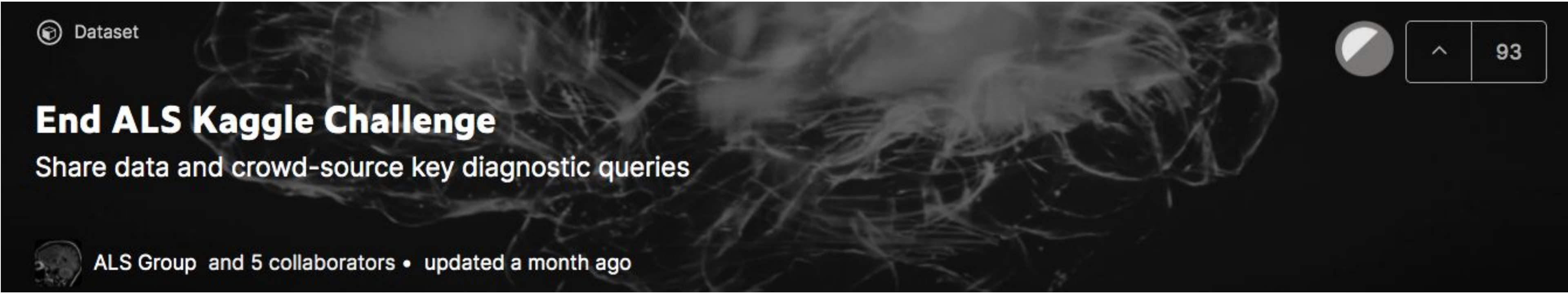
When the outlier is the signal



End ALS Kaggle Challenge

Share data and crowd-source key diagnostic queries

ALS Group and 5 collaborators • updated a month ago

Searching new genetic causes of ALS by aberrant gene expression analysis

15th May 2021

TUM - UCI team

Task 1

Does ALS have one mechanism of action?

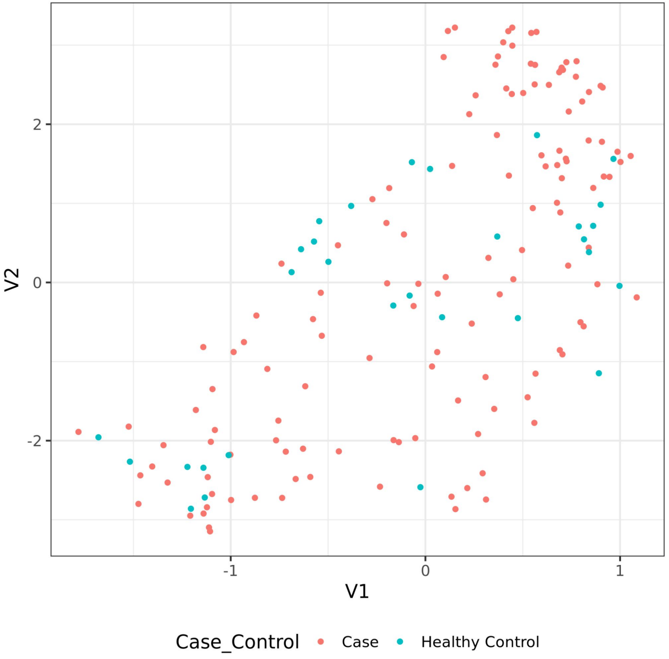
→ one pathway

Or is it caused by multiple independent or different mechanisms of action?

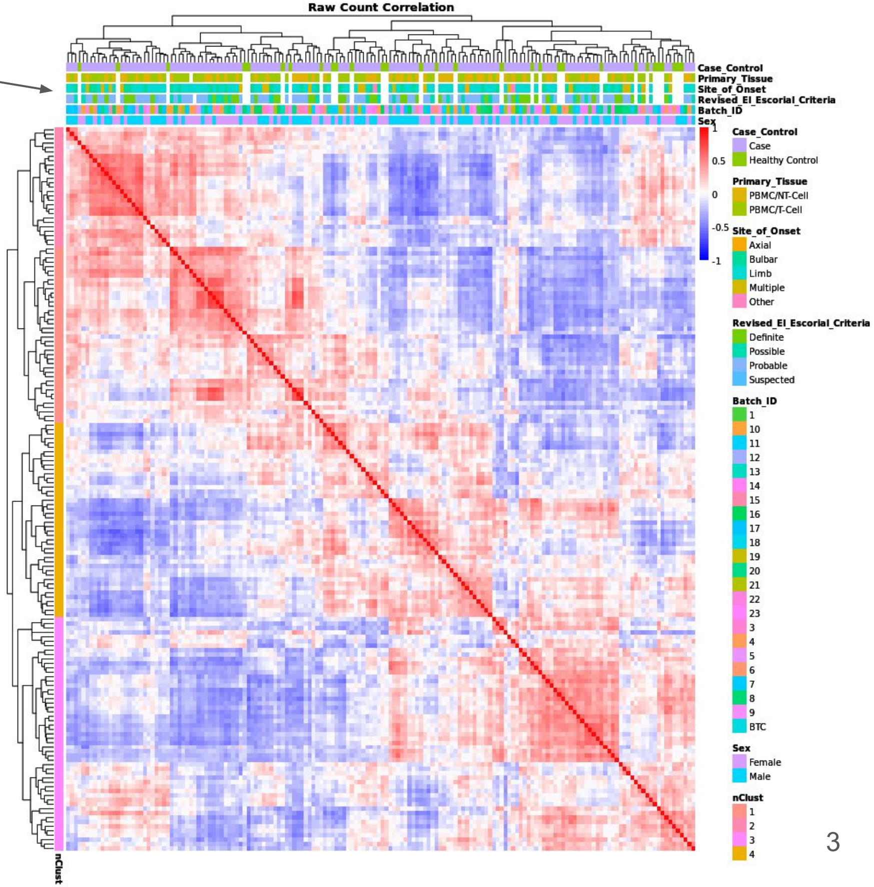
→ multiple pathways

Gene expression does not naturally cluster donor groups

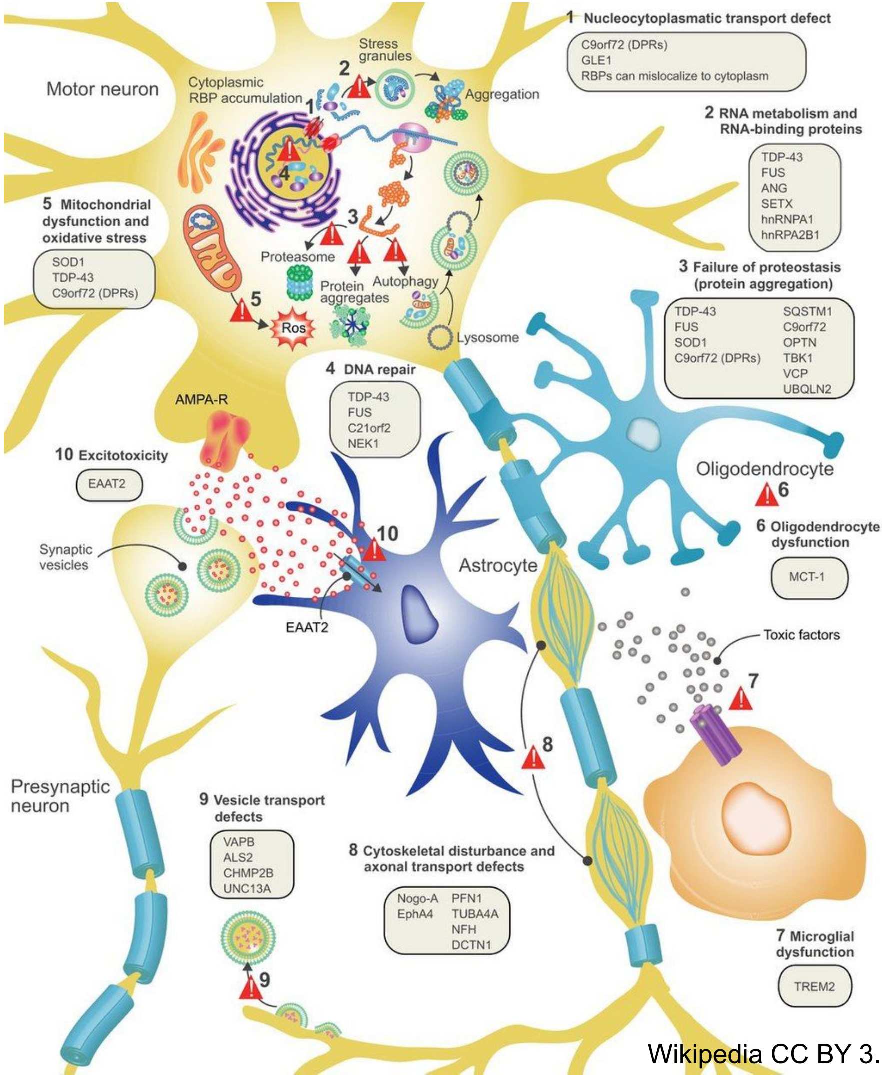
Cases and controls visualized in first two principal components (V1 and V2)



No covariate drives the sample clusters



In fact, mutations in ca. 40 genes have been implicated in ALS over various molecular pathways[1]



Wikipedia CC BY 3.0

1. Gregory et al. Curr Genet Med Rep (2020)
2. Hardiman O et al. Nat. Rev. Dis. Primers (2017)
3. van Damme et al. Disease Models and Mechanisms (2017)

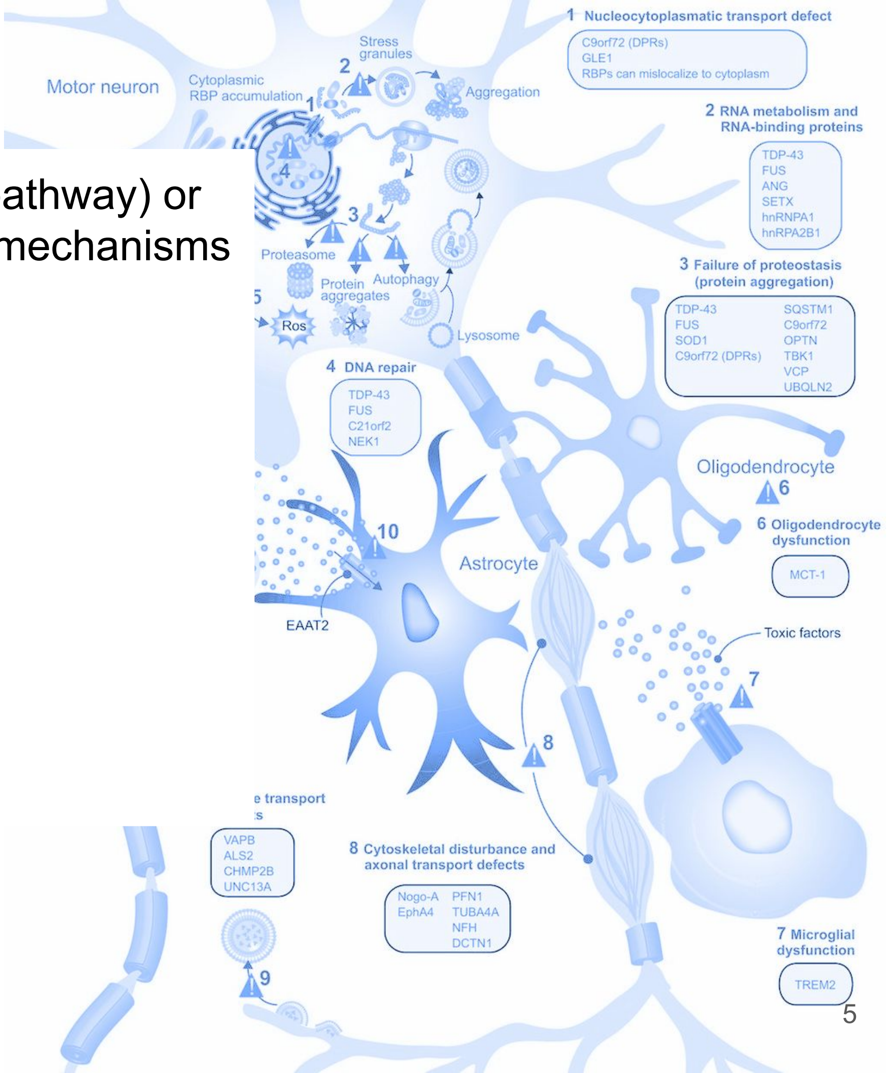
Does ALS have one mechanism of action (one pathway) or is it caused by multiple independent or different mechanisms of action (multiple pathways)?

➔ **Multiple pathways**

However:

Do we find evidence for new implicated genes?

…. and new pathways involved?

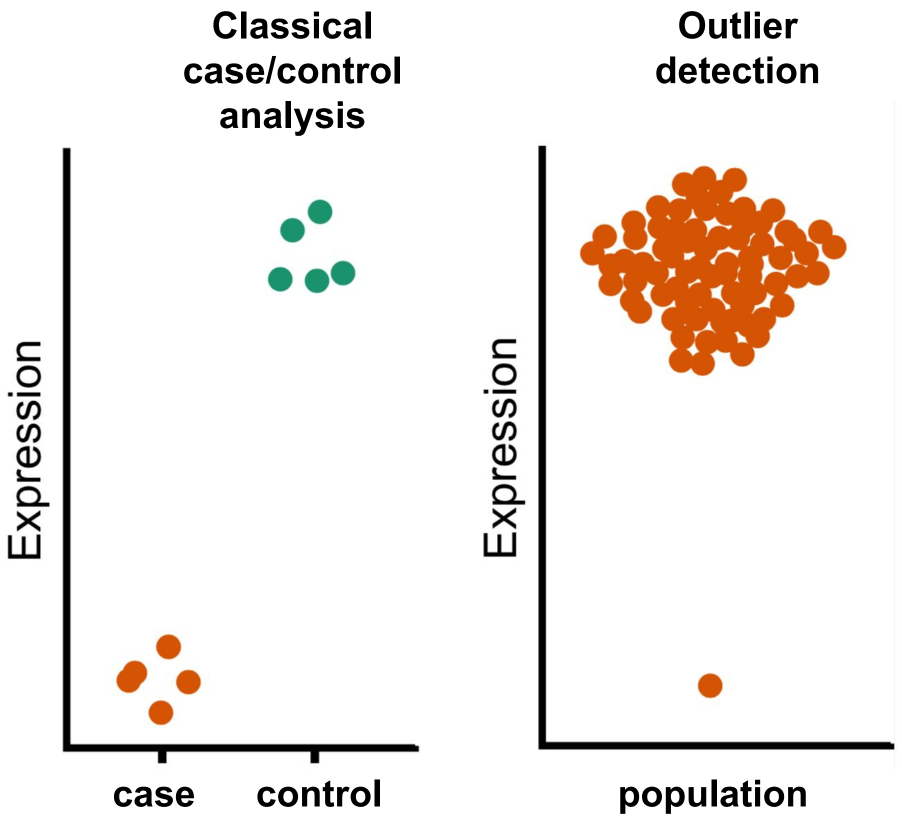


Outlier detection for gene discovery

- No common pattern of gene expression among patients
- Many pathways involved

➔ To search for new genes we instead ask: **What makes every patient unique?**

➔ This leads us to focus on **expression outlier detection**, instead of classical differential case/control expression analysis.

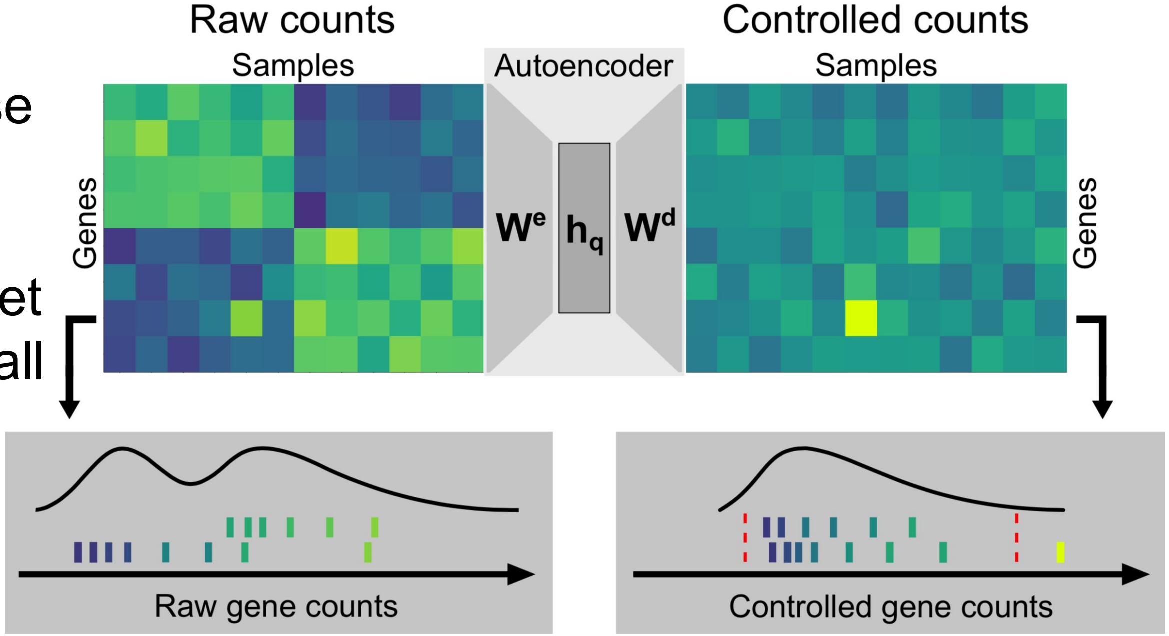


Classical case/control analysis

Outlier detection

A denoising autoencoder with a negative binomial loss to control for latent factors in RNA-Seq data

- Use a denoising autoencoder to automatically remove noise and confounding factors
- Negative Binomial loss
- Number of latent factors set to maximise precision-recall of artificial outliers

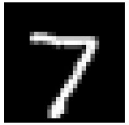


OUTRIDER, Brechtmann et al. AJHG (2018)

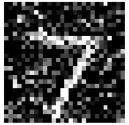
Selecting optimal latent representation

As with image processing denoising autoencoders (AEs), OUTRIDER is optimized to remove artificially injected noise

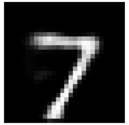
$$\mathbf{X}$$

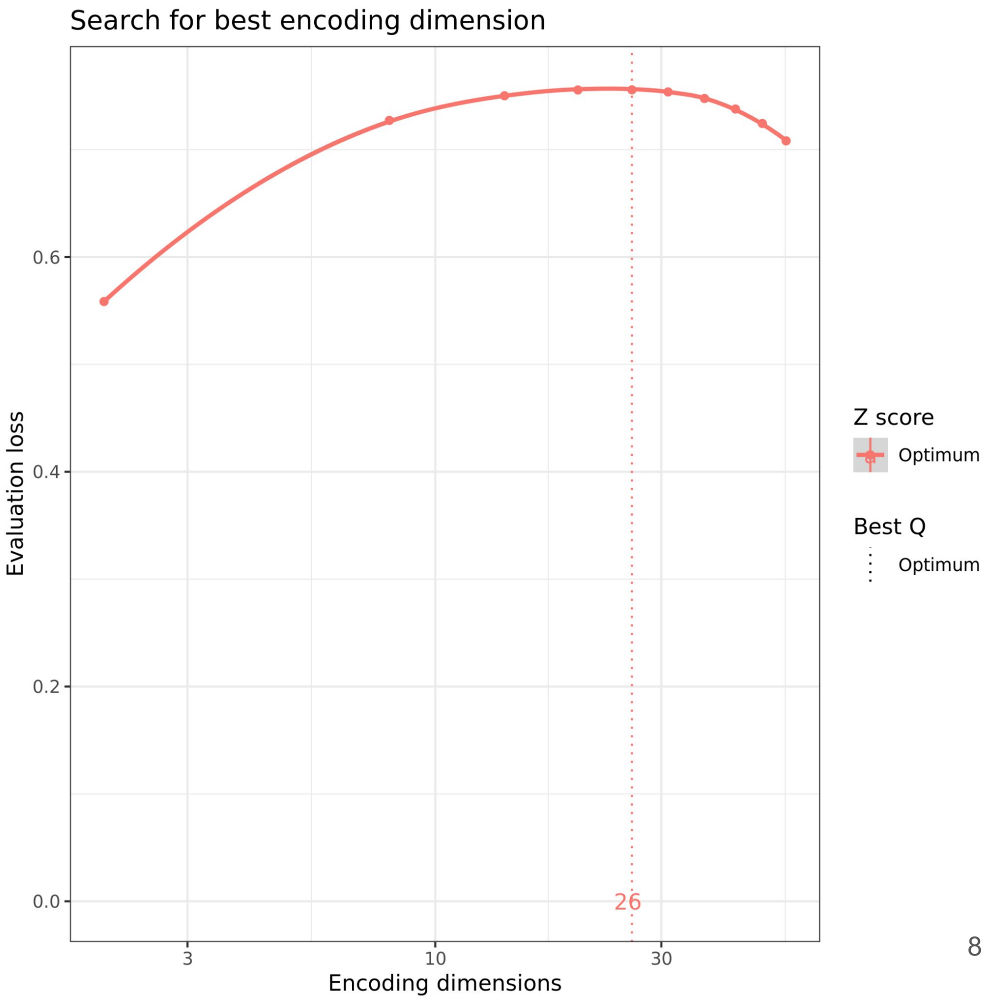


$$\mathbf{X}^{\mathbf{corrupt.}}$$



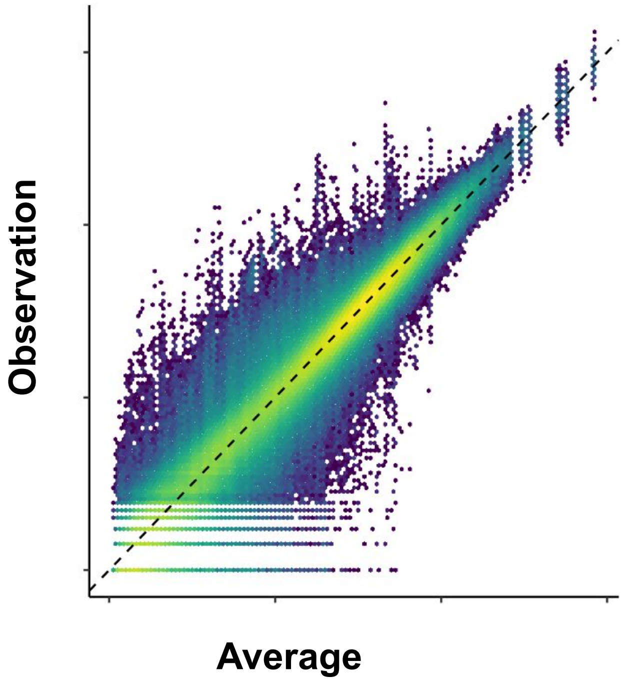
$$f(\mathbf{X}^{\mathbf{corrupt.}}, \boldsymbol{\theta})$$



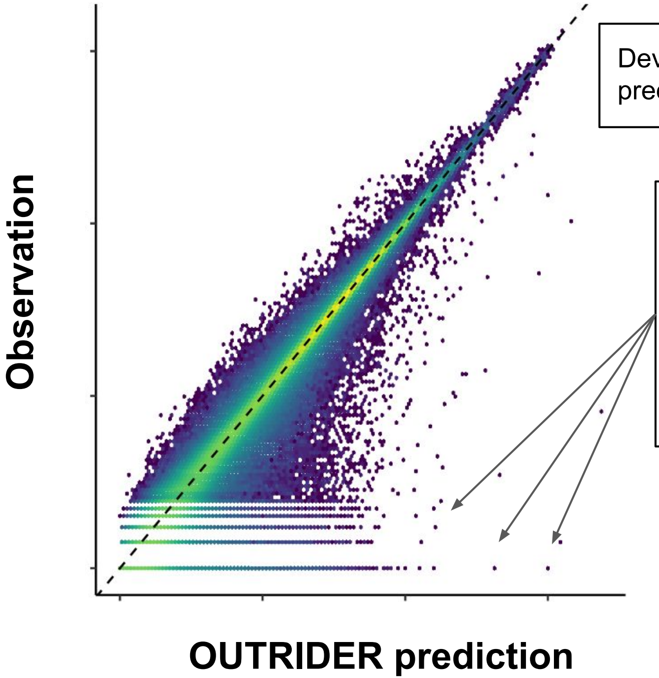


OUTRIDER accurately predicts expression of each gene per sample and reveals outliers

Before OUTRIDER



After OUTRIDER



Deviations from model predictions are outliers.

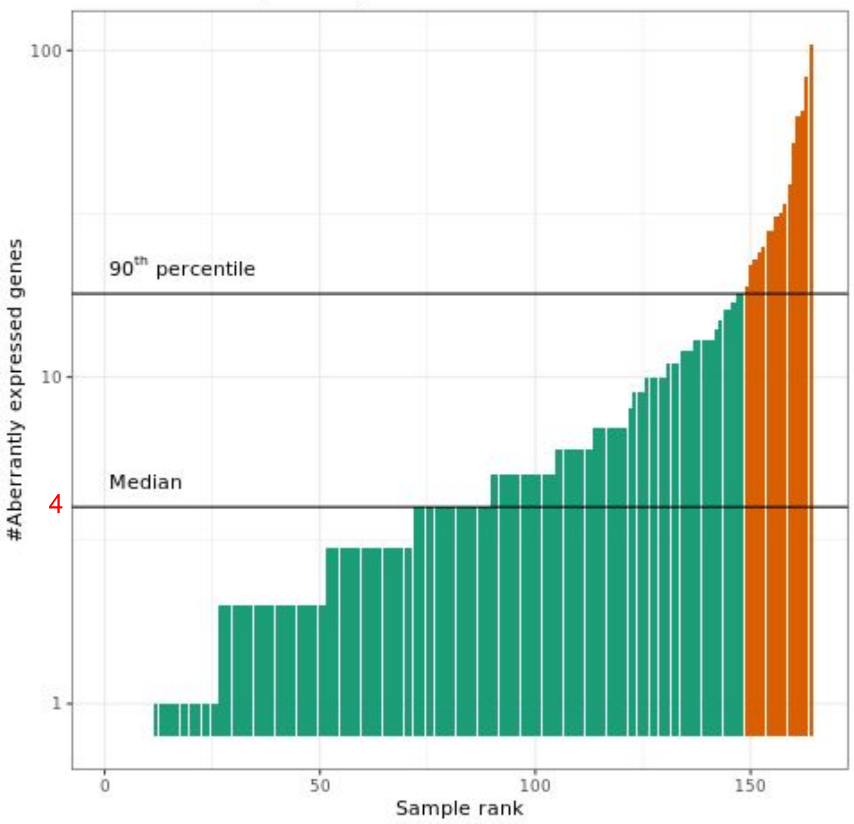
In these individuals, those genes are at levels far below what one would expect given the expression of all other genes….

🤔 Are rare genetic variants disrupting those genes in these individuals?

Scale: $\log_{10}$(count +1)

Individuals typically have 4 outlier genes

We observe a small number of aberrantly expressed genes per sample.



Samples ranked by number of aberrantly expressed genes

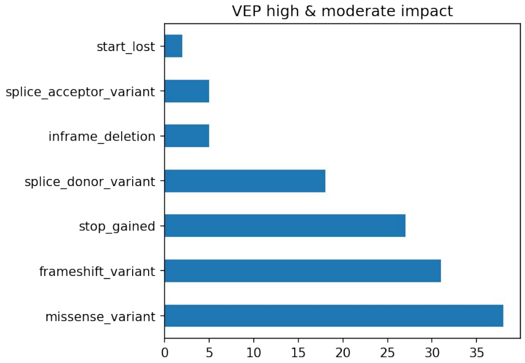
Genetically supported gene expression outliers

Filtering outliers for having **rare** and **deleterious** genetic variants (from the same sample with DNA sequencing), impacting coding or splicing.

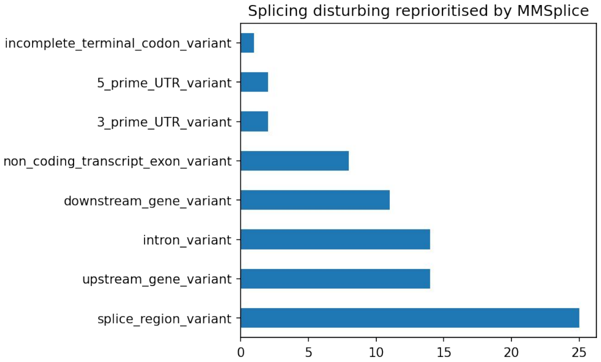
Frequency in general population < 0.1% (gnomAD[5])

At most 6 samples in the cohort

Standard VEP high and moderate impact annotations


VEP high & moderate impact

Deep-learning based aberrant splicing predictions[4]

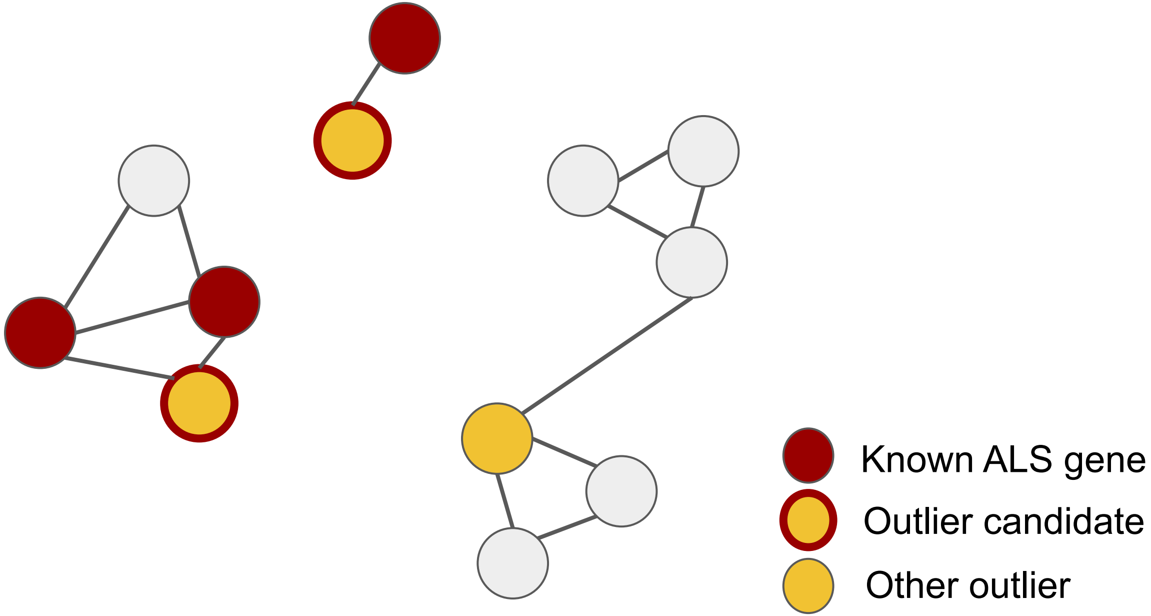

Splicing disturbing reprioritised by MMSplice

[4] MMSplice, Cheng et al. Genome Biology (2019)
[5] Karczewski et al. Nature (2020)

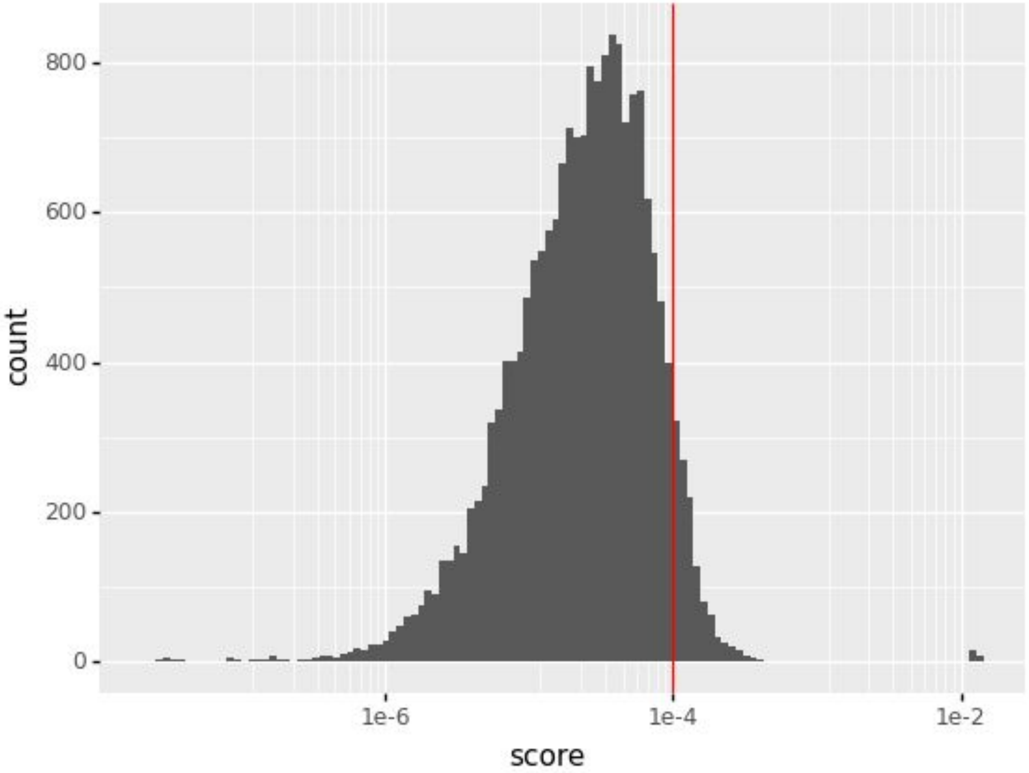
Gene network analysis to study the relationship between newly discovered and known ALS genes

level 1: Outliers in the network vicinity of ALS genes as new candidates

STRING https://string-db.org/
was used as a gene network.



Modeling network vicinity with random walks on gene networks



Vicinity to ALS genes modeled as the probability of visiting the gene by random walks starting from an ALS gene.

Genes with a prob. larger than 10^{-4} (**PPI score**) were considered interesting (right tail after red line).

Results

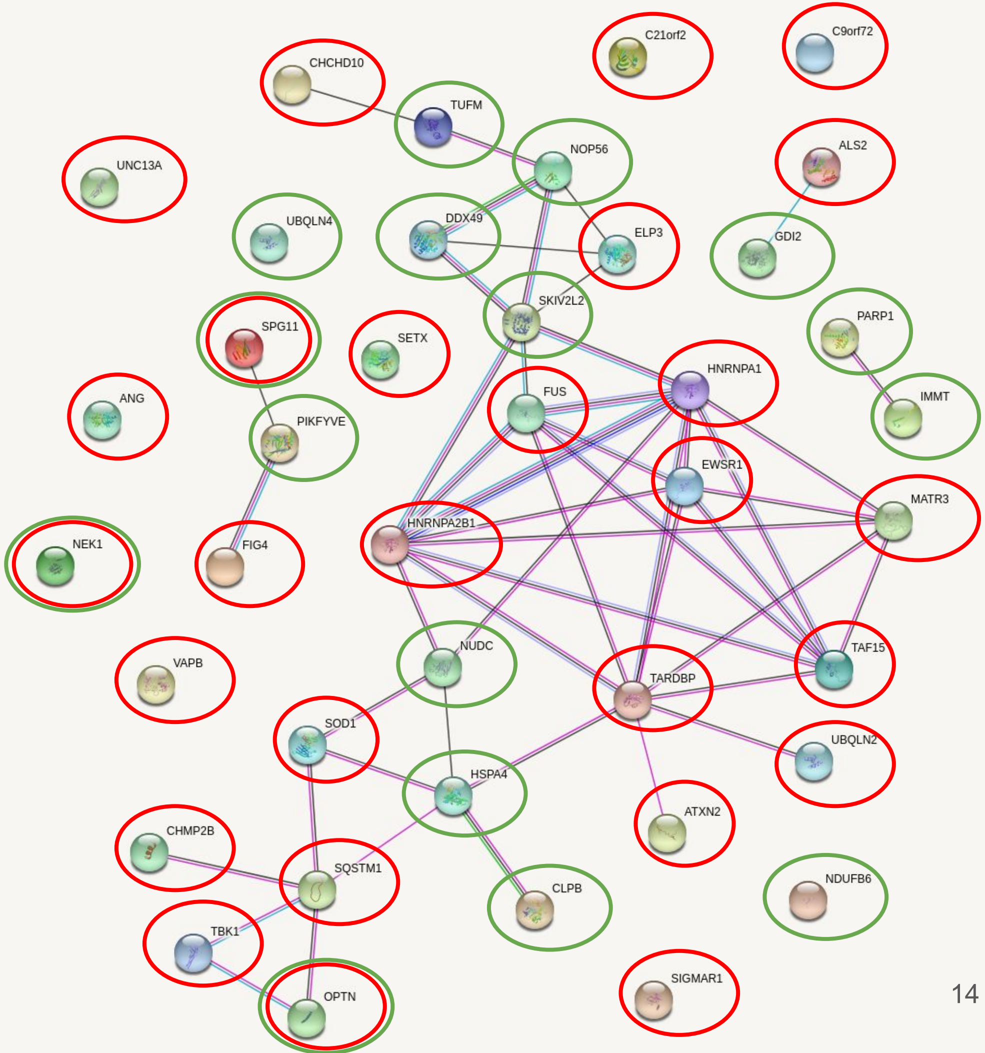
Network of known ALS genes and expression **outlier genes** containing a rare deleterious **variant** and a high **PPI score**.

Known ALS gene

expression outlier

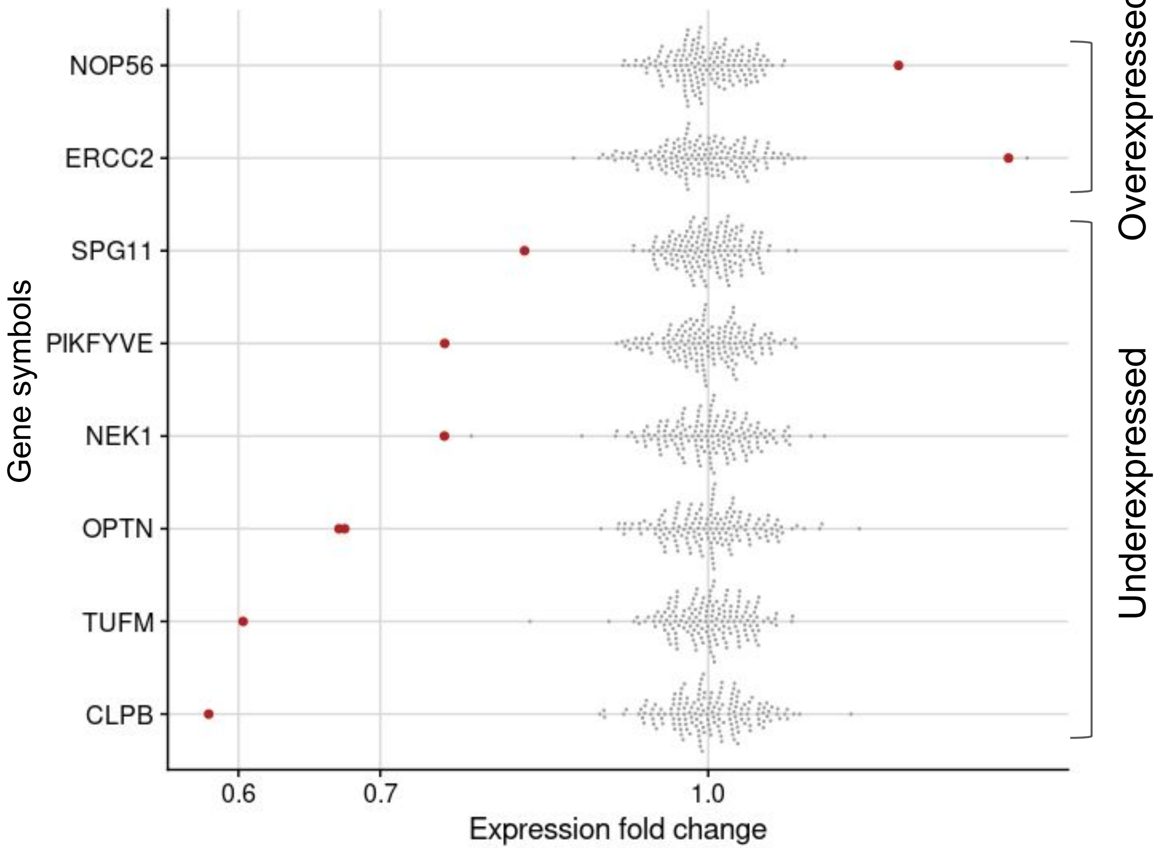
1. We found **16** expression outliers interacting with known ALS genes.

2. Some of them e.g. PIKFYVE connect known ALS genes



Identification of known genes and new interesting candidates

- Gene expression per sample
- Effect size: expression log2 fold change
- Outliers marked in red



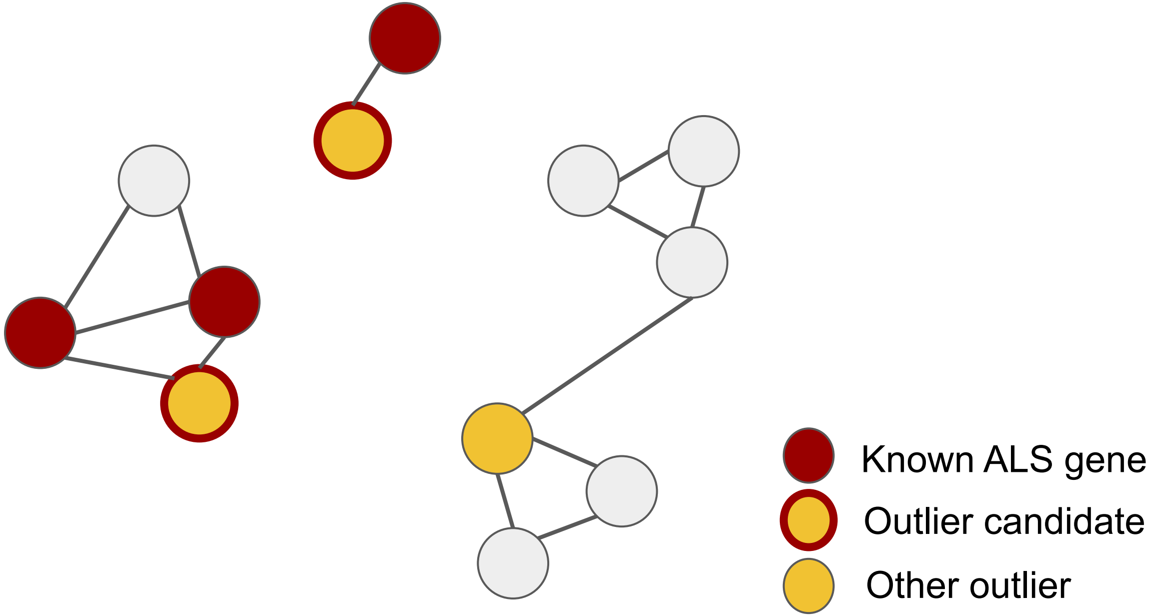
Identification of known genes and new interesting candidates

List of aberrantly expressed genes (outliers) containing at least one rare high impact variant.
The genes are either **known to cause ALS** (according to ALSoD) or **associated to other relevant diseases**. All genes are close to the established ALS genes in the gene network.

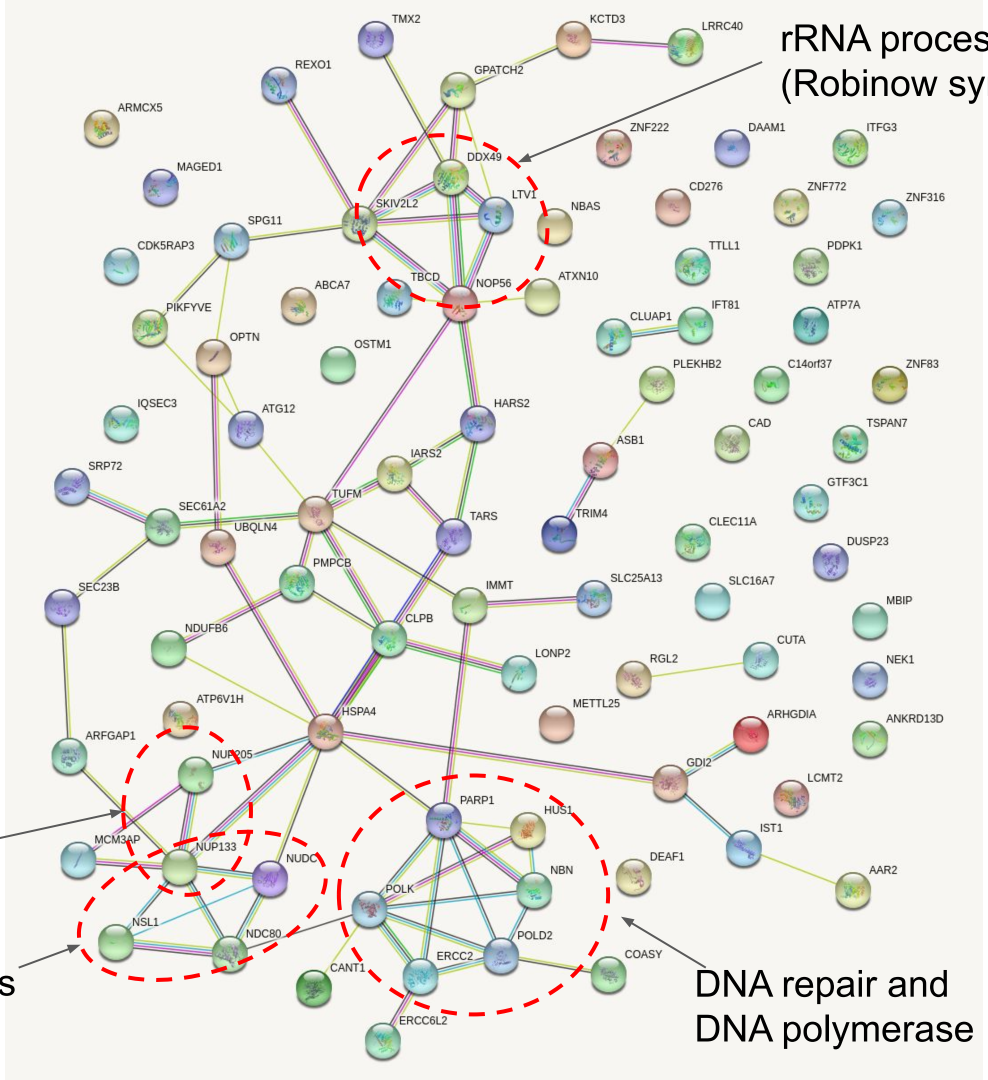
Sample	Gene	fold change	PPI score	Variant	Consequence	ClinVar	Comment
CASE.NEUEK191WYC	*NEK1*	0.75	1.22E-02	chr4:169424645:G>A	stop		Definitive ALS gene
CASE.NEUBK117YXL	*OPTN*	0.67	1.23E-02	chr10:13122390:C>A	stop		Definitive ALS gene
CASE.NEUZT557DHF	*OPTN*	0.67	1.23E-02	chr10:13112464:T>TAG	frameshift		Definitive ALS gene
CASE.NEUVX902YNL	*SPG11*	0.82	1.23E-02	chr15:44620189:C>A	splice donor	likely pathogenic	Tenuous ALS gene, variant predicted to cause aberrant splicing
CASE.NEULD354RZB	*NOP56*	1.23	1.61E-04	chr20:2655751:G>A	splice region		Variant predicted to cause aberrant splicing. Gene related to Ataxia.
CASE.NEUTA689LN5	*TUFM*	0.60	1.06E-04	chr16:28844814:G>A	stop	uncertain significance	Mitochondrial disease gene
CASE.NEUGW326BRV	*CLPB*	0.58	1.30E-04	chr11:72302312:G>A	stop	pathogenic	Mitochondrial disease gene
CASE.NEUME498PCJ	*PIKFYVE*	0.75	1.54E-04	chr2:208352730:A>AT	frameshift		Linked to neurodegeneration
CASE.NEURR881FKY	*ERCC2*	1.38	5.86E-05	chr19:45364832:CCTCA>C	splice donor	likely pathogenic	Causes neurological symptoms, e.g. spasticity and reflex abnormalities, and skin manifestations

Gene network analysis - level 2:
Identify clusters of outliers as new candidate pathways

STRING https://string-db.org/
was used as a gene network.



Further outliers with rare deleterious variants interact with each other indicating new implicated pathways



rRNA processing
(Robinow syndrome)

Nucleopores

Kinetochores

DNA repair and
DNA polymerase

Discussion / outlook

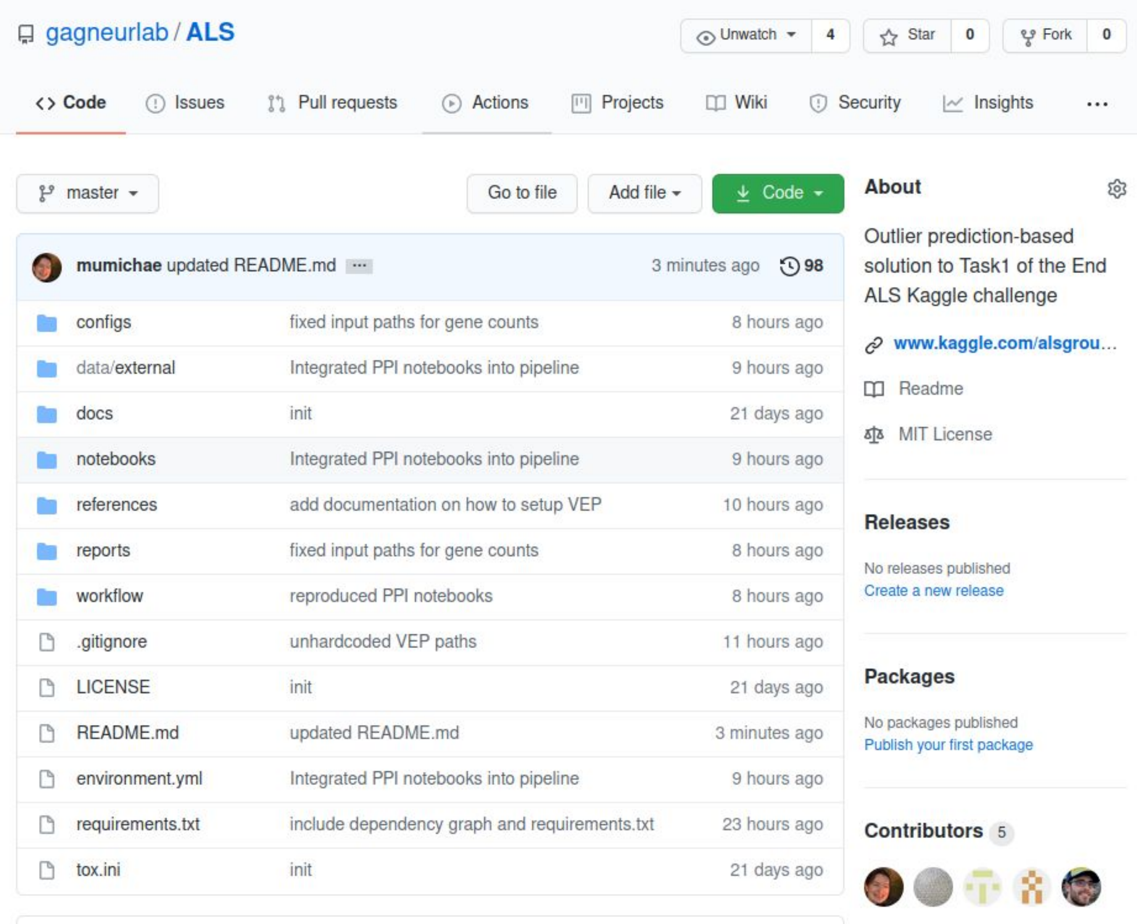
- The outlier analysis provides a new perspective, we believe more informative approach, for studying ALS.
- These new candidate genes could expand the understanding of pathways involved in the etiology of ALS.
- Future analysis would include:
    - Replicating the findings looking at WGS of the entire ALS dataset (other patients with damaging variants in the same genes).
    - Multi-omics outlier analysis: ATAC-seq, splicing, proteomics, to investigate the impact of gene regulatory control, splicing control and protein expression on ALS.
    - Integrative analysis of multi-omics data to obtain a holistic view.
    - Functional follow-ups and collaborations with experimental groups and experts in this area.

Conclusion

- We found variants associated with aberrant expression for known ALS genes, potentially characterising those affected patients (n = 4).
- We found new high impact variants in further cases in a gene potentially related to ALS, which would improve our catalogue of pathogenic variants.
- We found new candidate genes in known pathways.
- We found potential new pathways.
- Altogether, this gives a potential genetic explanation to **63 (46%)** of the patients and further supports a multi-causal and multi-mechanism view of ALS.

Code to reproduce the results

https://github.com/gagneurlab/ALS-kaggle



mumichae updated README.md ··· 3 minutes ago 98

configs	fixed input paths for gene counts	8 hours ago
data/external	Integrated PPI notebooks into pipeline	9 hours ago
docs	init	21 days ago
notebooks	Integrated PPI notebooks into pipeline	9 hours ago
references	add documentation on how to setup VEP	10 hours ago
reports	fixed input paths for gene counts	8 hours ago
workflow	reproduced PPI notebooks	8 hours ago
.gitignore	unhardcoded VEP paths	11 hours ago
LICENSE	init	21 days ago
README.md	updated README.md	3 minutes ago
environment.yml	Integrated PPI notebooks into pipeline	9 hours ago
requirements.txt	include dependency graph and requirements.txt	23 hours ago
tox.ini	init	21 days ago

About

Outlier prediction-based solution to Task1 of the End ALS Kaggle challenge

www.kaggle.com/alsgrou...

Readme

MIT License

Releases

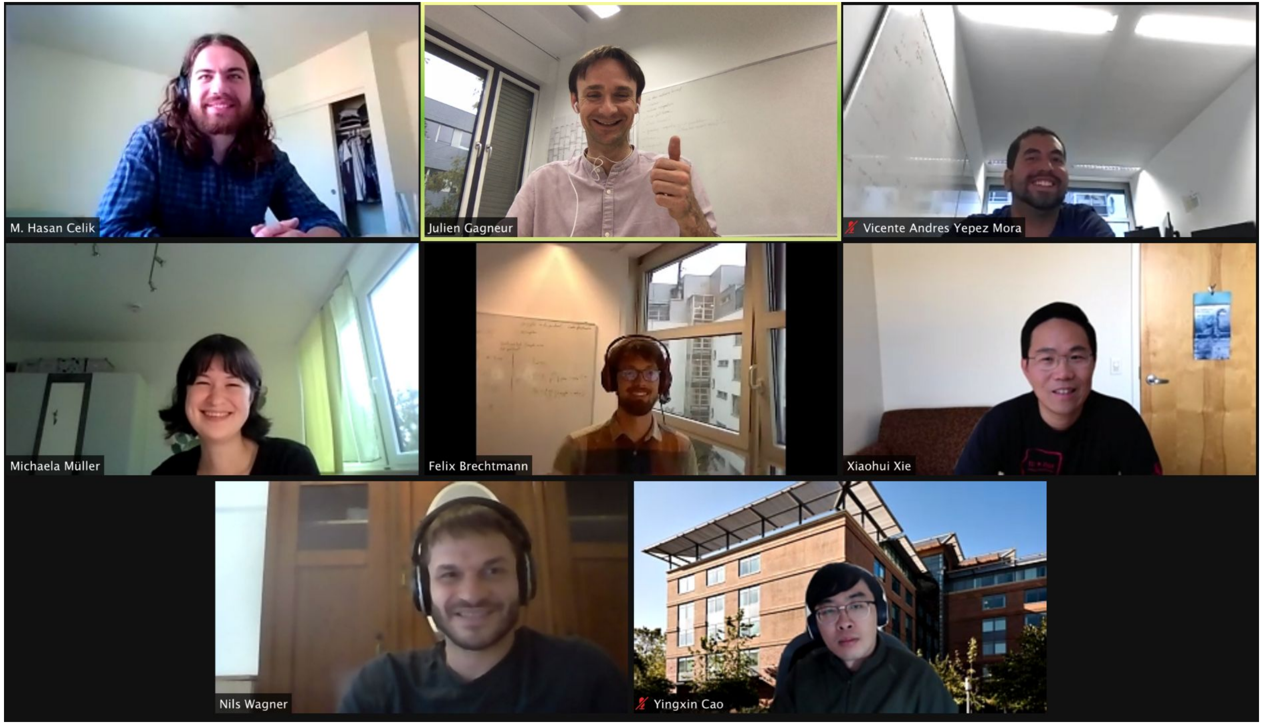
No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 5

The team



Felix Brechtmann[1], M. Hasan Çelik[2], Julien Gagneur[1], Florian Hölzlwimmer[1], Michaela Müller[1], Nils Wagner[1], Xiaohui Xie[2], Vicente Yépez[1], Michael Zech[1]

[1] Technical University of Munich [2]University of California, Irvine

Acknowledgement

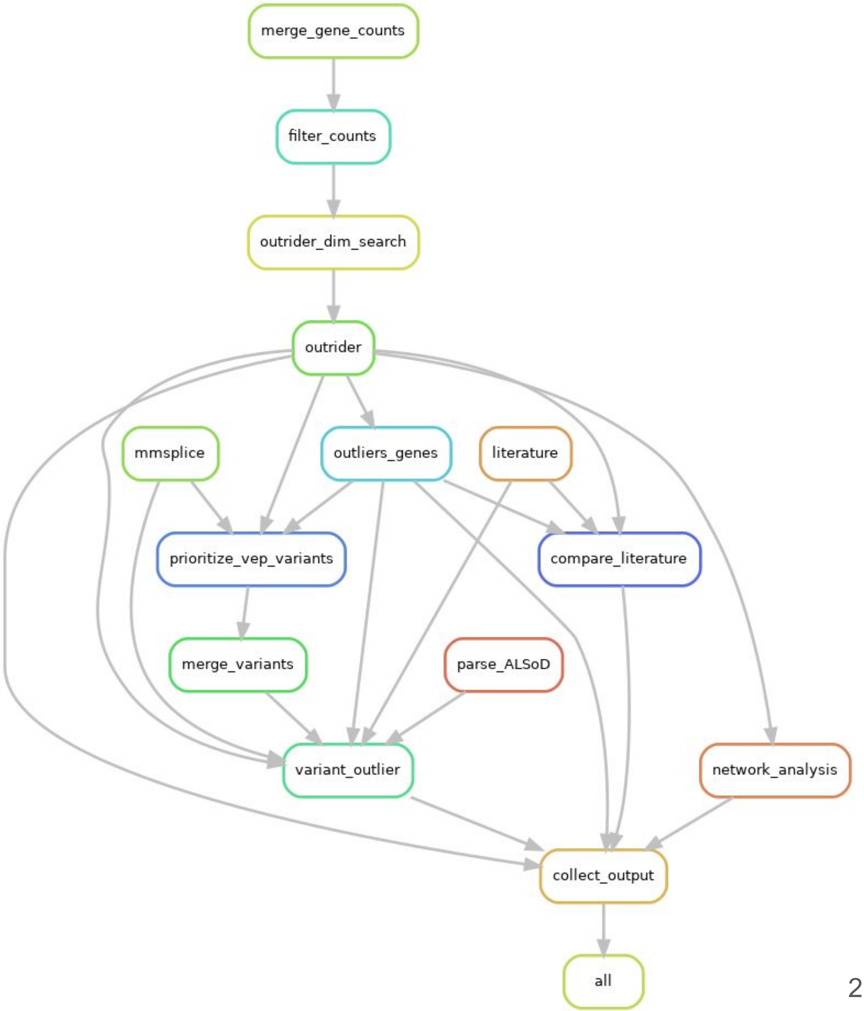
We thank

- Organizers of End ALS Kaggle Challenge for making the datasets available.
- Leslie Thompson and her group for helpful discussions.

Appendix

Analysis Workflow

- Reproducible pipeline in Snakemake
- Parallelized and robust
- Main steps:
  - Prepare gene counts
  - OUTRIDER analysis
  - Variant annotation
  - PPI network analysis
  - UMAP on expression space (not shown here)


merge_gene_counts
filter_counts
outrider_dim_search
outrider
mmsplice
outliers_genes
literature
prioritize_vep_variants
compare_literature
merge_variants
parse_ALSoD
variant_outlier
network_analysis
collect_output
all