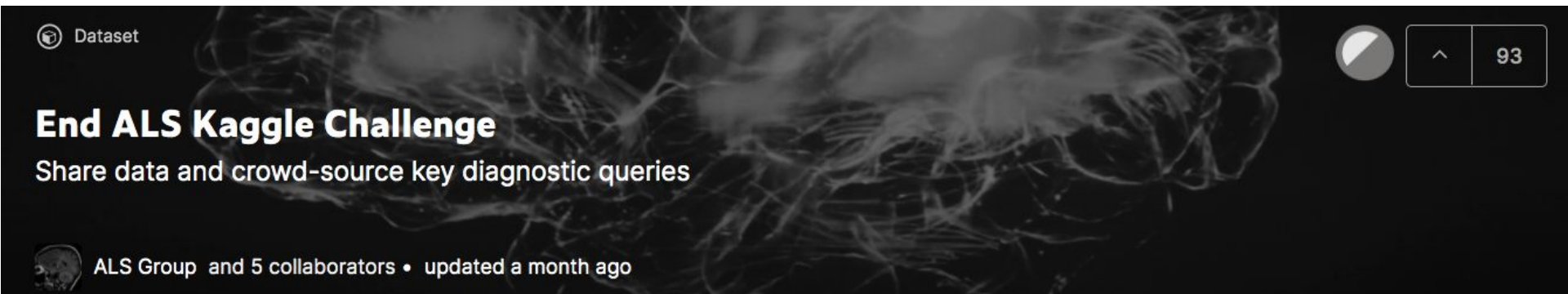


# When the outlier is the signal



Searching new genetic causes of ALS  
by aberrant expression analysis

15<sup>th</sup> May 2021

TUM - UCI team

# Task 1

Does ALS have one mechanism of action?

→ one pathway

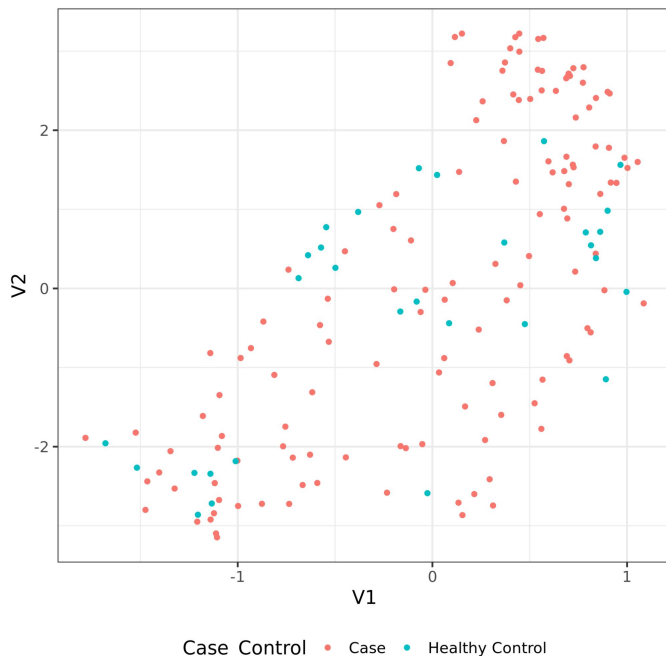
Or is it caused by multiple independent or different mechanisms of action?

→ multiple pathways

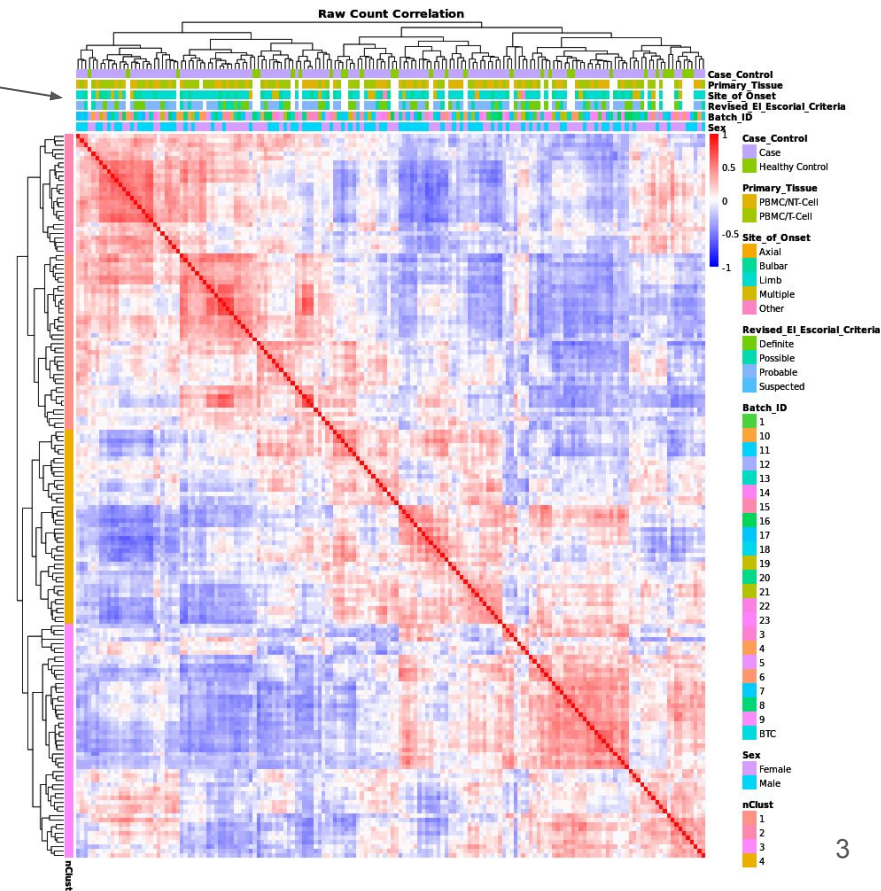
# Gene expression does not naturally cluster donor groups

## Cases and controls

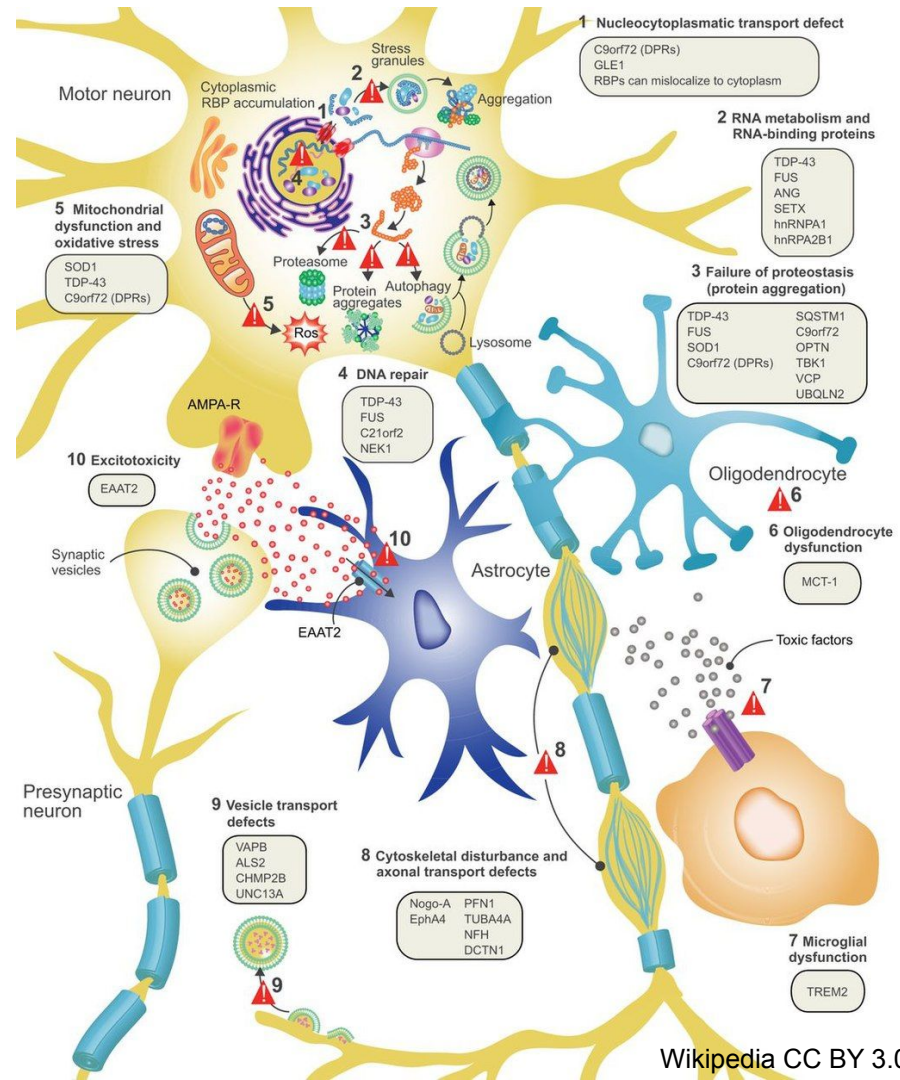
visualized in first two principal components (V1 and V2)



No covariate  
drives the  
sample clusters



**In fact, mutations in ca. 40 genes have been implicated in ALS over various molecular pathways<sup>1</sup>**



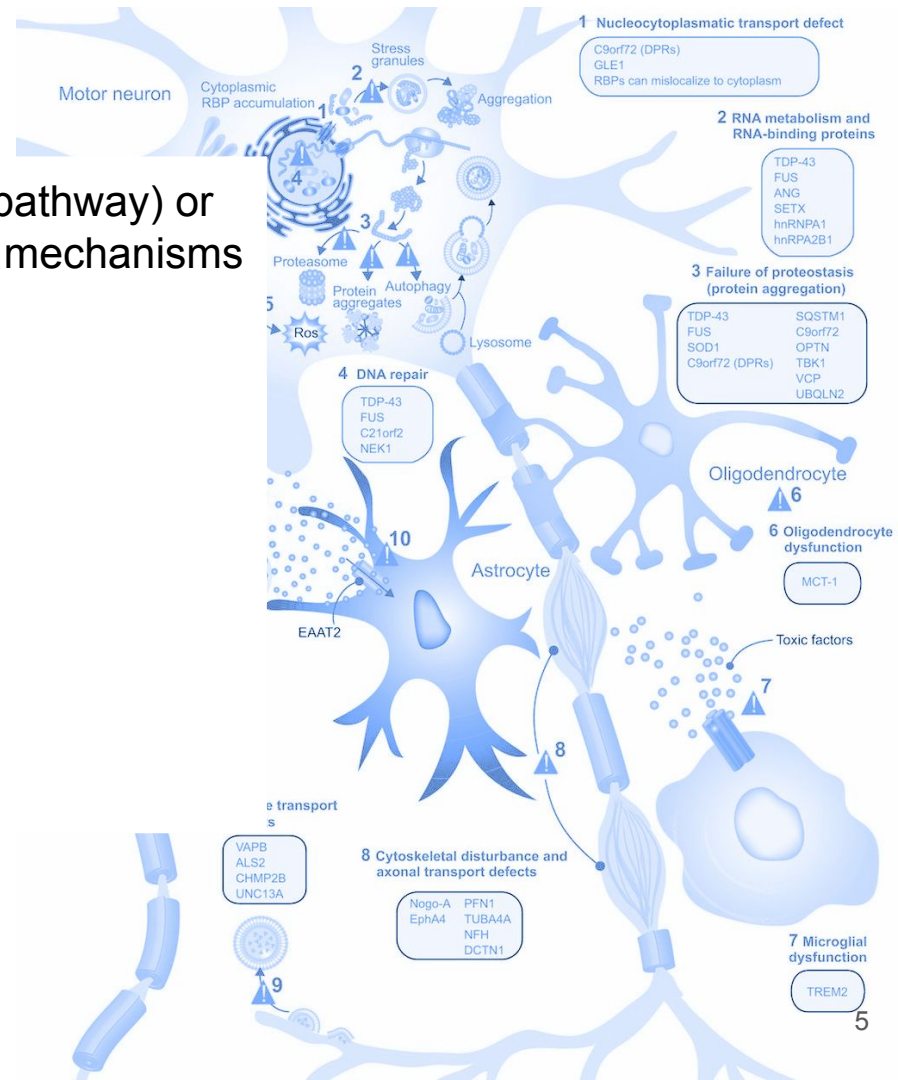
1. Gregory et al. Curr Genet Med Rep (2020)
2. Hardiman O et al. Nat. Rev. Dis. Primers (2017)
3. van Damme et al. Disease Models and Mechanisms (2017)

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**

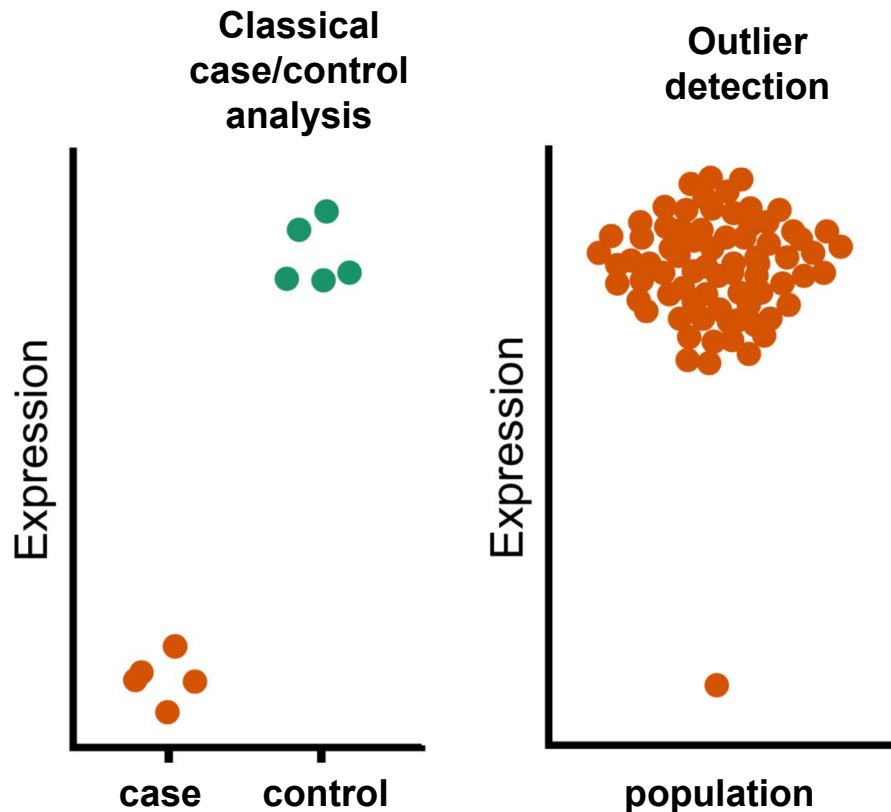
However:

Do we find evidence for new implicated genes?  
.... and new pathways involved?



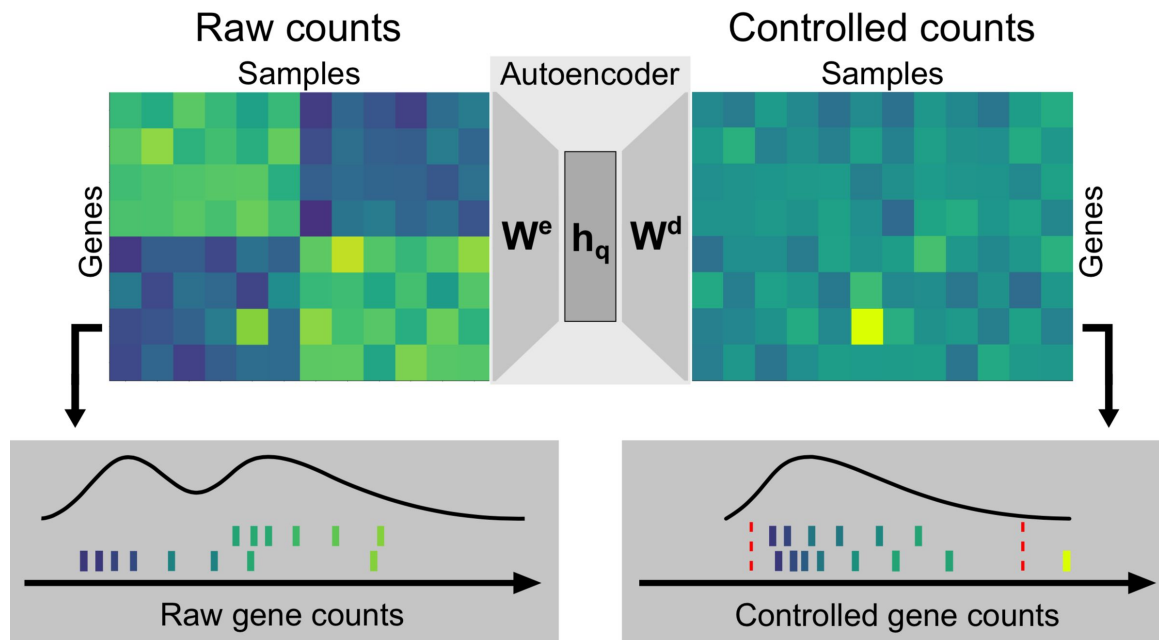
# Outlier detection for gene discovery

- No common pattern of gene expression among patients
  - Many pathways involved
- To search for new genes we instead ask:  
What makes every patient unique?
- Expression outlier detection



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

- Negative Binomial loss
- Number of latent factors set to maximise precision-recall of artificial outliers





# Hyperparameter optimization

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

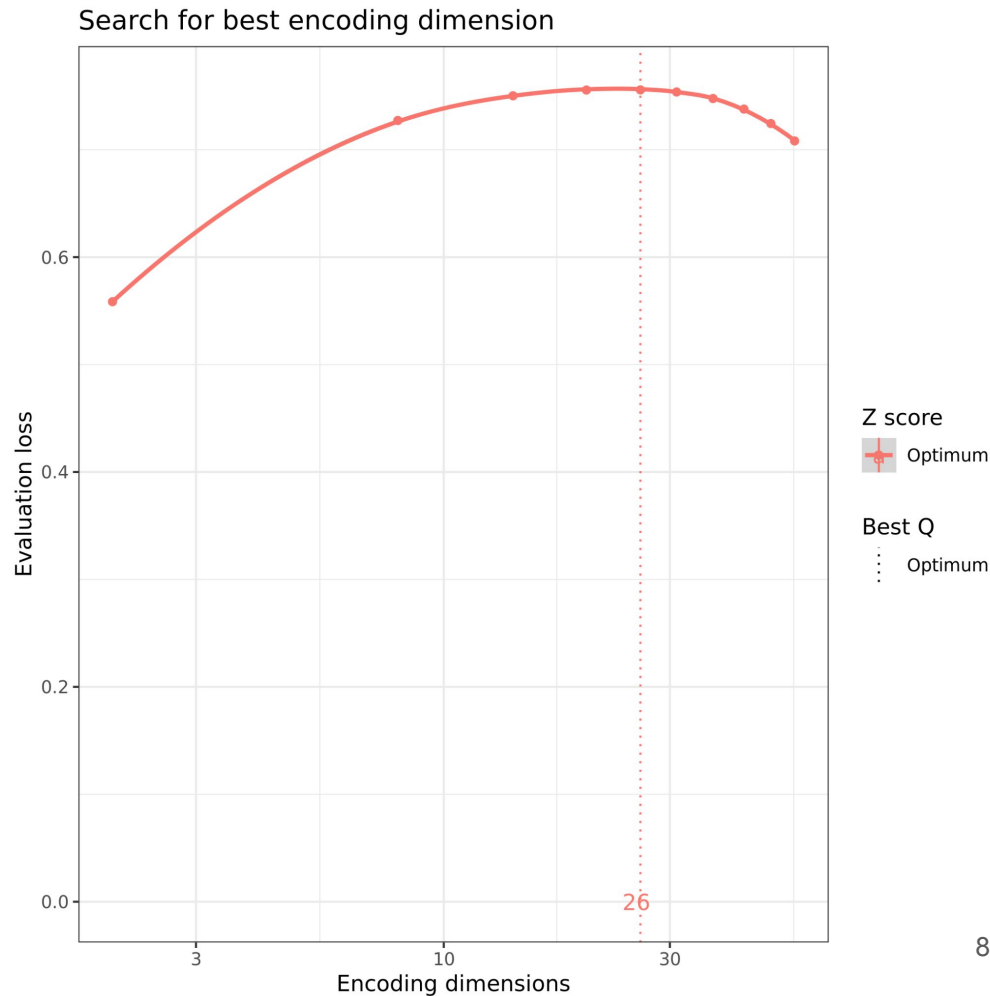
$\mathbf{X}$



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



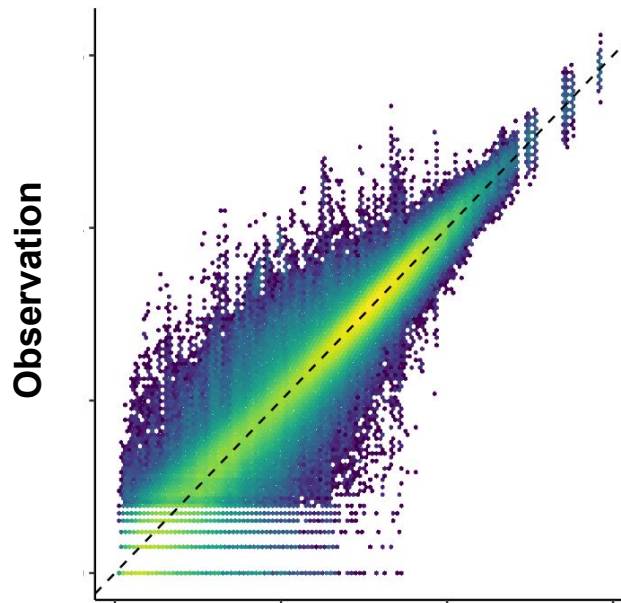
$f(\mathbf{X}^{\text{corrupt.}}, \theta)$





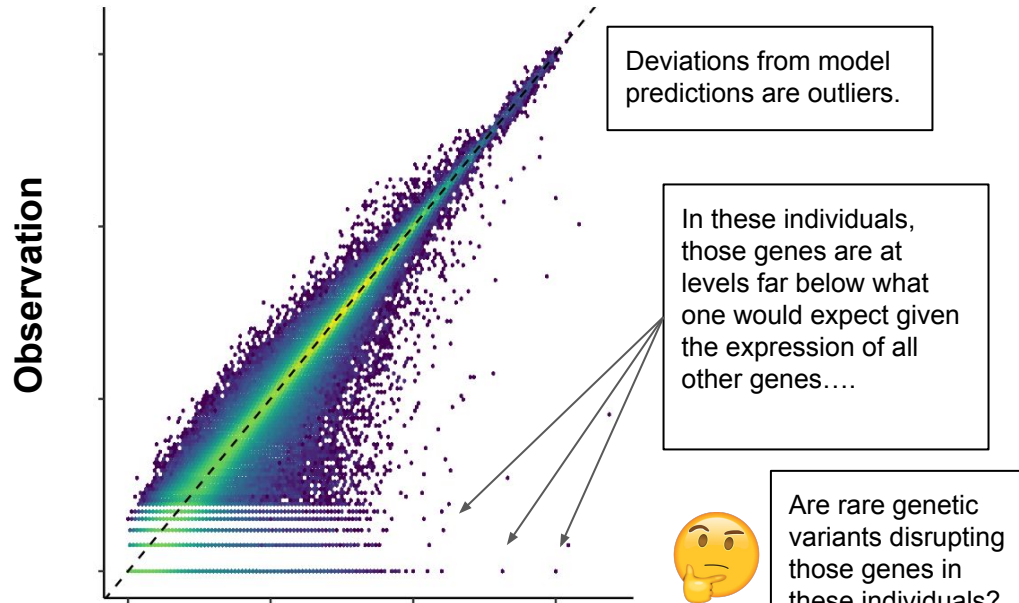
# OUTRIDER accurately predicts expression of each gene per sample and reveals outliers

Before OUTRIDER



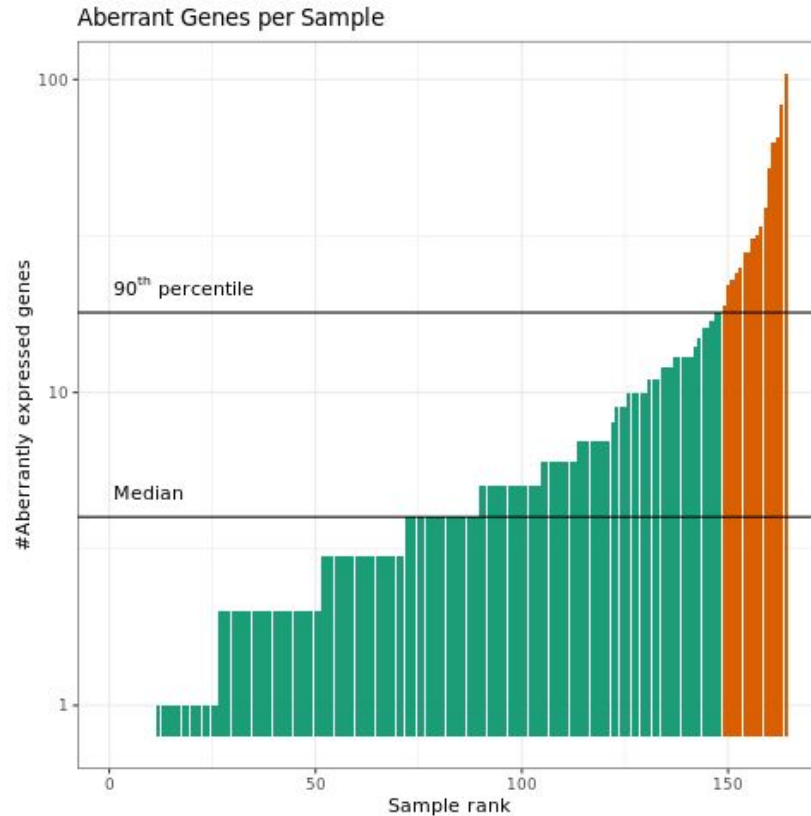
Average

After OUTRIDER



OUTRIDER prediction

# Individuals typically have 4 outliers



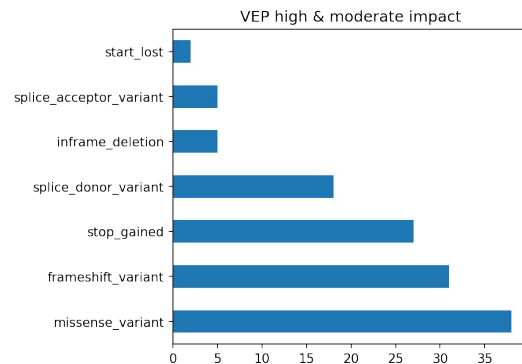
# Genetically supported outliers

Filtering outliers for having **rare** and **deleterious** genetic variants

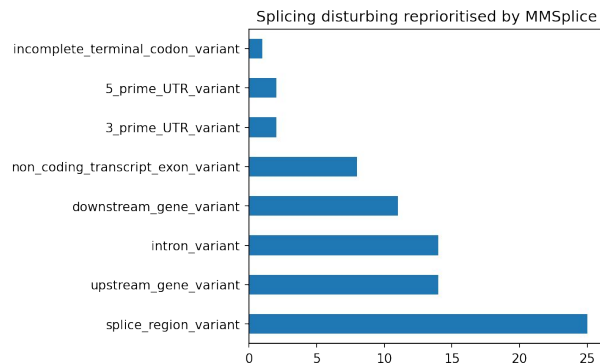
Frequency in general  
population < 0.1%  
(gnomAD<sup>5</sup>)

At most 6 samples in the  
cohort

Standard VEP high and  
moderate impact  
annotations



Deep-learning based  
aberrant splicing  
predictions<sup>4</sup>



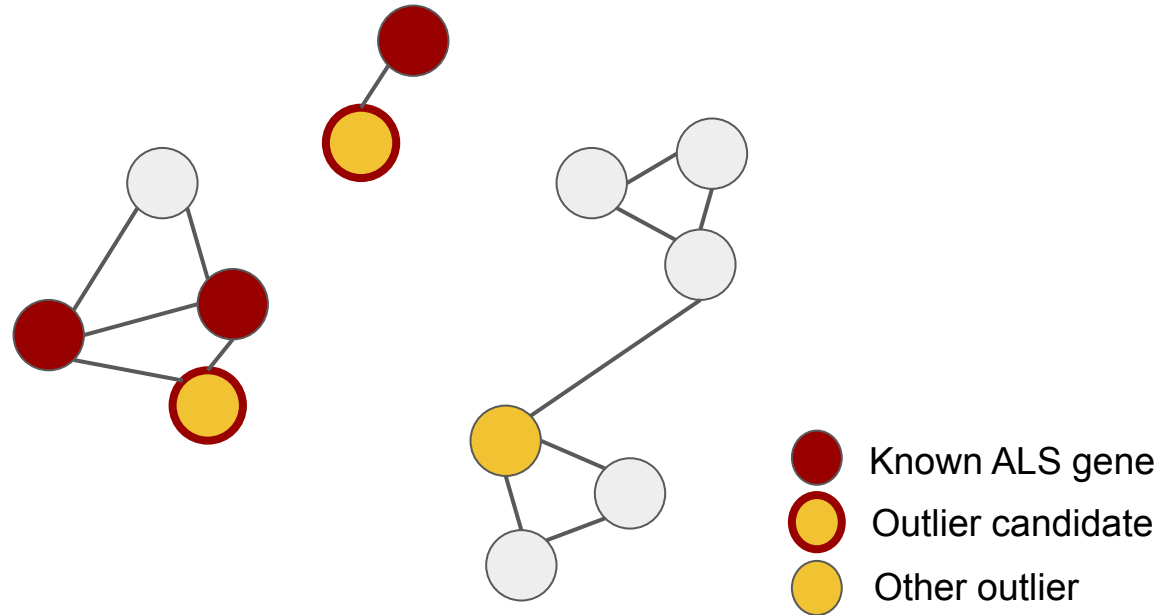
4. MMSplice, Cheng et al. Genome Biology (2019)

5. Karczewski et al. Nature (2020)

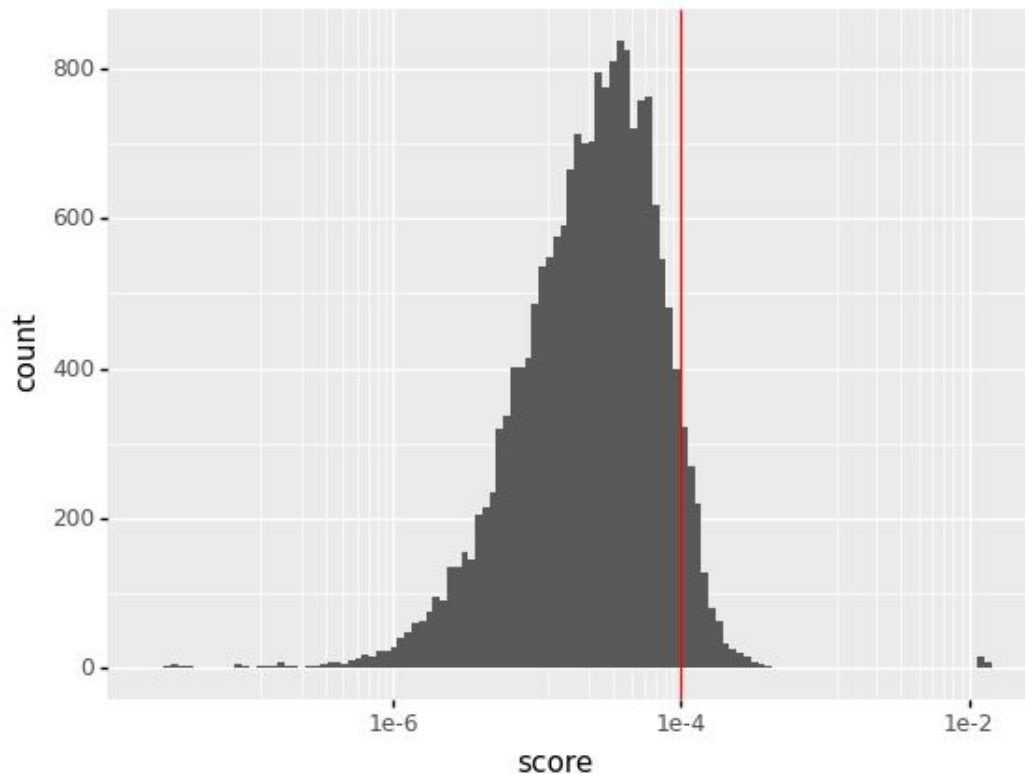
# Gene network analysis - level 1:

## 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





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

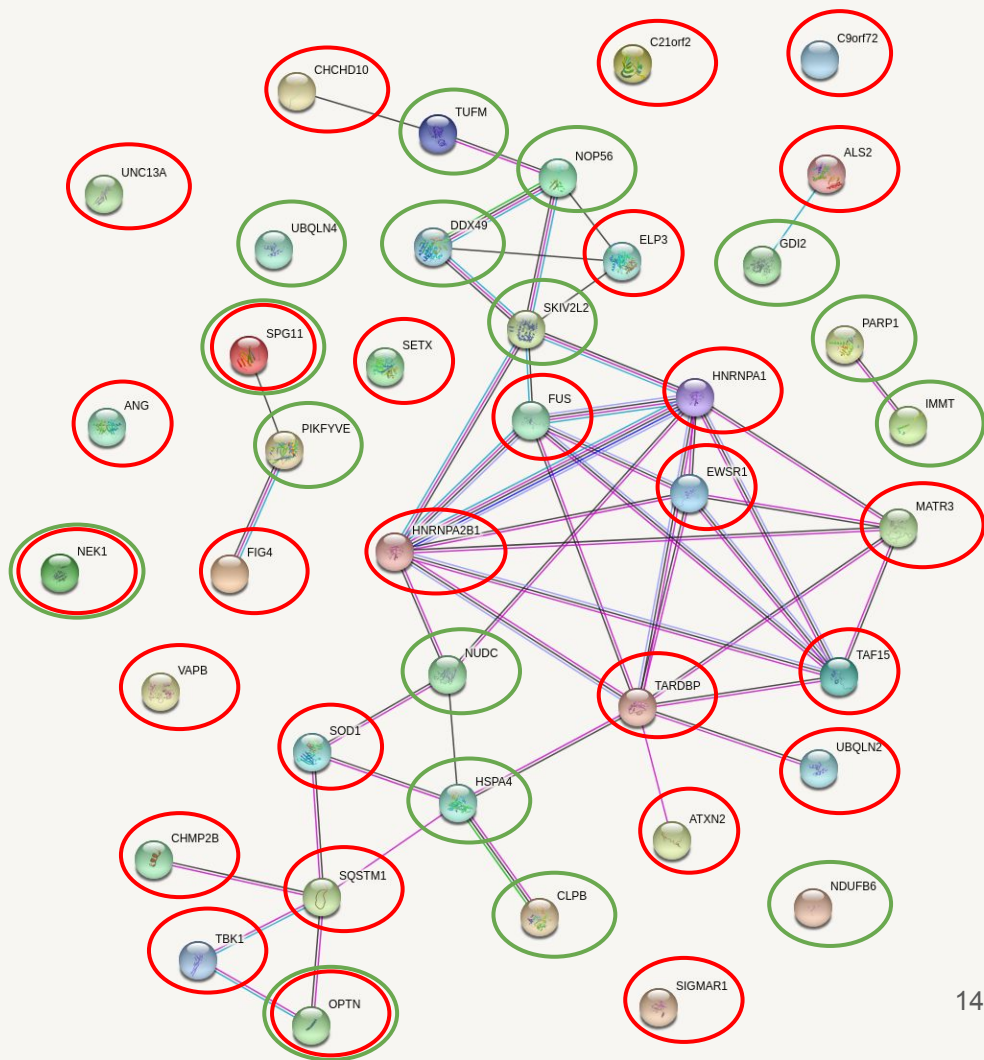
Genes with a prob. Larger than  $10^{-4}$  were considered interesting.

# Results

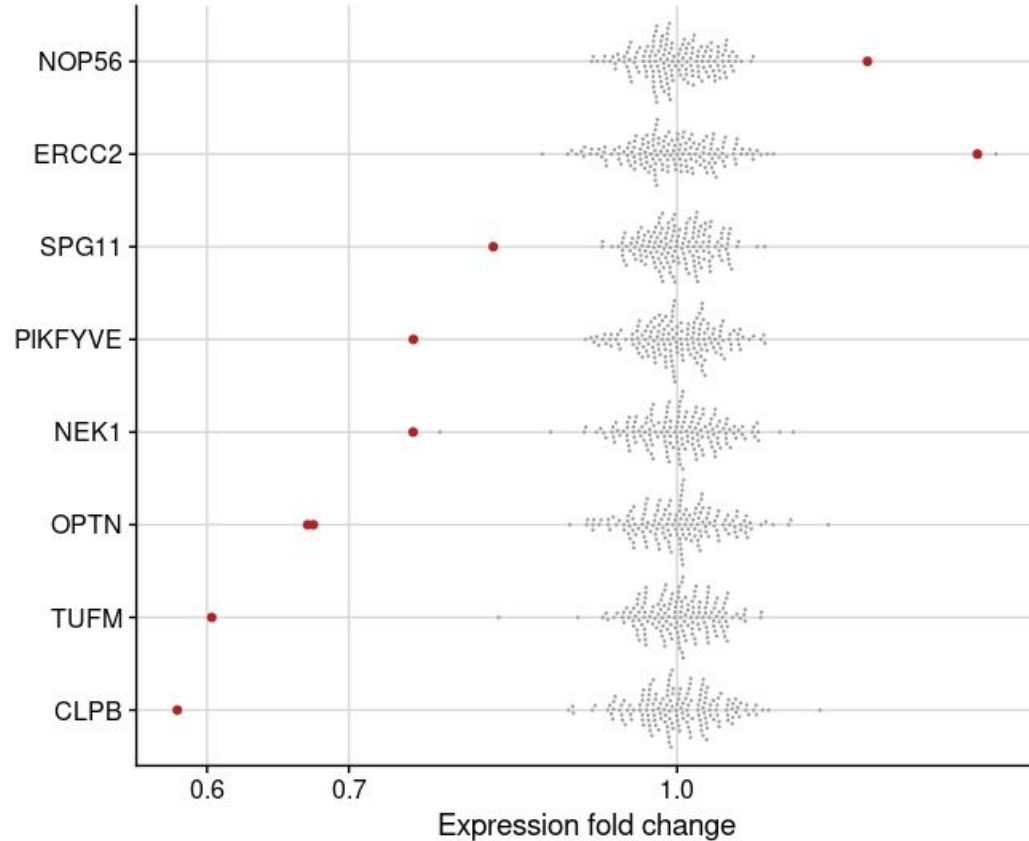
Network of known ALS genes and expression **outlier genes** containing a rare deleterious **variant** and a high **PPI score**.

-  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





# Identification of known genes and new interesting candidates

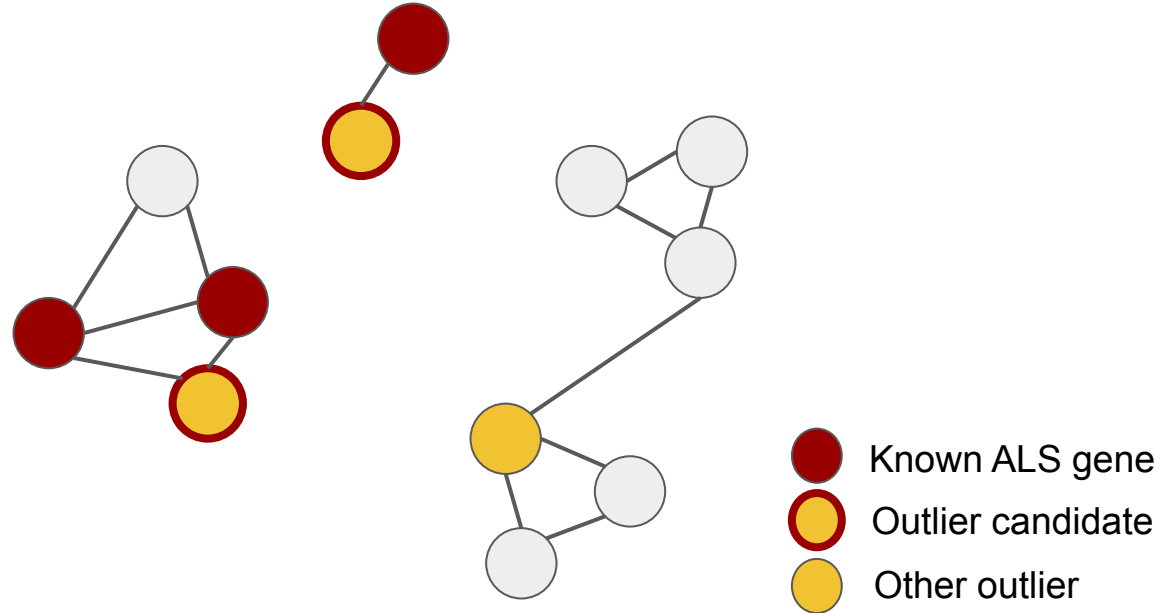
Aberrantly expressed genes containing a rare high impact variant.

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.

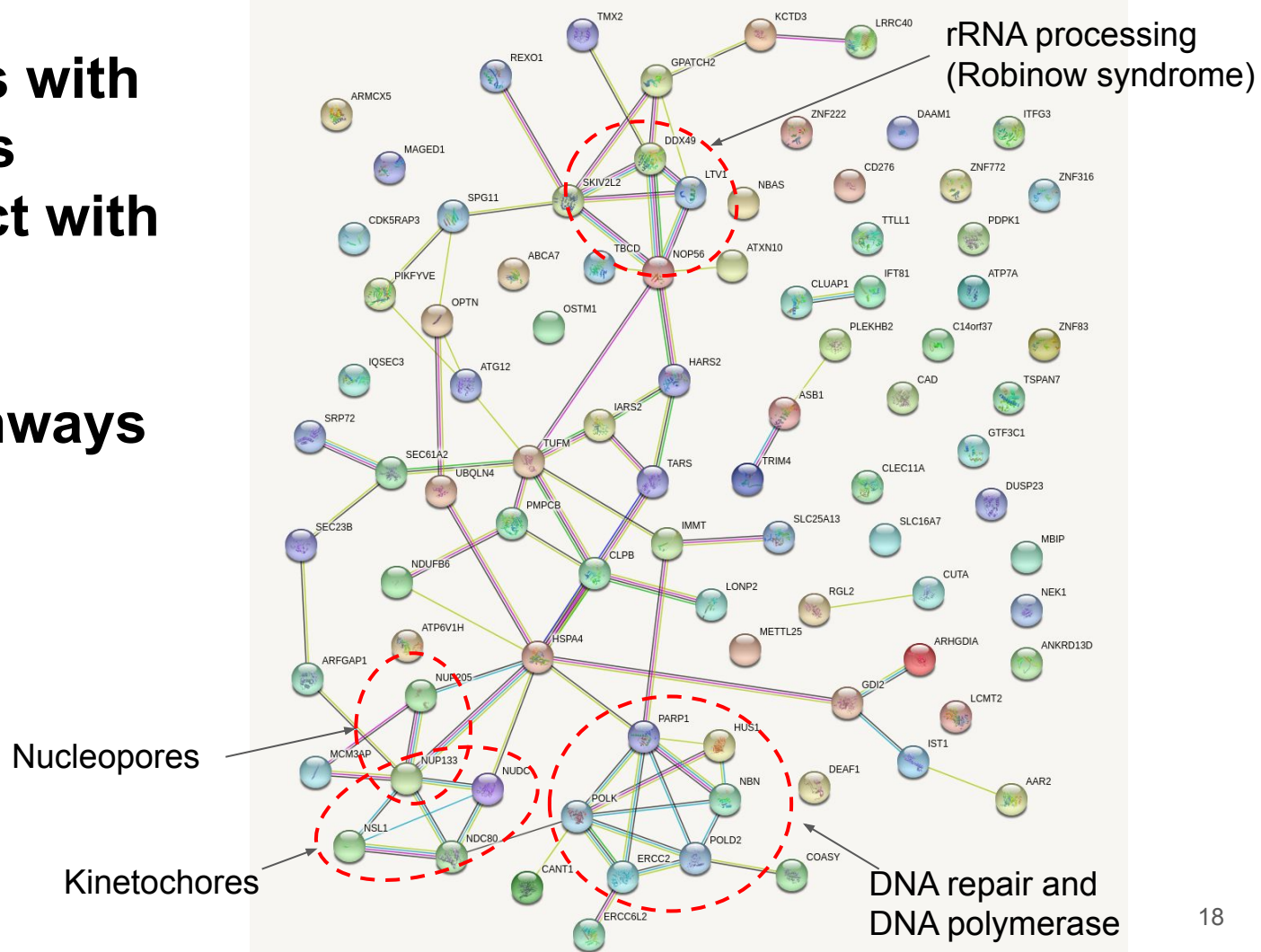
Sample	Gene	fold change	PPI score	Variant	Consequence	ClinVar	Comment
CASE.NEUEK191WYC	<i>NEK1</i>	0.75	1.22E-02	chr4:169424645:G>A	stop		Definitive ALS gene
CASE.NEUBK117YXL	<i>OPTN</i>	0.67	1.23E-02	chr10:13122390:C>A	stop		Definitive ALS gene
CASE.NEUZT557DHF	<i>OPTN</i>	0.67	1.23E-02	chr10:13112464:T>TAG	frameshift		Definitive ALS gene
CASE.NEUVX902YNL	<i>SPG11</i>	0.82	1.23E-02	chr15:44620189:C>A	splice donor	likely pathogenic	Tenuous ALS gene, variant predicted to cause aberrant splicing
CASE.NEULD354RZB	<i>NOP56</i>	1.23	1.61E-04	chr20:2655751:G>A	splice region		Variant predicted to cause aberrant splicing. Gene related to Ataxia.
CASE.NEUTA689LN5	<i>TUFM</i>	0.60	1.06E-04	chr16:28844814:G>A	stop	uncertain significance	Mitochondrial disease gene
CASE.NEUGW326BRV	<i>CLPB</i>	0.58	1.30E-04	chr11:72302312:G>A	stop	pathogenic	Mitochondrial disease gene
CASE.NEUME498PCJ	<i>PIKFYVE</i>	0.75	1.54E-04	chr2:208352730:A>AT	frameshift		Linked to neurodegeneration
CASE.NEURR881FKY	<i>ERCC2</i>	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: Clusters of outliers as new candidates

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**



# Discussion / outlook


- 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
  - Functional follow-ups

# 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 view of ALS

# Code to reproduce the results

<https://github.com/gagneurlab/ALS>

 **gagneurlab / ALS**

Unwatch 4

Star 0

Fork 0


[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#) [...](#)

master

Go to file

Add file

Code

 **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...](https://www.kaggle.com/alsgrou...)

Readme

MIT License

### Releases

No releases published






[Create a new release](#)

### Packages

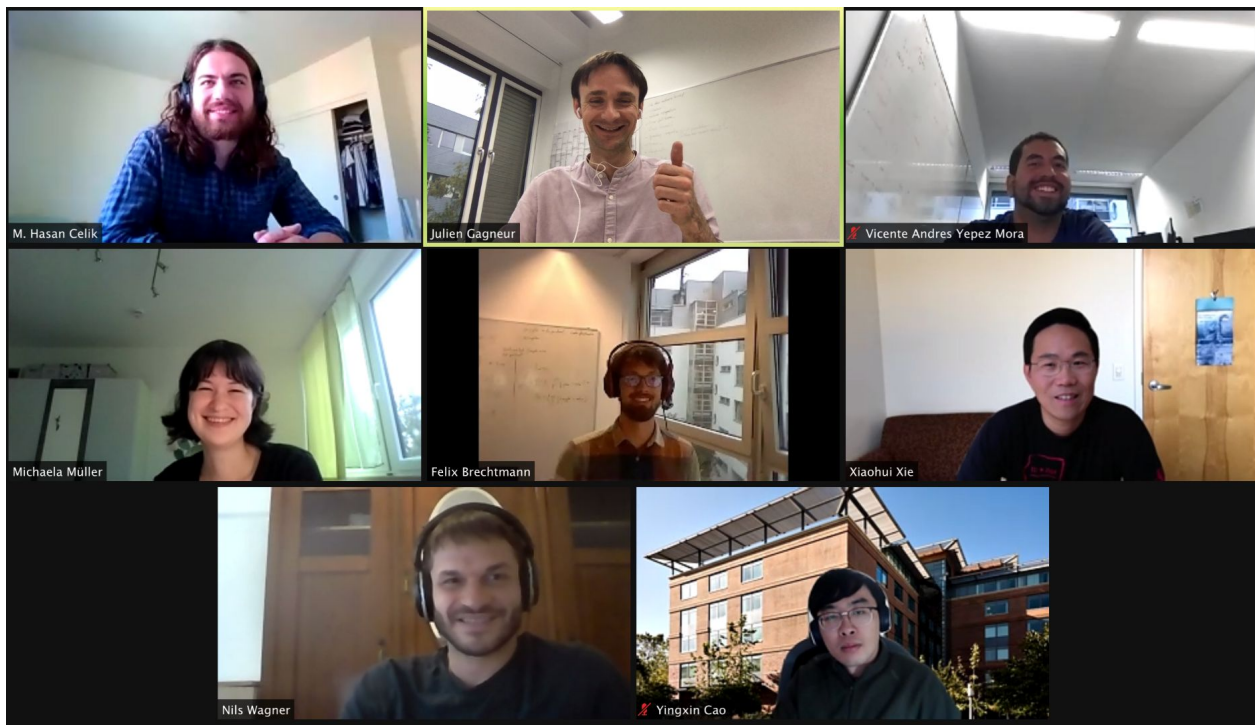
No packages published

[Publish your first package](#)

### Contributors 5



# The team



Felix Brechtmann<sup>1</sup>, Hasan Çelik<sup>2</sup>, Julien Gagneur<sup>1</sup>, Florian Hölzlwimmer<sup>1</sup>, Michaela Müller<sup>1</sup>, Nils Wagner<sup>1</sup>, Xiaohui Xie<sup>2</sup>, Vicente Yépez<sup>1</sup>, Michael Zech<sup>1</sup>



# 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)

