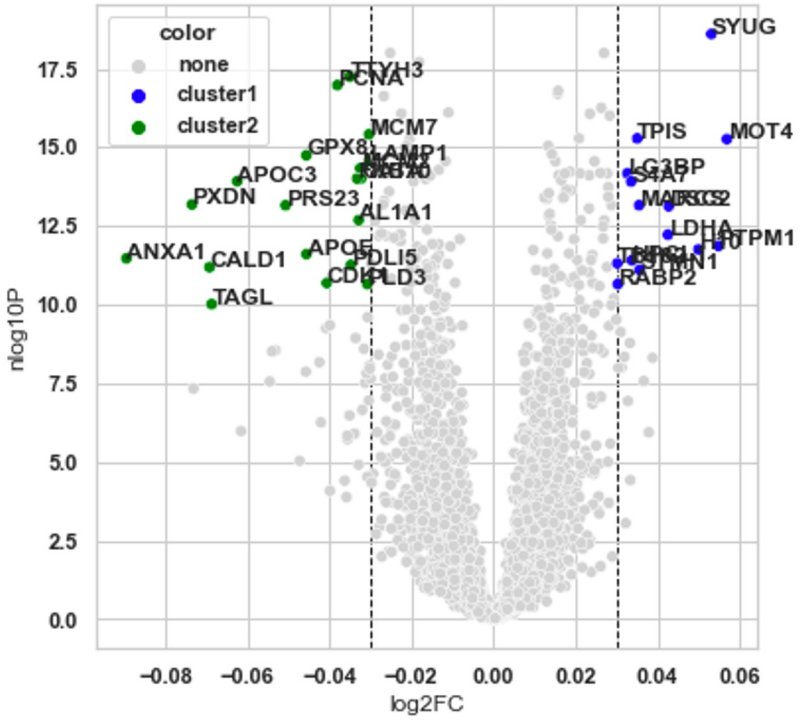
Glial cell

Differential expression analysis

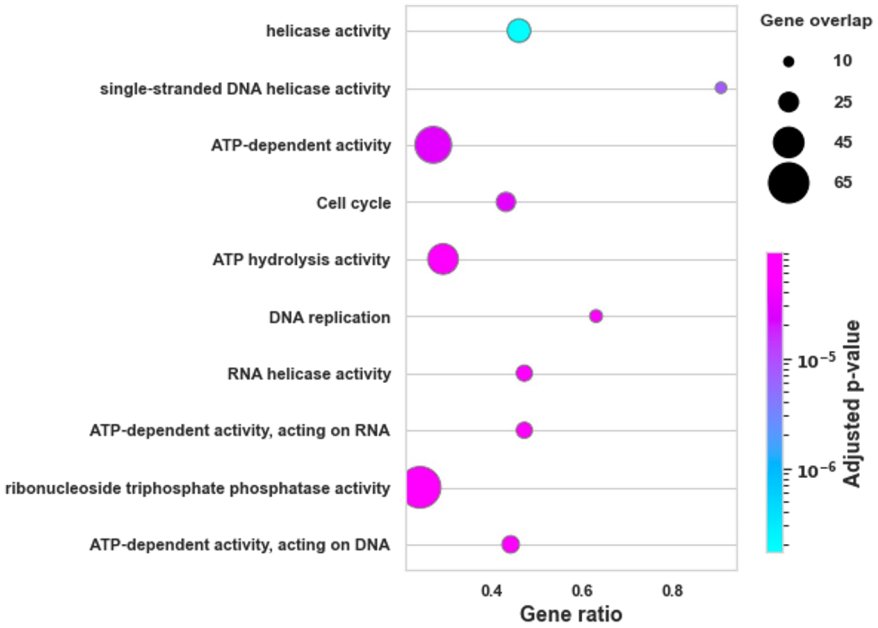
	Proteomics	**Transcriptomics**	**Epigenomics**
Total #	3,496	24,711	100,363
# differentially expressed at FDR=0.1	2,026	17,912	60,551
# in cluster 1	981	9,184	30,568
# in cluster 2	1,045	8,728	29,983

- T-test + Benjamini-Hochberg multiple test correction

Proteomics volcano plot



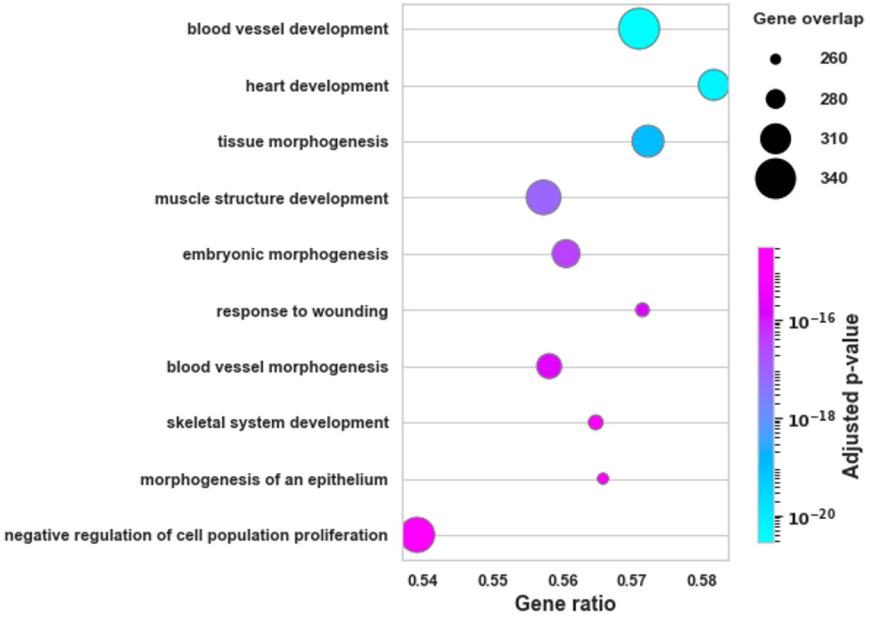
Proteomics GO terms



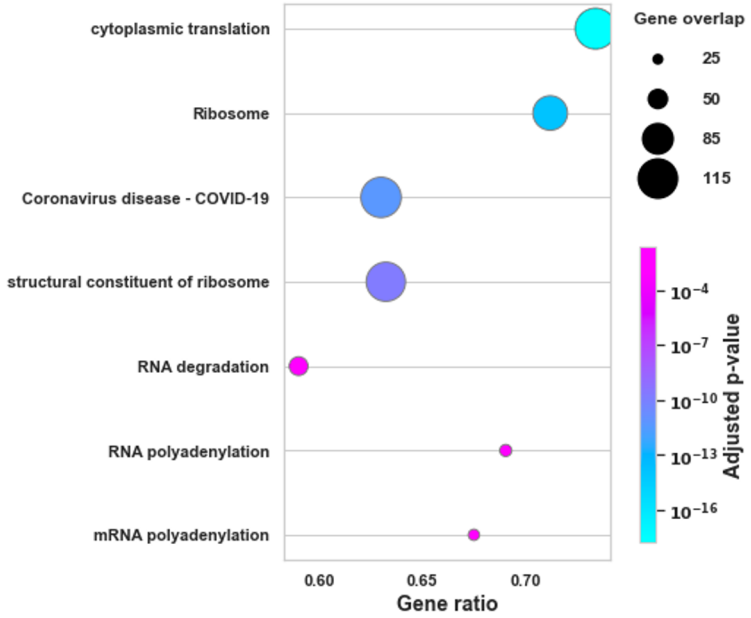
Cluster 1: helicase activity, ATP-dependent activity

Cluster 2: no significant GO terms against background

Transcriptomics GO terms

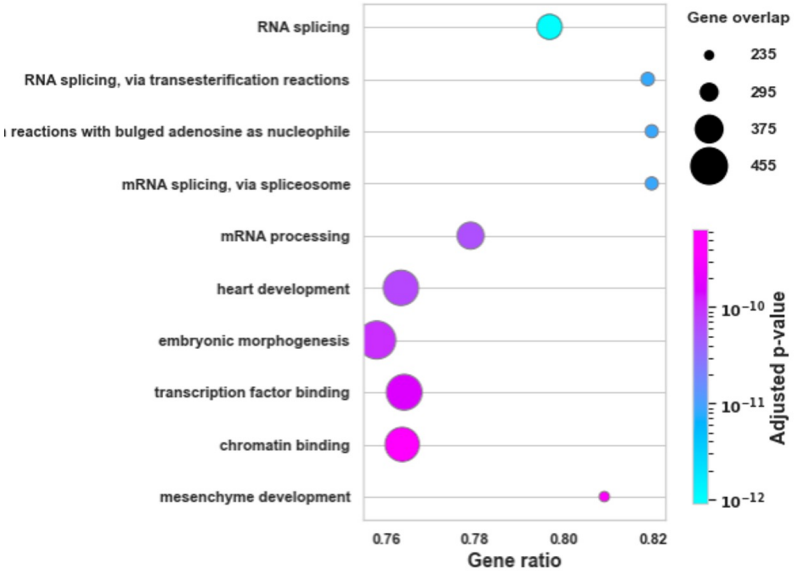


Cluster 1: development

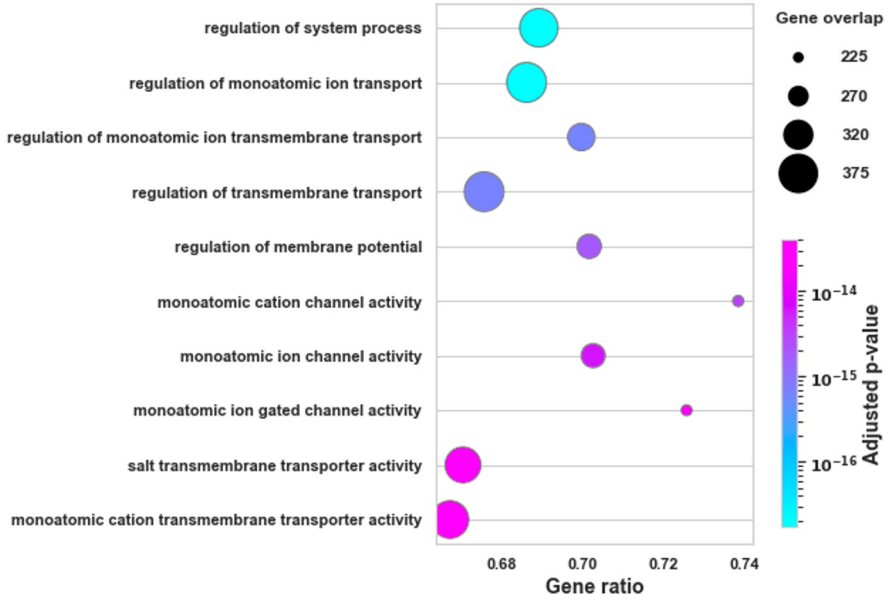


Cluster 2: translation, RNA degradation

Epigenomics GO terms



Cluster 1: transcription, development



Cluster 2: ion transport

GO terms across different modalities don’t show concordance

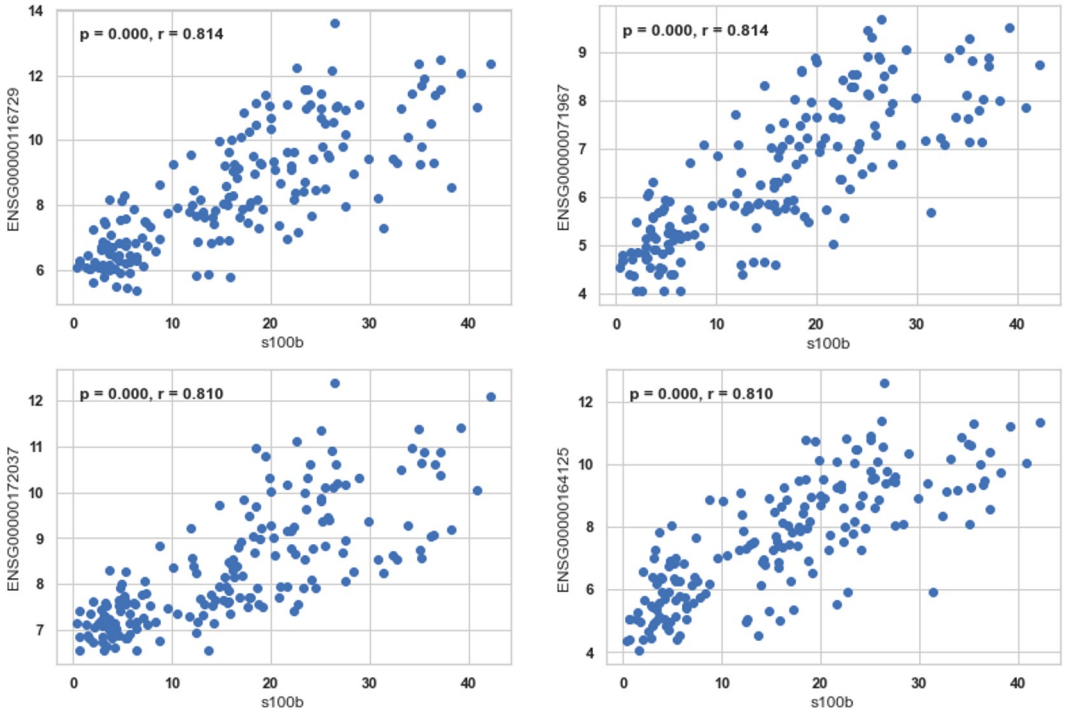
	Cluster 1	**Cluster 2**
Proteomics	Helicase activity, ATP-dependent activity	N/A
Transcriptomics	Development, response to wounding	Translation, RNA degradation
Epigenomics	RNA splicing, development	Ion transport

- Different modalities capture different snapshots of human biology
- Low correlation between protein and gene expression is expected

Clusters had high correlation with s100b

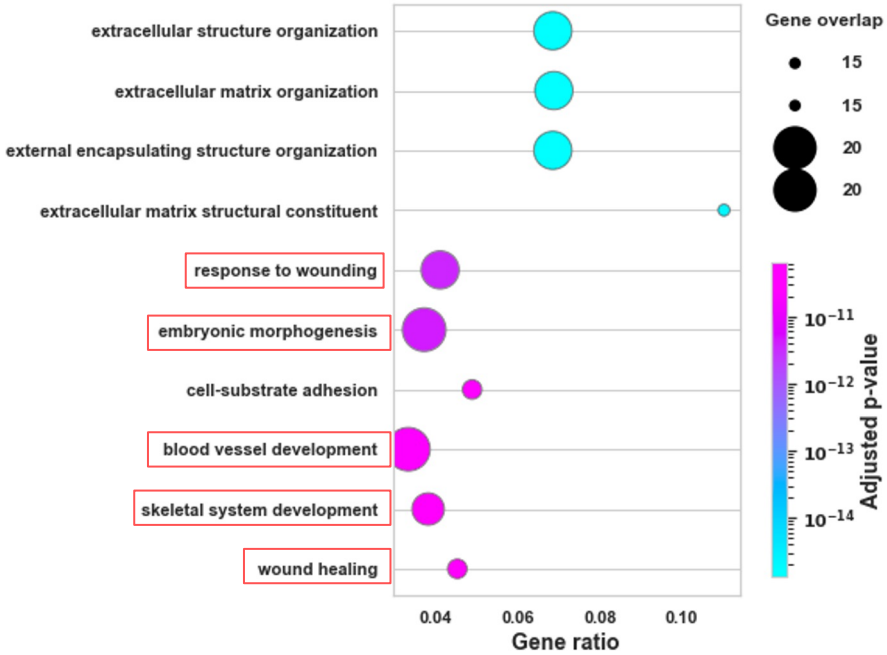
```
nefh
highly correlated genes: 0
overlap w/ DEGs: 0
isl1
highly correlated genes: 0
overlap w/ DEGs: 0
nkx6
highly correlated genes: 0
overlap w/ DEGs: 0
tuj1
highly correlated genes: 0
overlap w/ DEGs: 0
s100b
highly correlated genes: 286
overlap w/ DEGs: 286
```

r threshold = 0.7

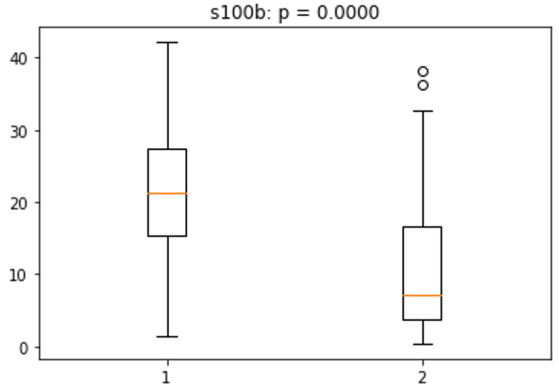


Clusters had high correlation with s100b

GO terms for genes highly correlated with s100b



Overlaps with cluster 1, where there was higher s100b expression



Accounting for s100b

- Uncontrollable differences between batches drive heterogeneity in data, masking relevant biological differences

Attempted here:

- From limma: remove batch effect
- Removing principal components that are most highly correlated with s100b

Other avenues:

- Incorporating s100b bias into SNF model
- Fuse another status matrix with covariates we want to remove

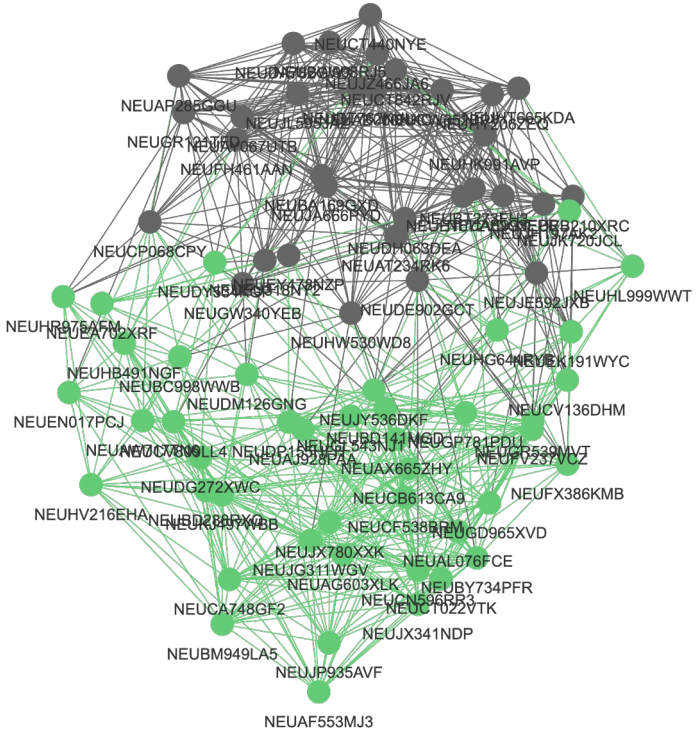
Regress out s100b

Remove unwanted s100b batch effect from each modality

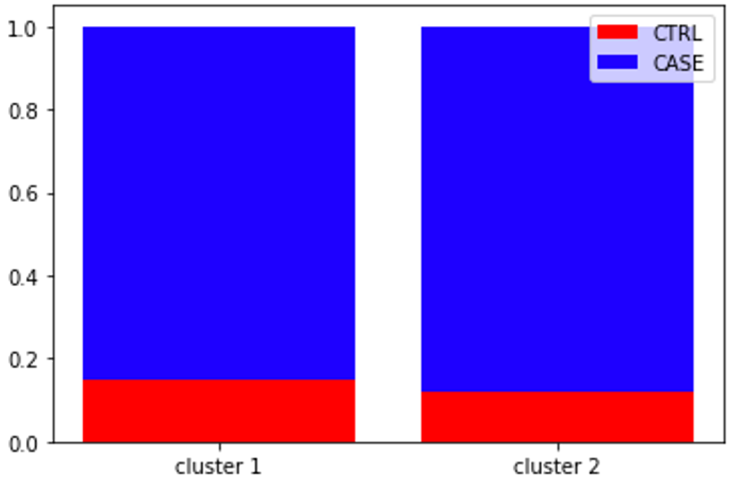
- Remove patients with missing s100b data (196 → 182)
- Apply removeBatchEffect to all three modalities
  - Fit a linear model to each feature in the dataset using the covariates we want to remove
  - Use residuals from that linear model instead of original data
- Rerun SNF and differential expression analysis

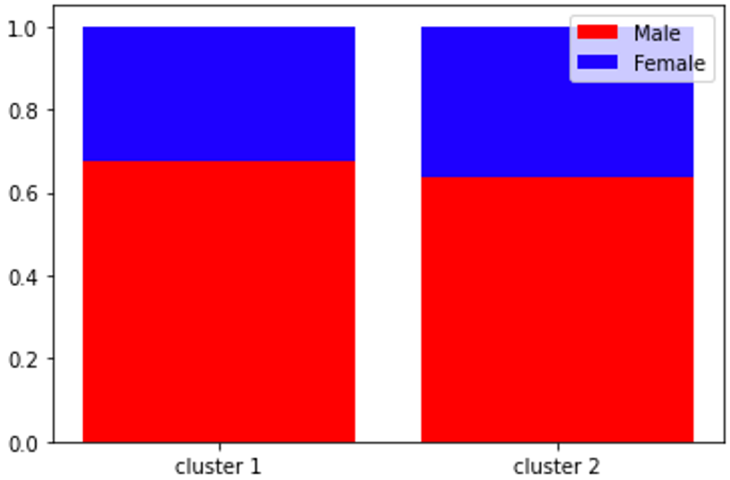
SNF network visualization

- 182 patients
- 2 clusters
  - Cluster 1 size: 108
  - Cluster 2 size: 74
- NMI (case/control): 0.00135
- Silhouette score = 0.181

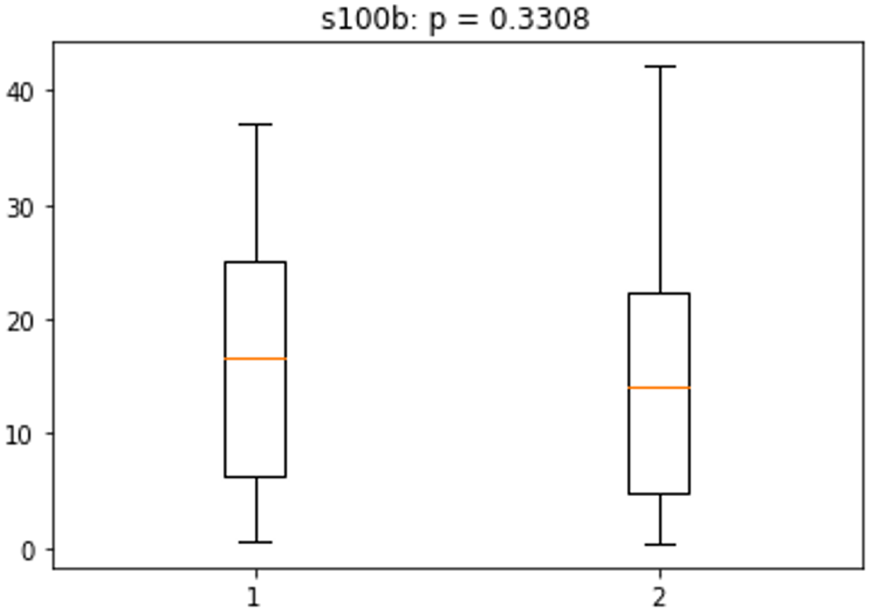


No cluster separation based on case/control or sex





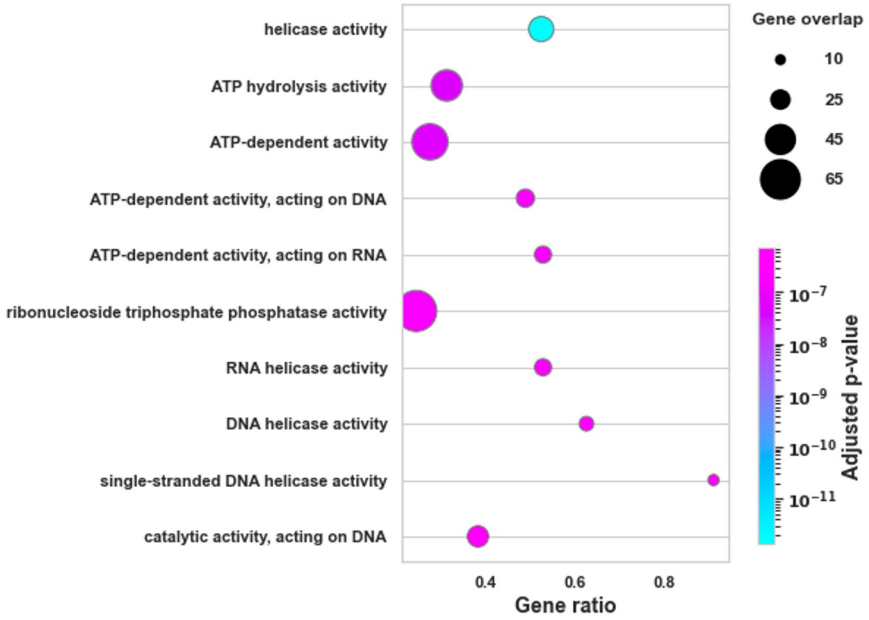
s100b was successfully regressed out



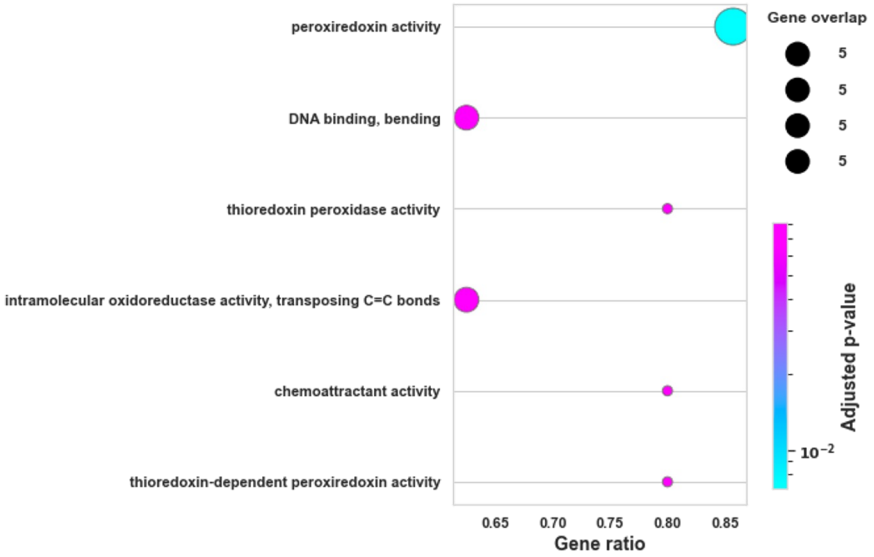
Differential expression analysis

	Proteomics	**Transcriptomics**	**Epigenomics**
Total #	3,496	24,711	100,363
# differentially expressed at FDR=0.1	1,733	17,290	56,405
# in cluster 1	882	9,401	28,820
# in cluster 2	851	7,889	27,585

Proteomics GO terms

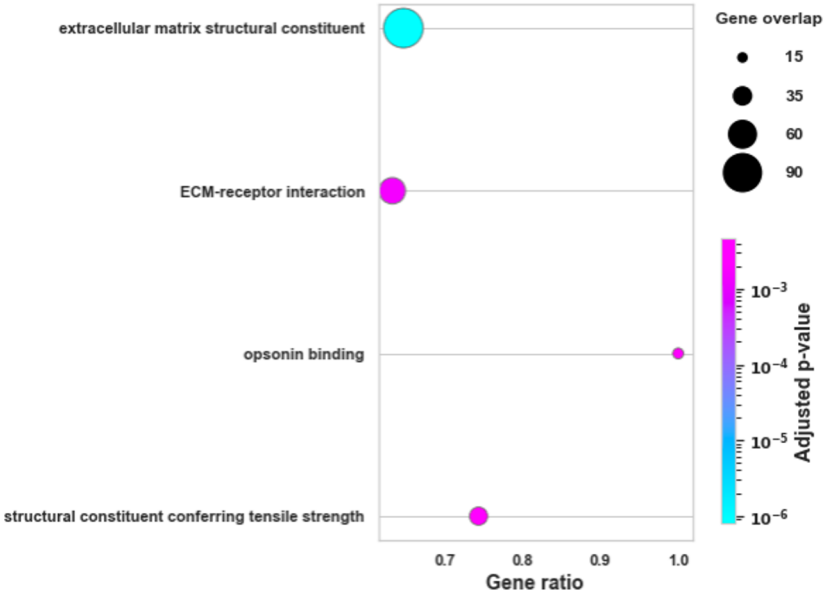


Cluster 1: helicase activity, ATP-dependent activity

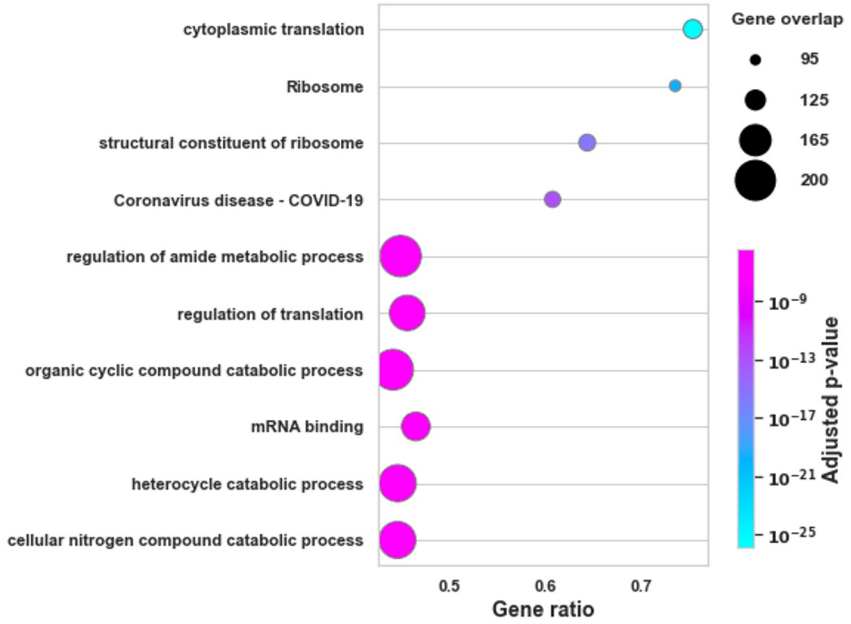


Cluster 2: peroxiredoxin

Transcriptomics GO terms

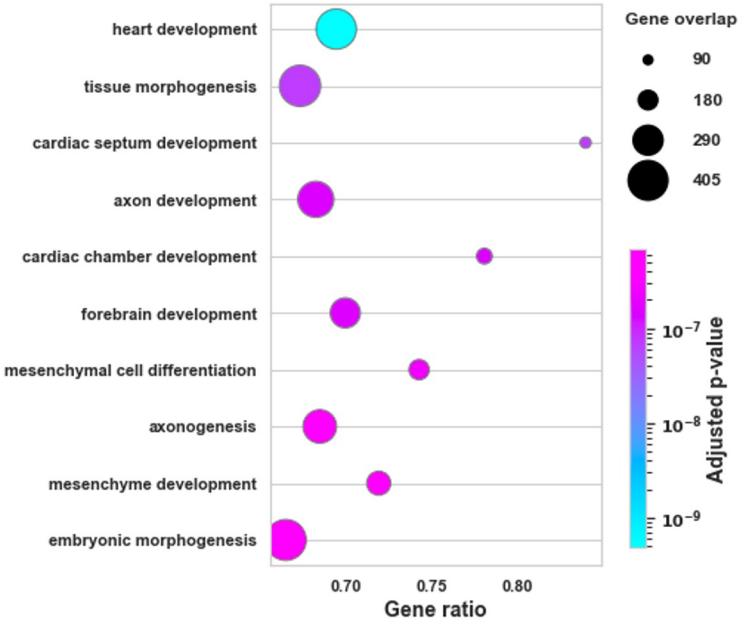


Cluster 1: ECM

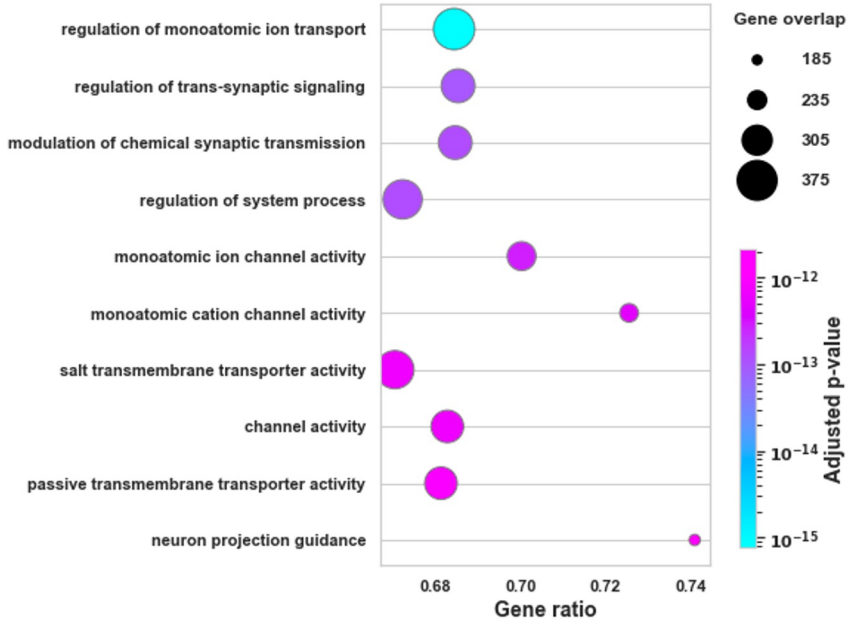


Cluster 2: translation, catabolic processes

Epigenomics GO terms



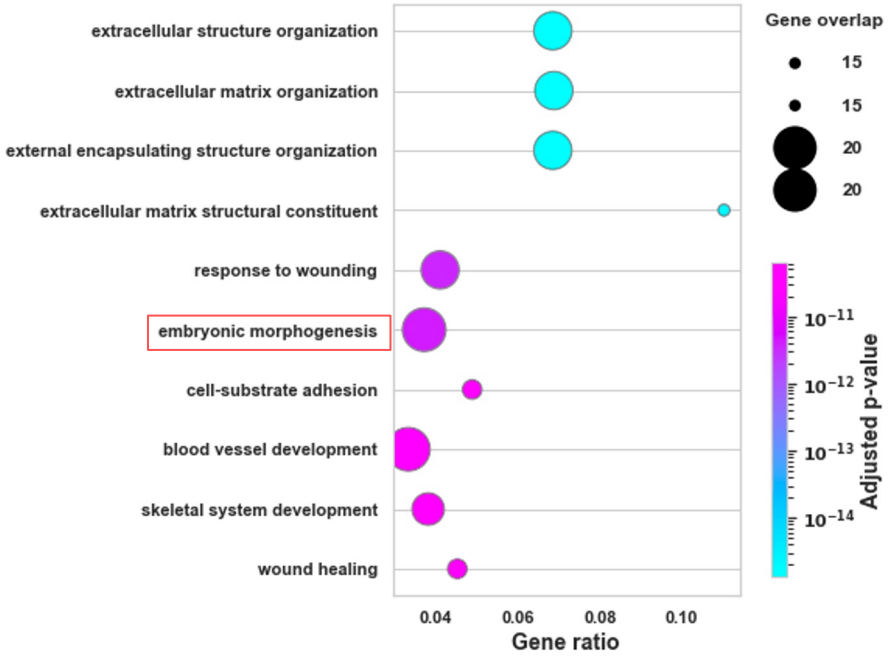
Cluster 1: development



Cluster 2: ion transport

Clusters had low correlation with s100b

GO terms for genes highly correlated with s100b



Little overlap with either cluster

Top GO terms of new clusters align with original SNF clusters

	Cluster 1	**Cluster 2**
Proteomics	Helicase activity, ATP-dependent activity	peroxiredoxin
Transcriptomics	Development	Translation, catabolic processes
Epigenomics	Development	Ion transport

- Clusters have similar top GO terms to original despite s100b being regressed out

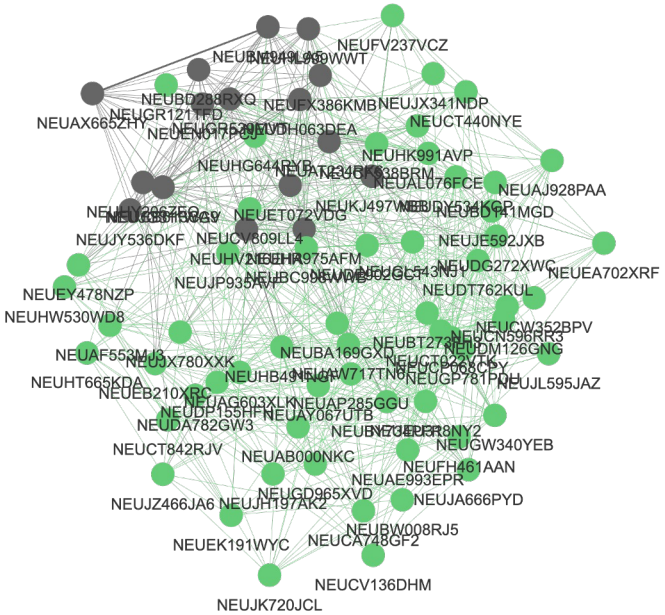
PCA to remove s100b

Remove principal components highly correlated with s100b from each modality

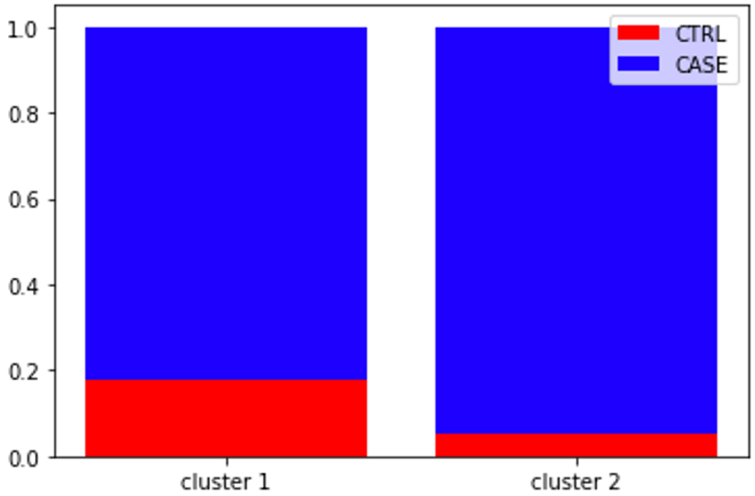
- Remove patients with missing s100b data (196 → 182)
- Keep 50% of features with highest variance
- Run PCA and get top 50 PCs for each modality
- Remove PCs with |correlation with s100b| > 0.3
- Rerun SNF on remaining PCs

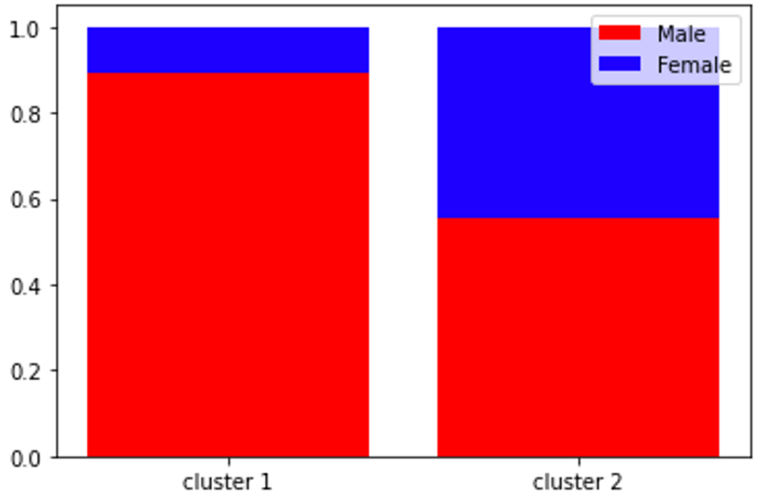
SNF network visualization

- 182 patients
- 2 clusters
  - Cluster 1 size: 135
  - Cluster 2 size: 47
- NMI (case/control): 0.0544
- Silhouette score = 0.033

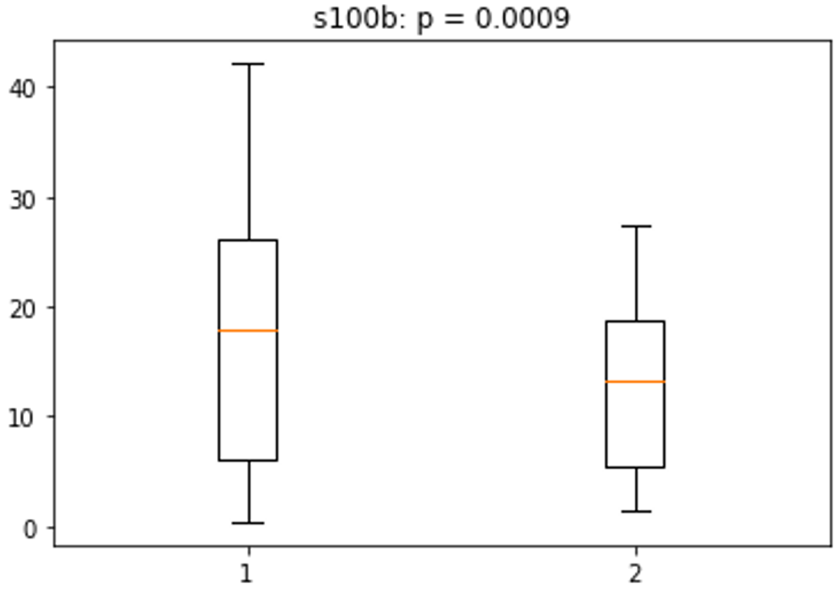


More apparent cluster separation by case/control and sex





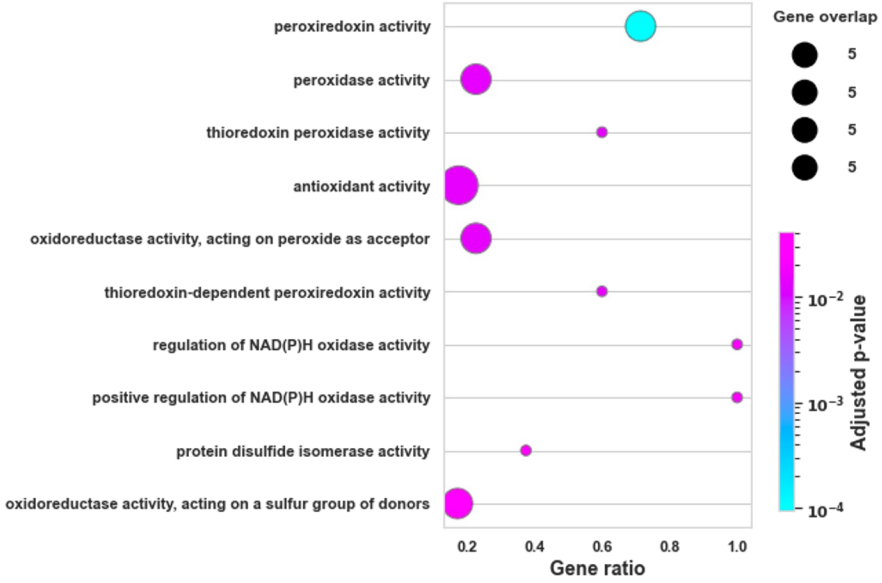
s100b dependency was lowered but still significant



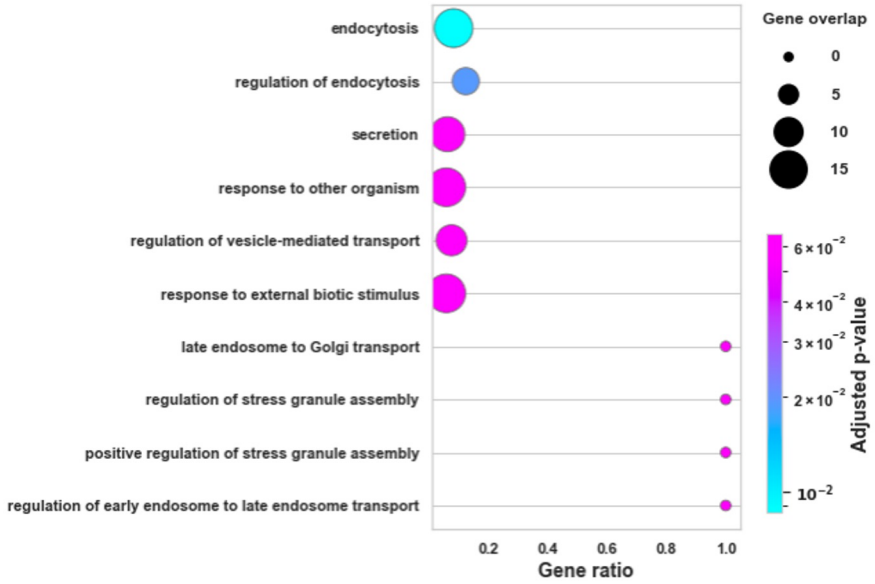
Differential expression analysis

	Proteomics	**Transcriptomics**	**Epigenomics**
Total #	3,496	24,711	100,363
# differentially expressed at FDR=0.1	412	11,199	28,145
# in cluster 1	187	5,557	13,787
# in cluster 2	225	5,642	14,358

Proteomics GO terms

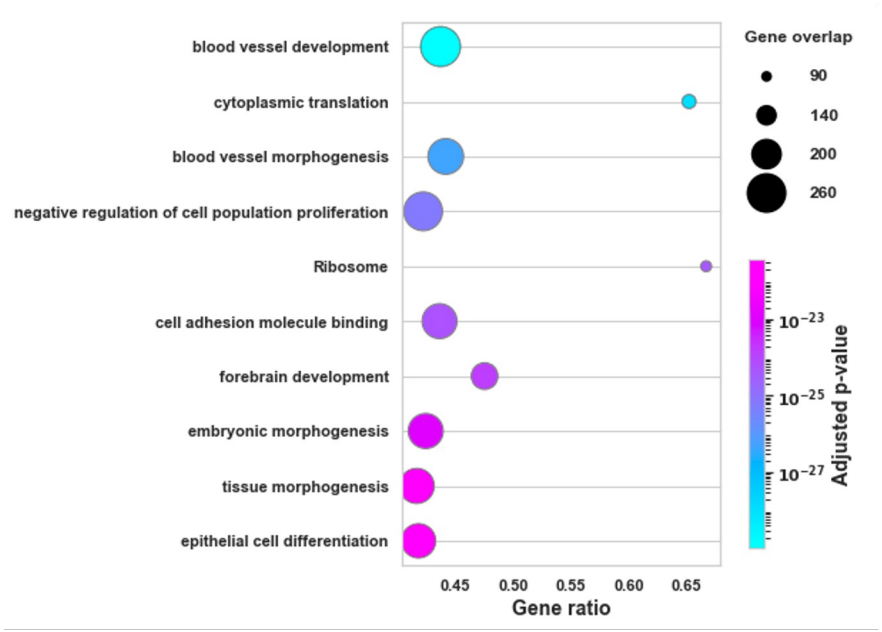


Cluster 1: peroxiredoxin, oxidoreductase

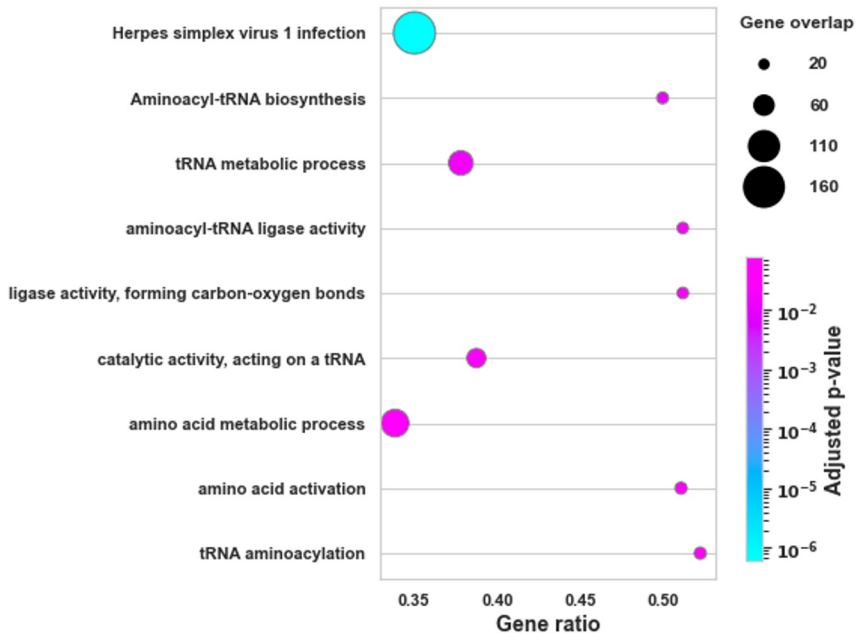


Cluster 2: endocytosis, vesicle-mediated transport

Transcriptomics GO terms

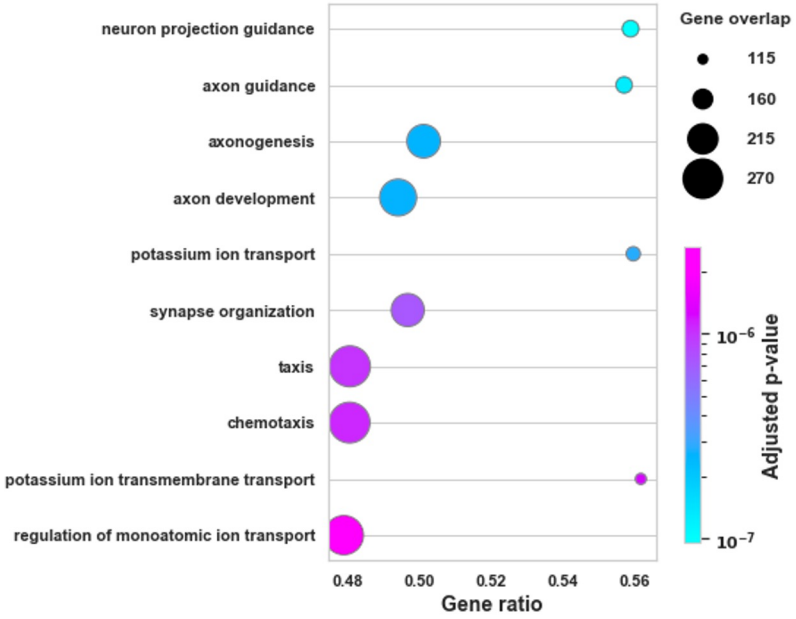


Cluster 1: development, translation

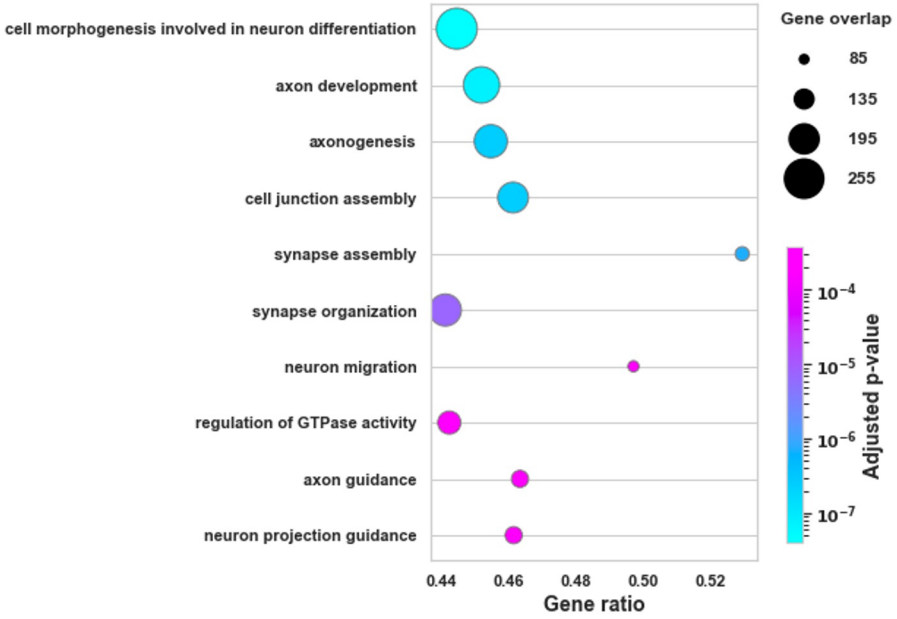


Cluster 2: tRNA/amino acid processes

Epigenomics GO terms: poor separation



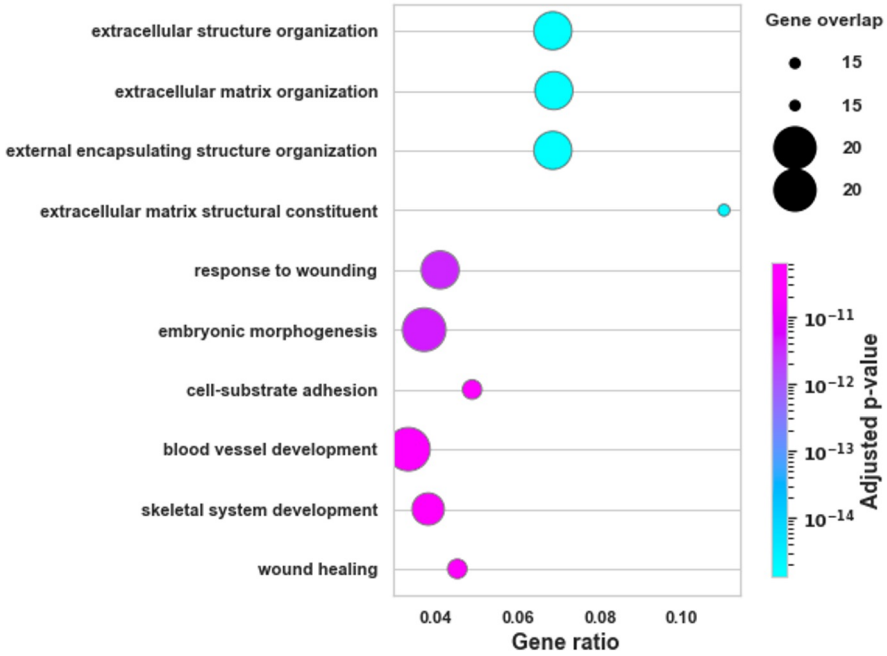
Cluster 1: axon, ion transport



Cluster 2: axon, synapse, neurons

Clusters had low correlation with s100b

GO terms for genes highly correlated with s100b



Little overlap with either cluster

Top GO terms of PCA clusters align poorly with regressed SNF clusters

	Cluster 1	**Cluster 2**
Proteomics	Peroxiredoxin, oxidoreductase	endocytosis, vesicle-mediated transport
Transcriptomics	Development, translation	tRNA/amino acid processes
Epigenomics	Axon, ion transport	Axon, synapse, neuron processes

- This may be because cluster 2 is now smaller (47 vs. 74 for regressed SNF)

Discussion of PCA SNF

- sklearn PCA is not deterministic ➔ varying results based on random_state
  - Occasionally outputted 3 clusters
- PCA SNF separated clusters more based on case/control and sex
  - May want to remove sex-related signals: drop PCs highly correlated with sex
  - Determine whether PCA SNF is able to separate based on case/control alone
- Few controls present in dataset ➔ PCA SNF case separation may be unreliable
  - Intrinsically difficult to separate based on case/control
  - Young motor neurons used in dataset- may be unrepresentative of ALS development
  - Regressing out s100b via limma did not yield separation based on case/control

Future Work

Future Work

- Apply SNF to only cases
  - Originally used for cancer patients to predict survival
- Use network results in a continuous manner
  - Trajectory inference from single-cell omics data: compare properties of cells over pseudotime
  - Make ordering of patients within each cluster ➔ does the ordering correspond to ordering by clinical progression?
- How are patients with similar genetic backgrounds spread across the network?
  - Some patients have mutations in known ALS genes (SOD1, c9orf72)
- Add clinical progression data as a separate modality
- Incorporate other datasets

References

Balendra, R., & Isaacs, A. M. (2018). C9orf72-mediated ALS and FTD: multiple pathways to disease. *Nature Reviews Neurology*, *14*(9), 544–558. https://doi.org/10.1038/s41582-018-0047-2

Chierici, M., Bussola, N., Marcolini, A., Francescatto, M., Zandonà, A., Trastulla, L., Agostinelli, C., Jurman, G., & Furlanello, C. (2020). Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.01065

Deconinck, L., Cannoodt, R., Saelens, W., Deplancke, B., & Saeys, Y. (2021). Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology*, *27*, 100344. https://doi.org/10.1016/j.coisb.2021.05.005

Gregory, J. M., Fagegaltier, D., Phatnani, H., & Harms, M. B. (2020). Genetics of Amyotrophic Lateral Sclerosis. *Current Genetic Medicine Reports*, *8*(4), 121–131. https://doi.org/10.1007/s40142-020-00194-8

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Smyth, G., & de Graaf, C. (2020, November 8). *removeBatchEffect: Remove Batch Effect in limma: Linear Models for Microarray Data*. Rdrr.io. https://rdrr.io/bioc/limma/man/removeBatchEffect.html

Wang, B., Jiang, J., Wang, W., Zhou, Z.-H., & Tu, Z. (2012). Unsupervised metric fusion by cross diffusion. *IEEE Xplore*, 2997–3004. https://doi.org/10.1109/CVPR.2012.6248029

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, *11*(3), 333–337. https://doi.org/10.1038/nmeth.2810