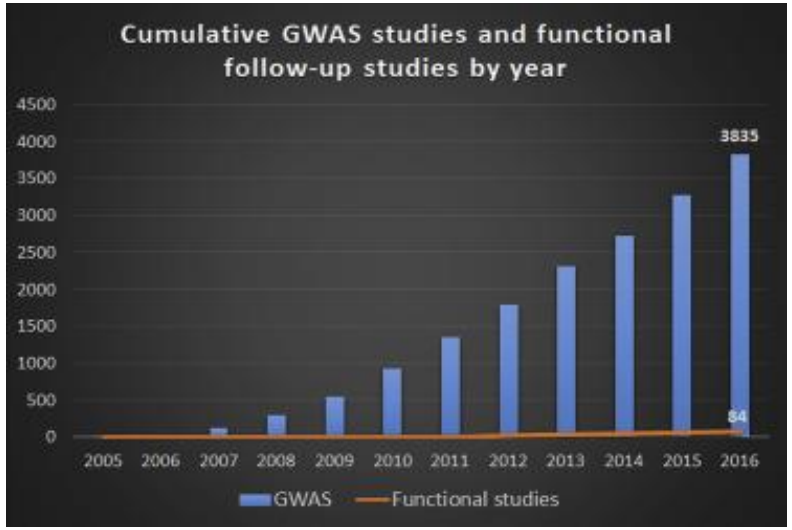
Linking variants to genes and variant effect prediction

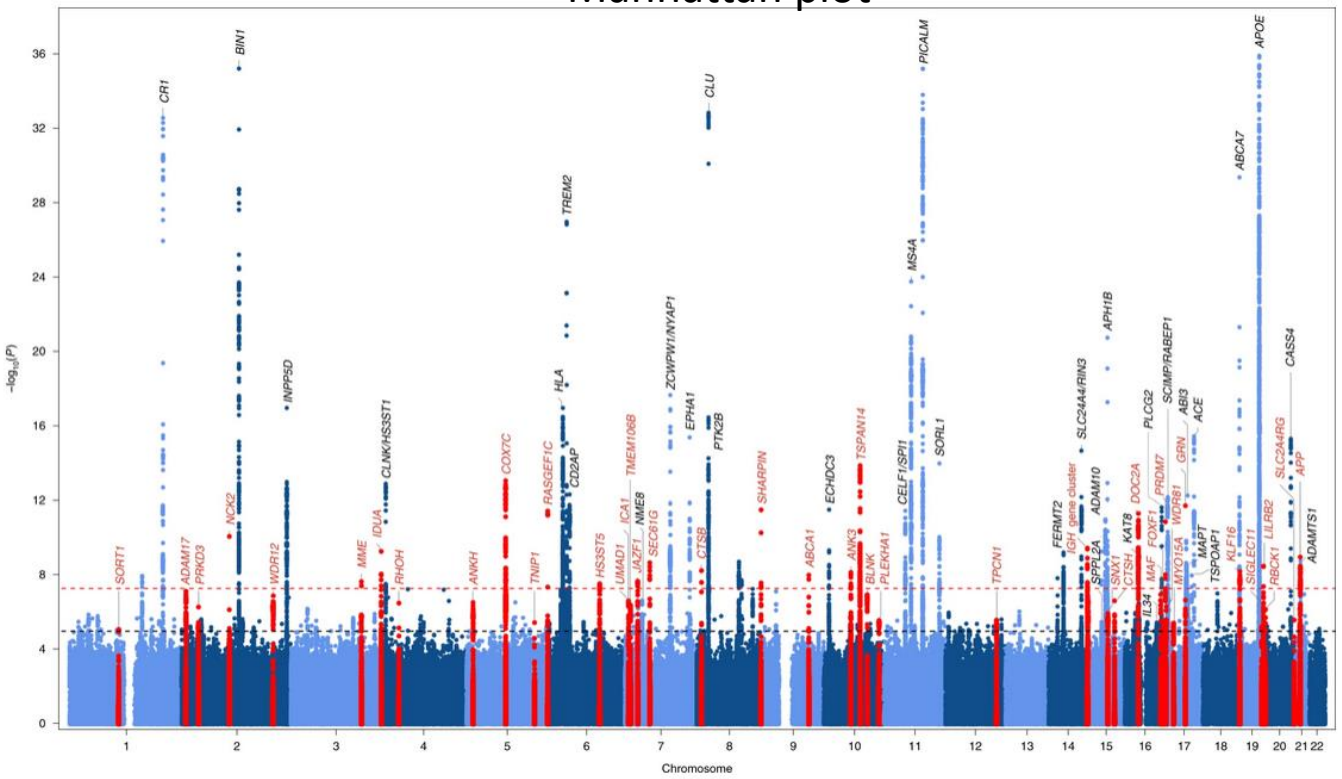
Gerard Bouland
The Delft Bioinformatics Lab, TU Delft
Human Genetics, LUMC

A lot of risk variants but not a lot of understanding



ref: The Post-GWAS Era: From Association to Function (https://doi.org/10.1016/j.ajhg.2018.04.002)

Manhattan plot



ref: New insights into the genetic etiology of Alzheimer's disease and related dementias (https://www.nature.com/articles/s41588-022-01024-z)

How does a genome-wide association study find variants?

rs12548

```
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTAAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTAAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTAAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTTAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTAAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTTAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTTAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTTAGAGGAGTCATT
AAACCTCAGCCTGCTTCTCGTCCGGGTTGTAAGAGGAGTCATT
```

Genomes of individuals with Alzheimer's

Genomes of healthy individuals

Chi-sq = 25.043

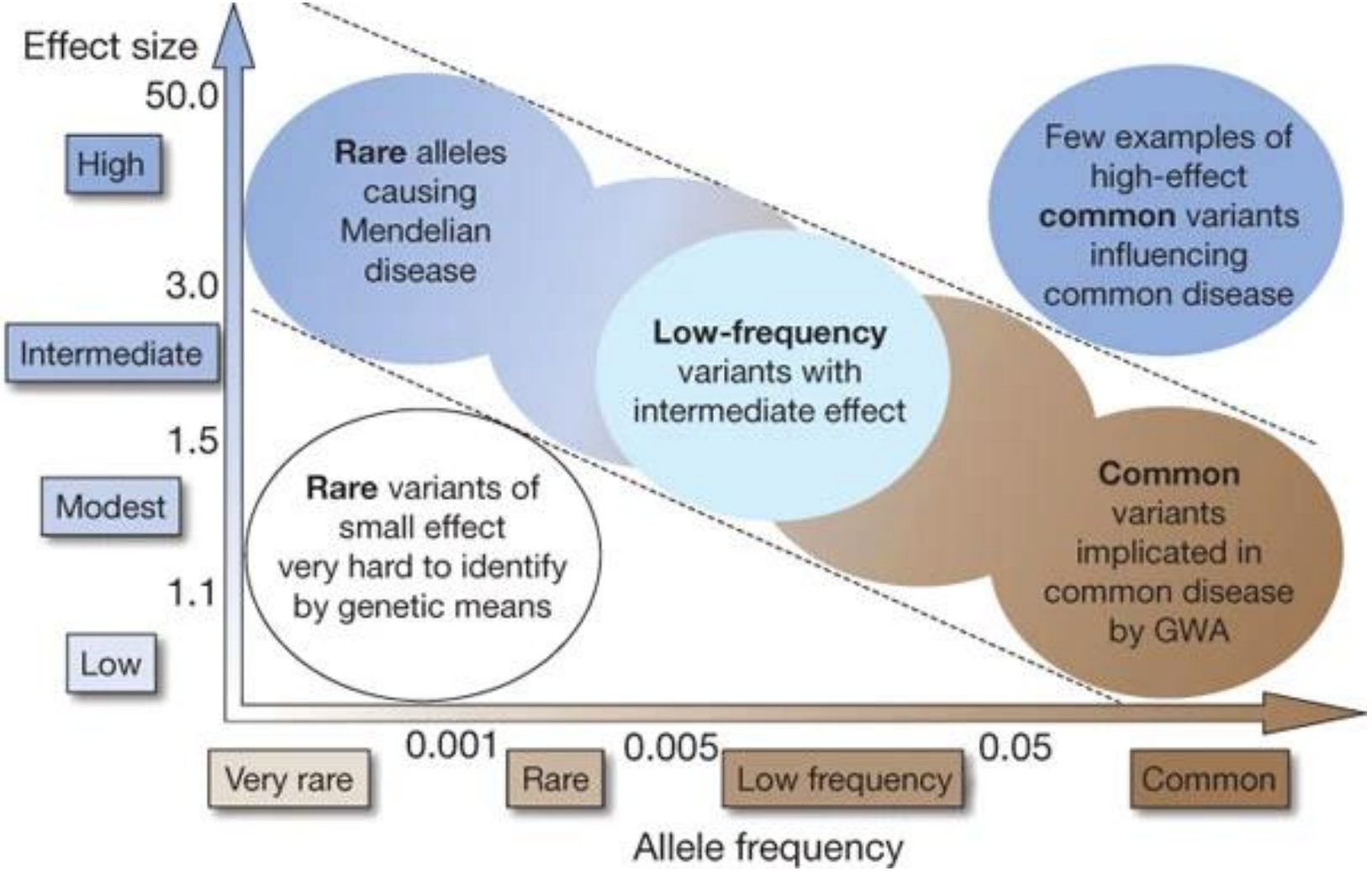
P = 5.61×10^{-7}

Observed	Allele A (N = 50)	Allele T (N = 40)
Alzheimer's (n =50)	40	10
Healthy (n = 40)	10	30

Expected	Allele A (N = 50)	Allele T (N = 40)
Alzheimer's (n =50)	27.8	22.2
Healthy (n = 40)	22.2	17.8

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

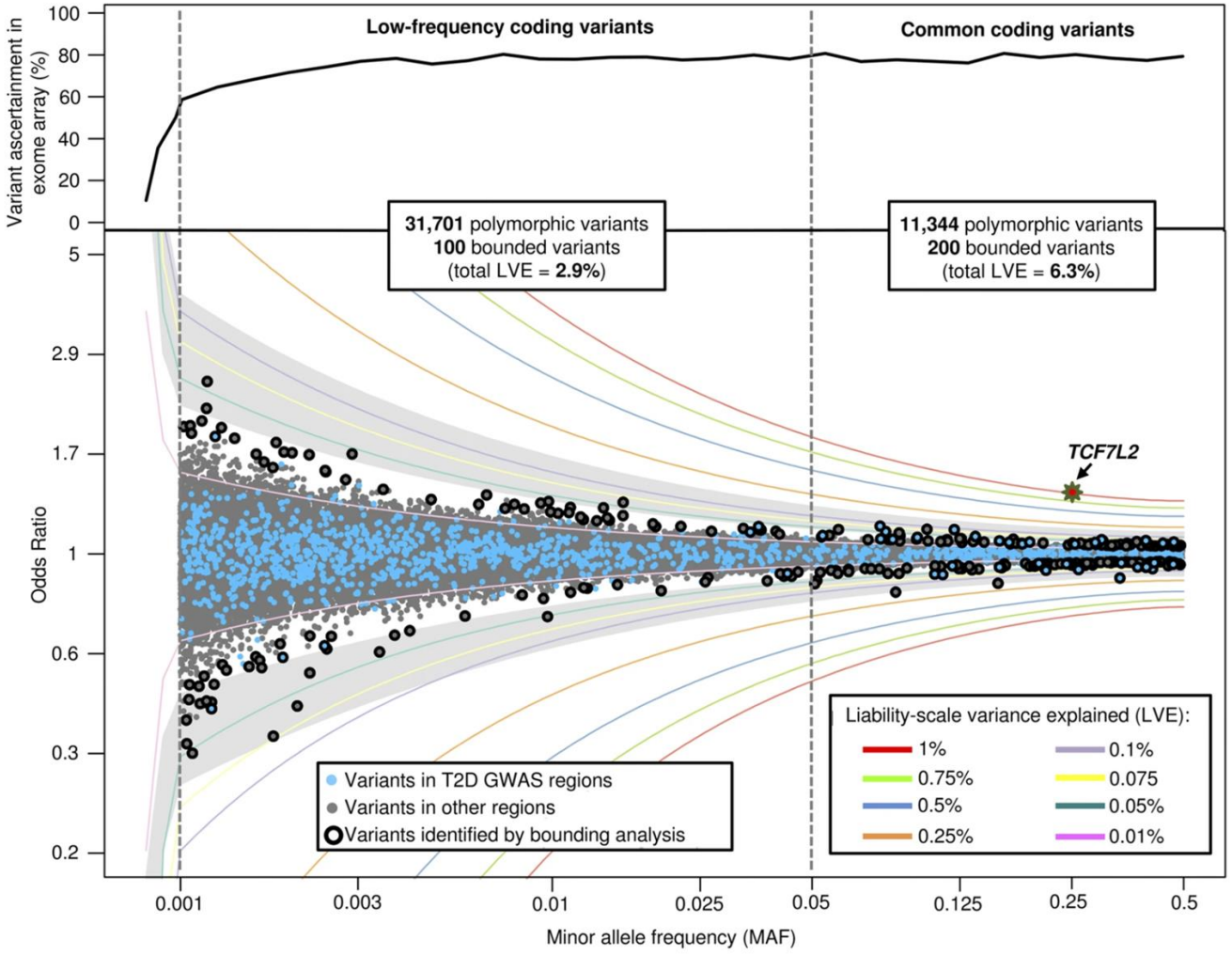
Most common variants have small effects



Manolio et al 2009, PMID: 19812666

Most common variants have small effects

Type 2 diabetes



Fuchsberger, et al. 2016

Multiple hypothesis testing

- Usually, millions of variants are tested in GWAS studies.
- If we assume significance at $P = 0.05$, then you allow that there is a 5% chance that the association you found is random.
- So, if you find a 100 times a $P = 0.05$, then at least 5 of them are likely to be based on chance.
- Each study has hundreds of $P < 0.001$ purely by statistical chance (no real relationship to disease)
- "Genome-wide significance" often set at $P = 5x10^{-8}$ (= .05 / 1 million tests)

Disease associated variants

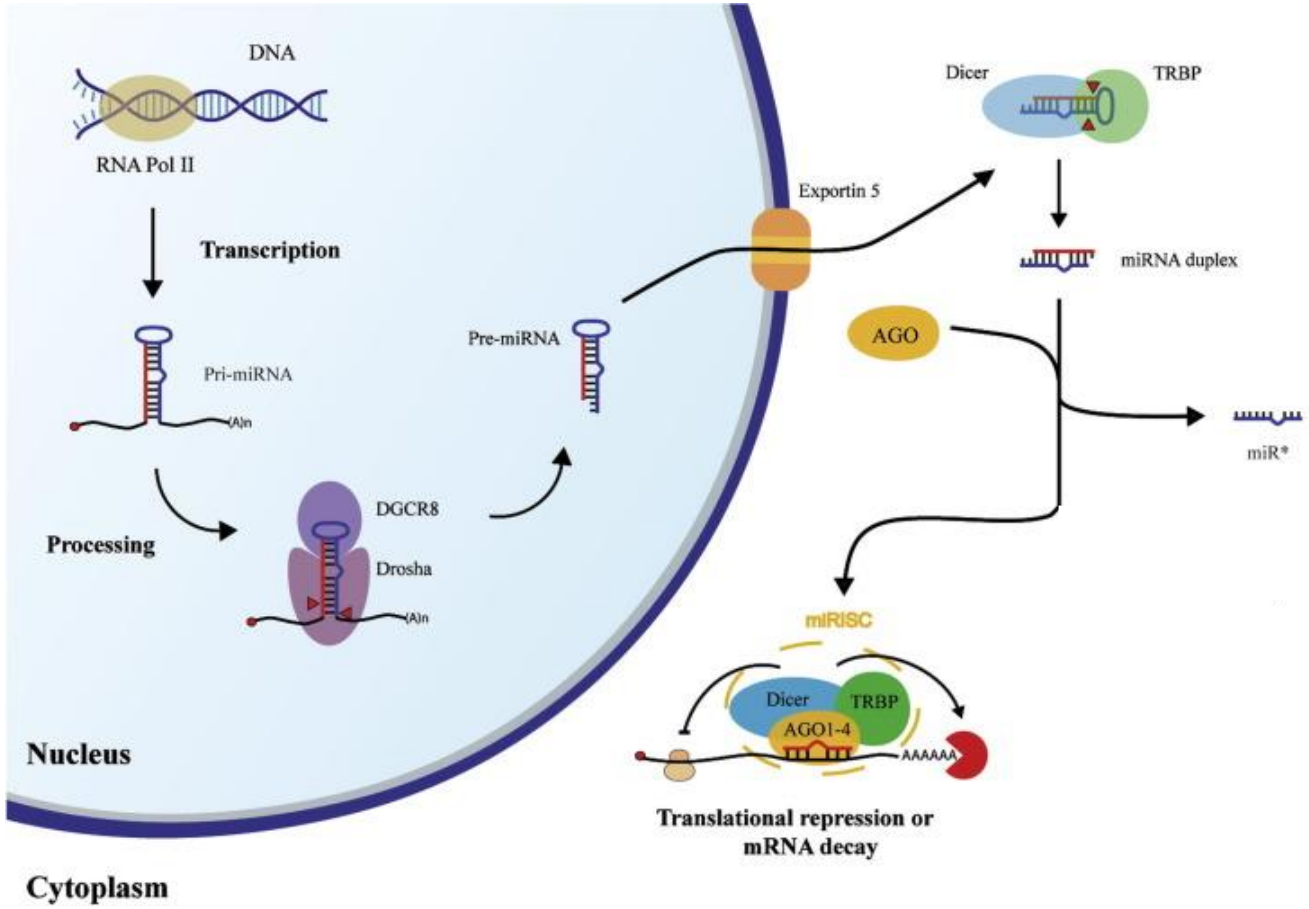
- Alzheimer's (~ 80 variants)
- Schizophrenia (~290 variants)
- Type 2 Diabetes (~240 variants)
- Chronic Kidney Disease (~40 variants)
- Bipolar disorder (~ 65 variants)

- Only 5% of disease-associated SNPs are in gene coding sequences
- So, what is happening with the other 95%???

Biological mechanisms of non-coding variants

1. Variant in 3’-UTR region, altering microRNA binding site
2. Variants altering splice sites
3. Variants altering transcription factor binding site
4. Variants altering DNA methylation

Variant in 3’-UTR region, altering microRNA binding site



Saraiva et al, 2017

	Predicted consequential pairing of target region (top) and miRNA (bottom)
Position 21-28 of HMGA2 3' UTR hsa-let-7f-5p	5' ...GCCAACGUUCGAUUUCUACCUCA... 3' UUGAUAUGUUAGAUGAUGGAGU

https://www.targetscan.org

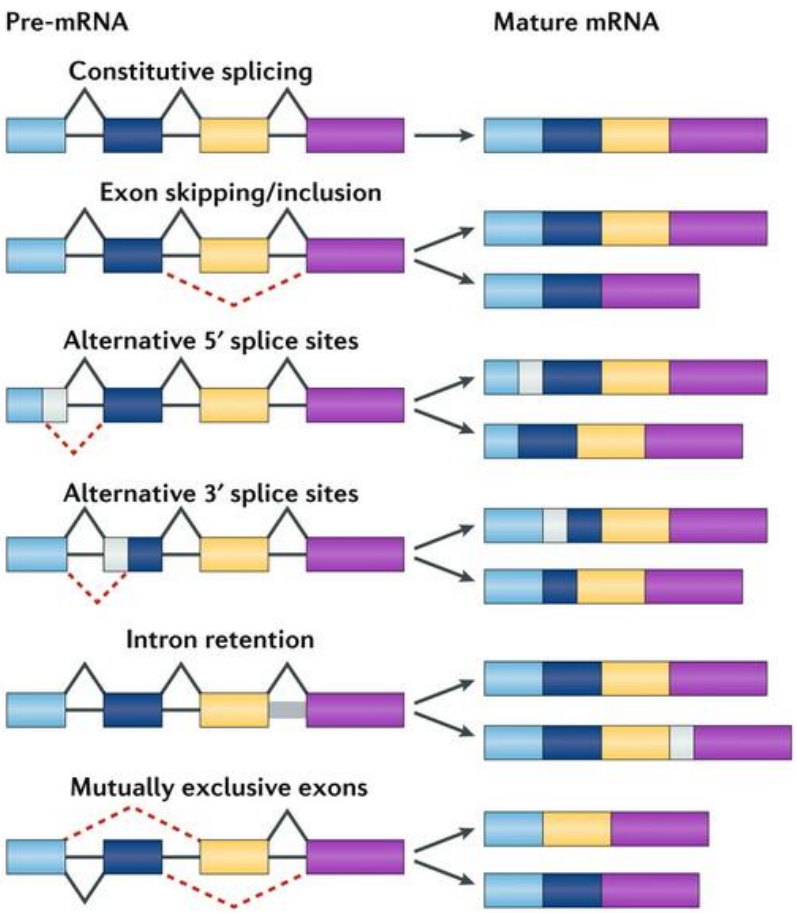
Total N	Significance threshold	N in 3’UTR (% to the UTR; % to the total)
57,671	p value<9×10^{-6}	1,652 (2.9%)
	p value<5×10^{-8}	1,041 (1.8%)

Steri et al, 2019

Variant in 3’-UTR region, altering microRNA binding site

- Alzheimer’s Disease
  - 10 known variants (Ghanbari et al, 2016)
- Autism
  - 21 known variants (Vaishnavi et al, 2014)
- Metastasis in osteosarcoma
  - 1 known variant (Zhang S et al, 2016)

Variants altering splice sites



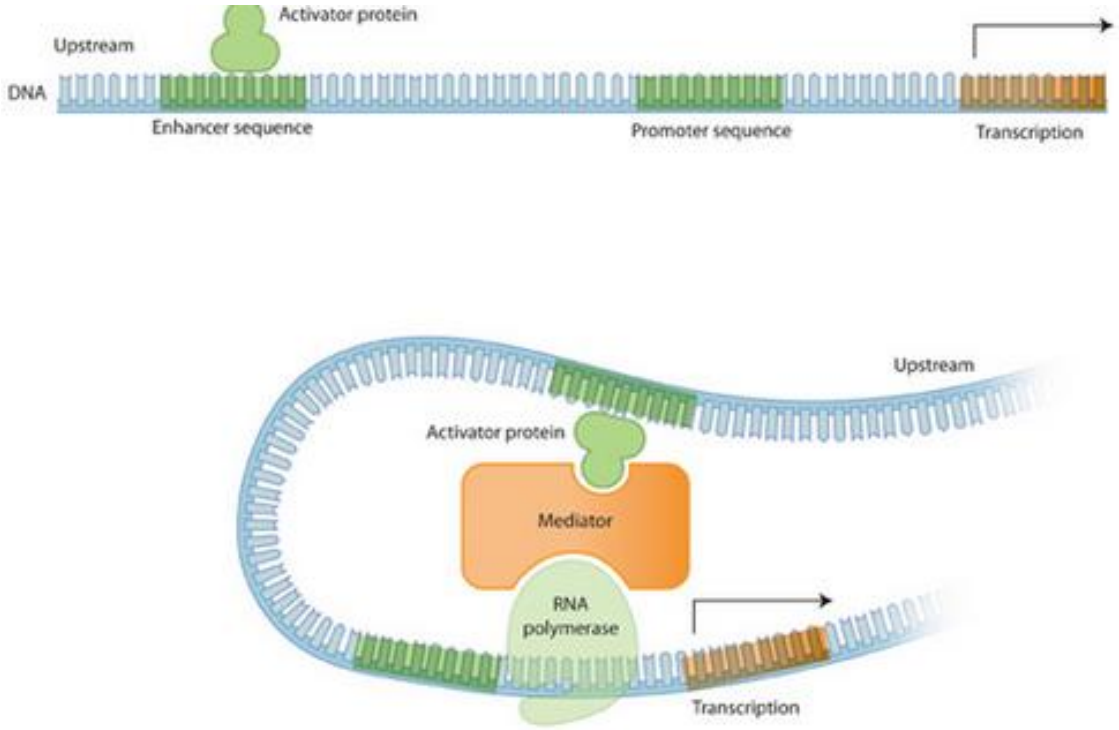
Frankiw et al, 2019

- >99% splice sites GT and AG in the donor and acceptor sites, respectively. (Kurmangaliyev et al, 2013)
- Most disease-causing mutations of splice sites occur at these dinucleotides (Kurmangaliyev et al, 2013)

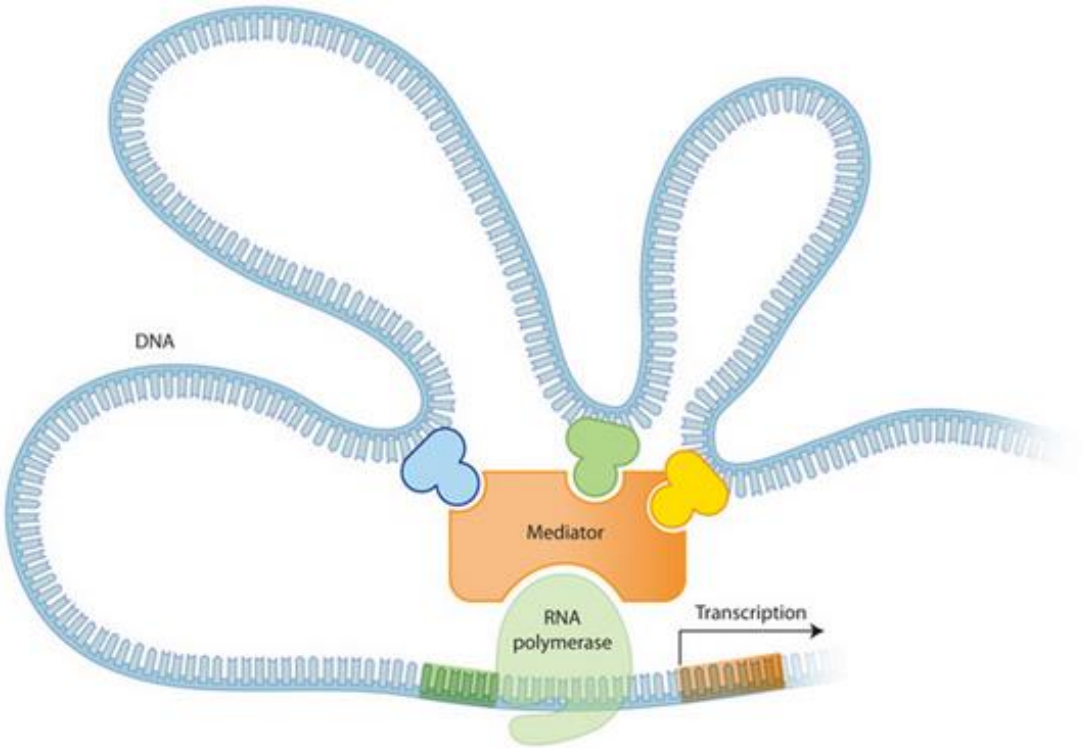
Variants altering splice sites

- Adipose-related traits
  - Seven exon-exon junctions in four genes (*AKTIP*, *DTNBP1*, *FTO* and *UBE2E1*) (**Ma et al, 2015**)
- Alzheimer's Disease
  - 21 genes (**Raj, et al, 2018**)
- Schizophrenia
  - Four genes (*NEK4*, *FXR1*, *SNAP91* or *APOPT1)(**Takata et al, 2017**)*
- Breast Cancer
  - Six genes (*BABAM1, DCLRE1B/PHTF1, PEX14, RAD51L1, SRGAP2D and STXBP4*) (**Caswell et al, 2015**)

Variants altering transcription factor binding site



www.nature.com/scitable



www.nature.com/scitable

Variants altering transcription factor binding site

- 8% of all polymorphisms reside in TFBS, yet it is estimated that of all disease associated polymorphisms 31% reside in TFBS

- BMI
  - *ARID5B*, *IRX3, IRX5 (***Claussnitzer et al, 2015**)
- Chronic kidney disease
  - *TCF7L2 (***Köttgen et al, 2009**)
- Diabetes
  - *NEUROD1*, *MTNR1B (***Lyssenko et al, 2009**)
- Hypertension
  - *PHOX2*, *SCG2(**Wen et al, 2007***)
- Many many more

Variants altering DNA methylation

- Methylation most often happens at CpG sites

- A methylated cytosine represses gene expression and can even lead to a more condensed chromatin structure.

- 23% of all polymorphisms are CpG site related SNPs (**Zhou et al, 2015**)

- Cardiovascular disease (**Jiantao Ma et al 2022**)
  - *APOB*
  - *SREBF1*
  - *PM20D1*

Summary of biological mechanisms

1. Variant in 3’-UTR region, altering microRNA binding site
2. Variants altering splice sites
3. Variants altering transcription factor binding site
4. Variants altering DNA methylation

Finding expression quantitative trait loci (eQTLs)

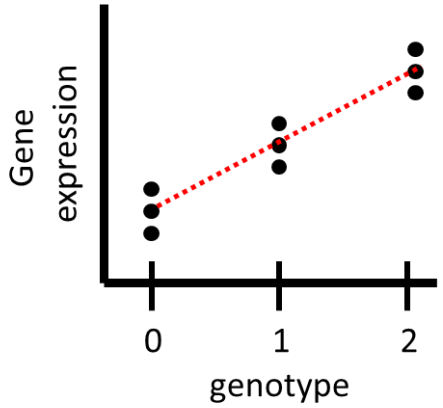
- Expression quantitative trait loci (eQTL) is a SNP that is associated with the expression of a gene.

- An eQTL always forms a pair with an eGene.

- eQTLs are usually identified using linear models.

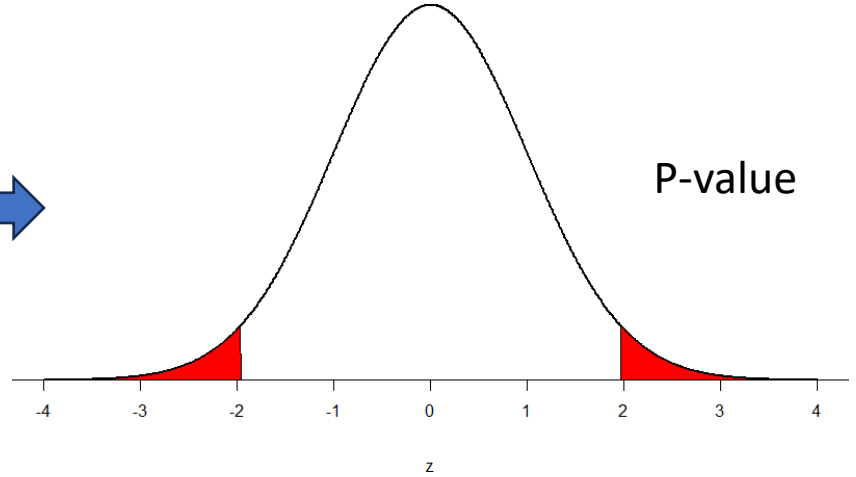
Finding expression quantitative trait loci (eQTLs)

1. Encode genotypes as 0,1,2
   - A/A = 0
   - A/T = 1
   - T/T = 2
2. Associate genotype with gene expression (often linear model)
   - expression = βG + e



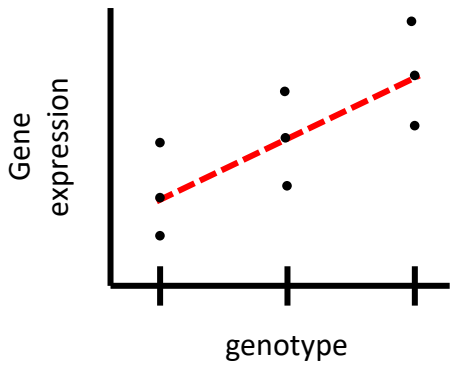
$$Z = \frac{\beta}{\text{Standard Error of } \beta}$$

Wald test where,
H0 : β = 0



Finding expression quantitative trait loci (eQTLs)

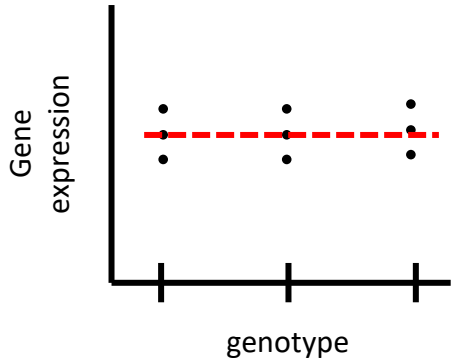
$$Z = \frac{\beta}{\text{Standard Error of } \beta}$$



β = 1
SE β = 0.33

Z: 1 / 0.33 = **3**

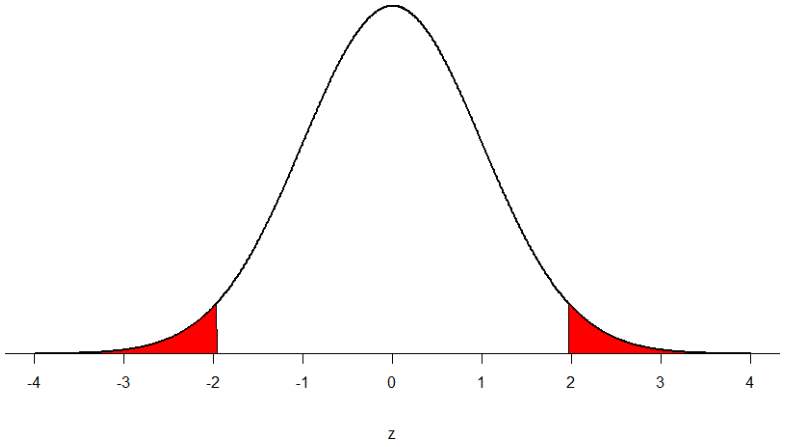
Genotype seems to be predictive of gene expression, but there is still some variability. Small p-value

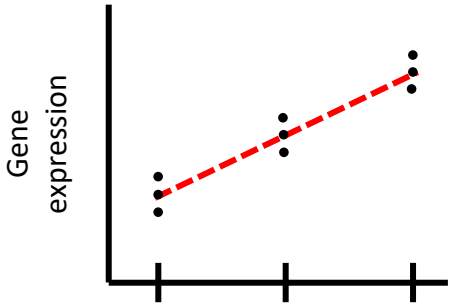


β = 0
SE β = 0.33

Z: 0 / 0.33 = **0**

Genotype says nothing about gene expression. P-value = 1





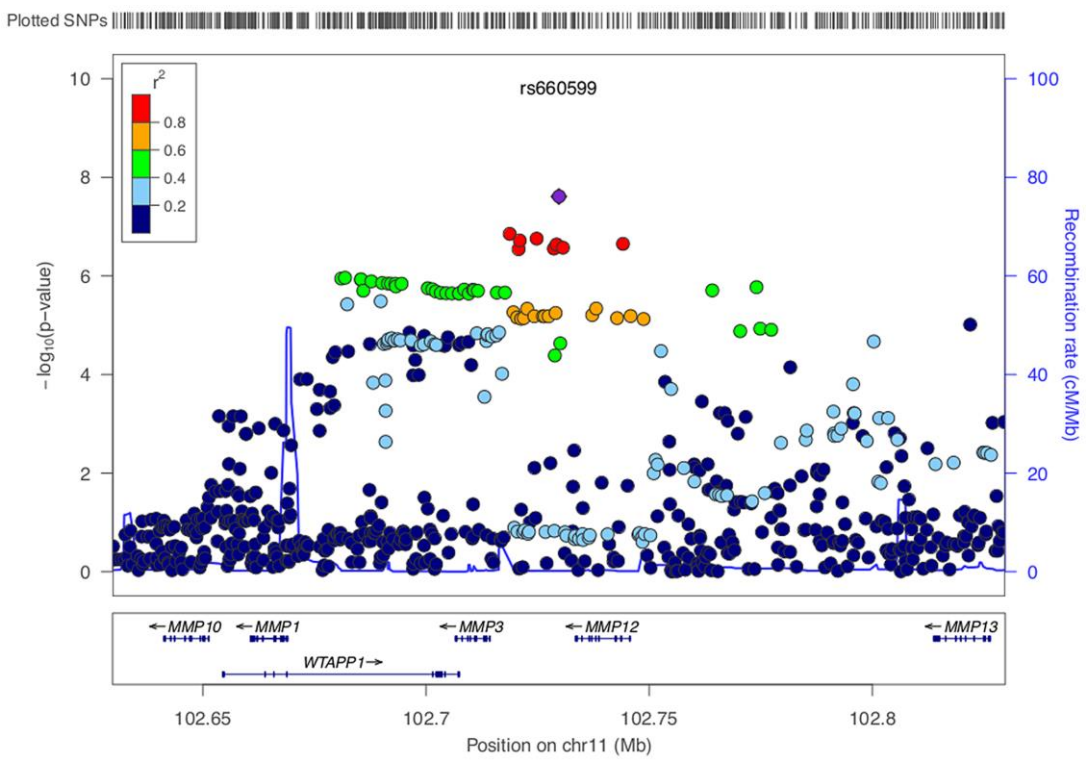
β = 1
SE β = 0.2

Z: 1 / 0.20 = **5**

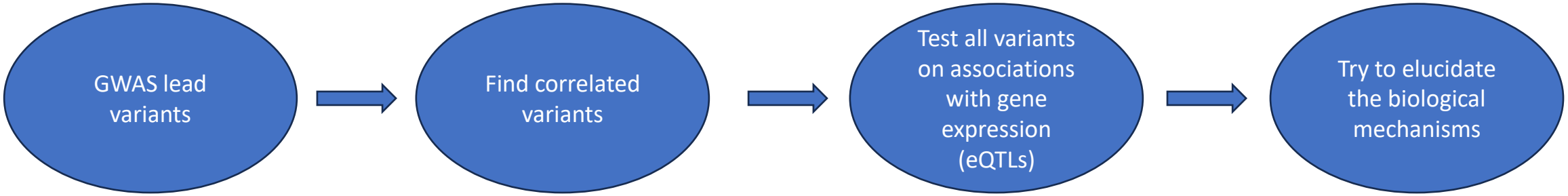
Genotype seems to be predictive of gene expression, very little variability. Very small p-value

Finding expression quantitative trait loci (eQTLs)

- Usually, for every gene SNPs within 1MB of the TSS of the respective gene are tested. (Why? See biological mechanism.)
- Because many SNPs are correlated with each other, a gene is often associated with a whole LD-block.
- Before multiple testing, “independent signals” are identified using clumping.
- So instead of correcting for the total number of SNPs, you correct for the total number of “independent signals”



Combining GWASs with eQTLs



GWAS papers / GWAS catalog

Based on reference panels (e.g. 1000 genomes, HRC)

GTEx / your own data

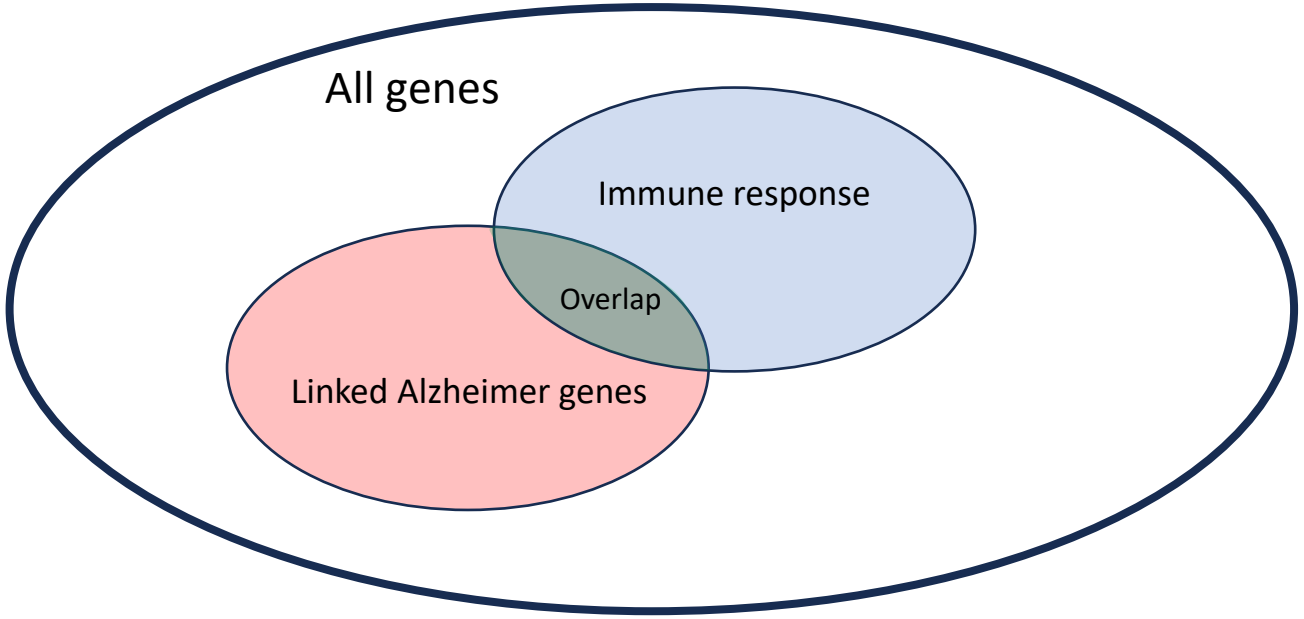
Consequence types (CADD, ensembl)

- 3’-UTR variants
- TFBS variants
- Splice variants
- Methylation
- Histone modifications
- etc

Gene set enrichment analysis

- Assume now that we linked all Alzheimer's variants to genes.
- What could be next thing that we would want to investigate?
- We can test whether the genes are involved in the same pathway or active in the same biological process, molecular function, cellular component or are they interacting with the same metabolite, and many more!

Gene set enrichment analysis



GO:0006955 JSON

immune response

Biological Process

Definition (GO:0006955 GONUTS page)

Any immune system process that functions in the calibrated response of an organism to a potential internal or invasive threat.

408,913 annotations

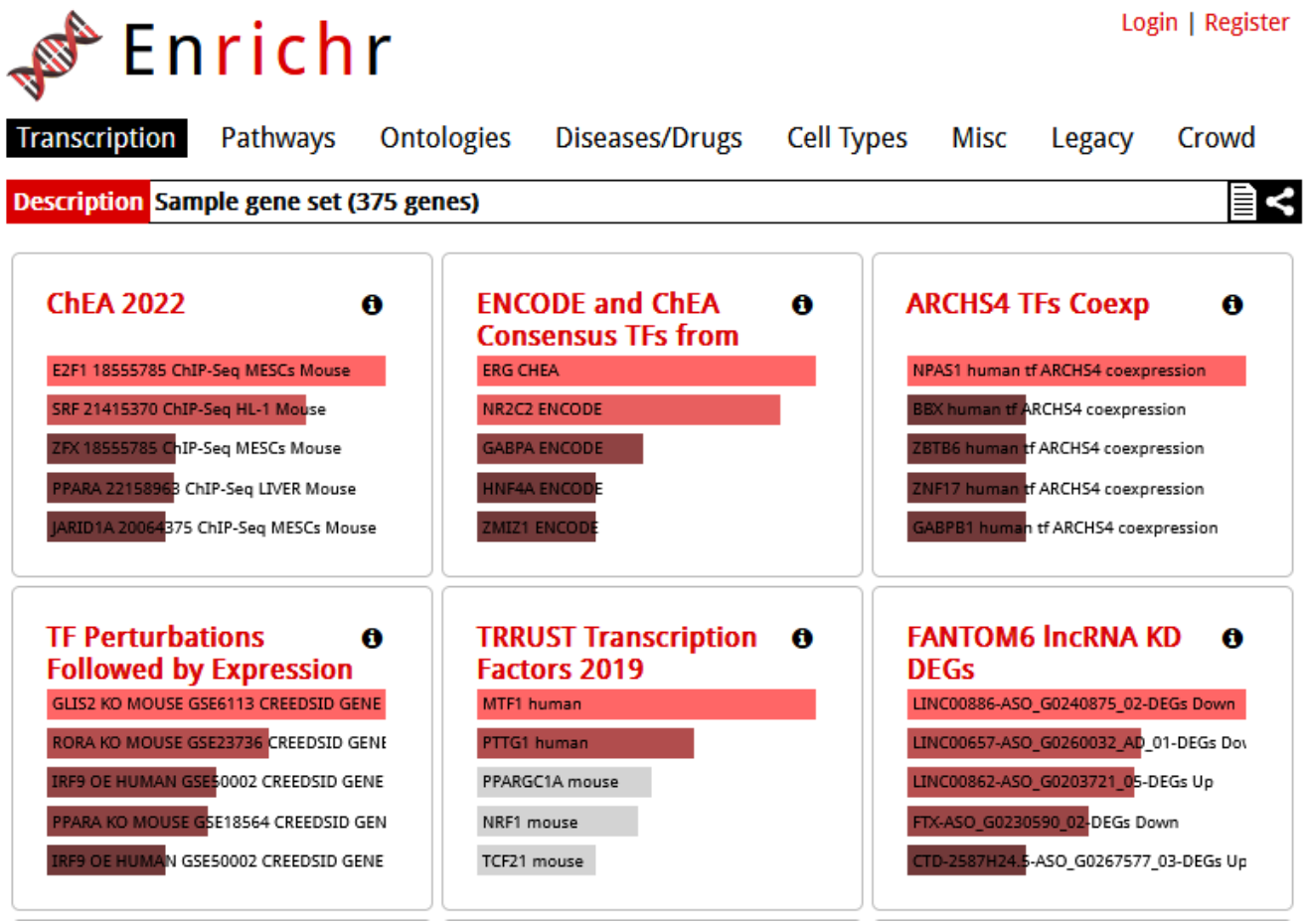
https://www.ebi.ac.uk/QuickGO/term/GO:0006955

Observed	Not Alzheimer Genes (N = 50)	Alzheimer Genes (N = 40)
Not in immune (n =50)	40	10
In Immune (n = 40)	10	30

Expected	Not Alzheimer Genes (N = 50)	Alzheimer Genes (N = 40)
Not in immune (n =50)	27.8	22.2
In Immune (n = 40)	22.2	17.8

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

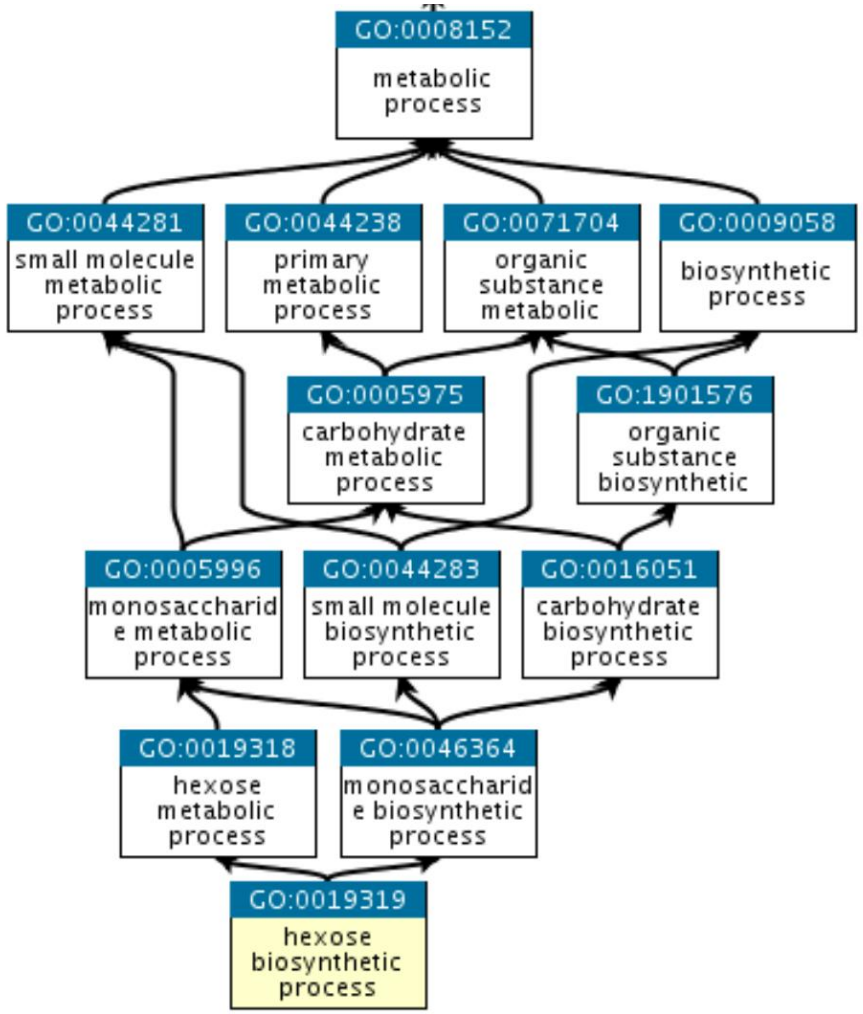
Gene set enrichment analysis



225 different libraries

https://maayanlab.cloud/Enrichr/

Gene ontology



https://geneontology.org

Hierarchy of terms

- Biological processes
- Cellular Component
- Molecular function

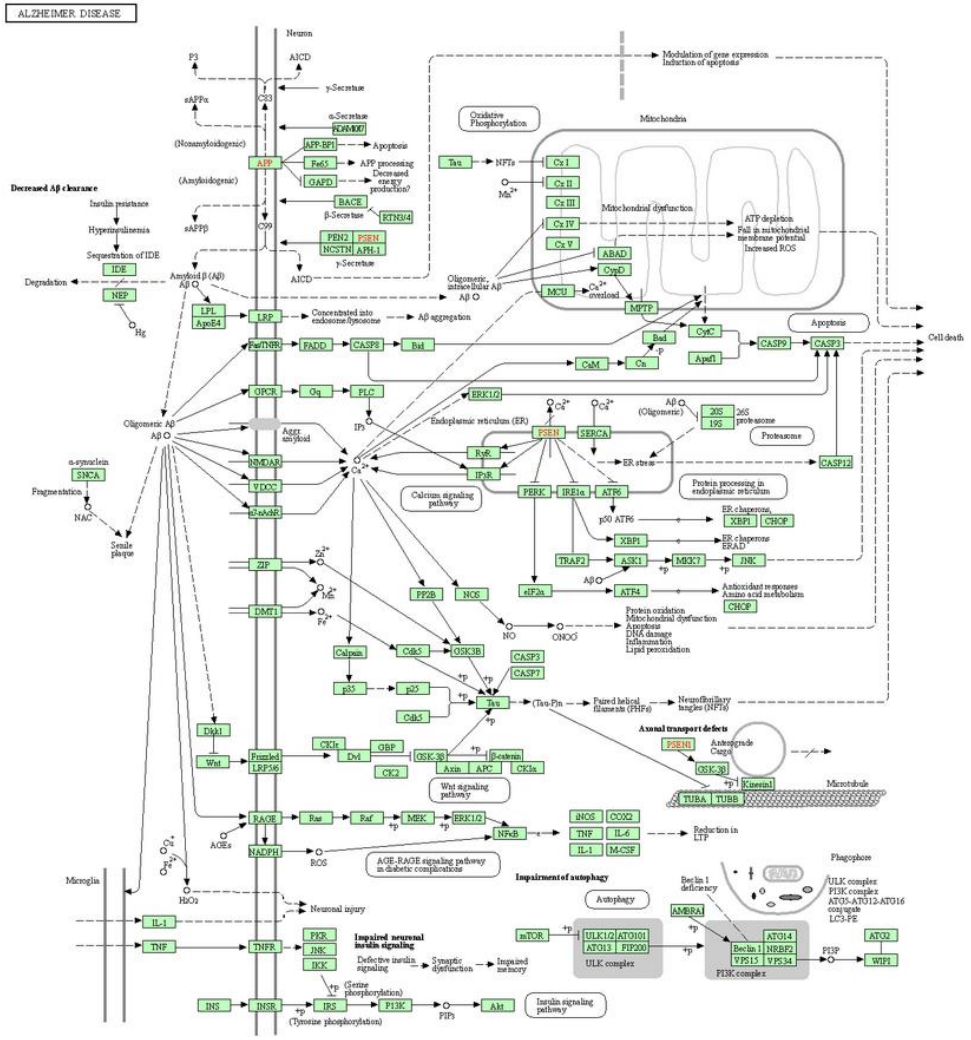
Very extensive (~7.028 terms)

But a lot of overlap between terms.

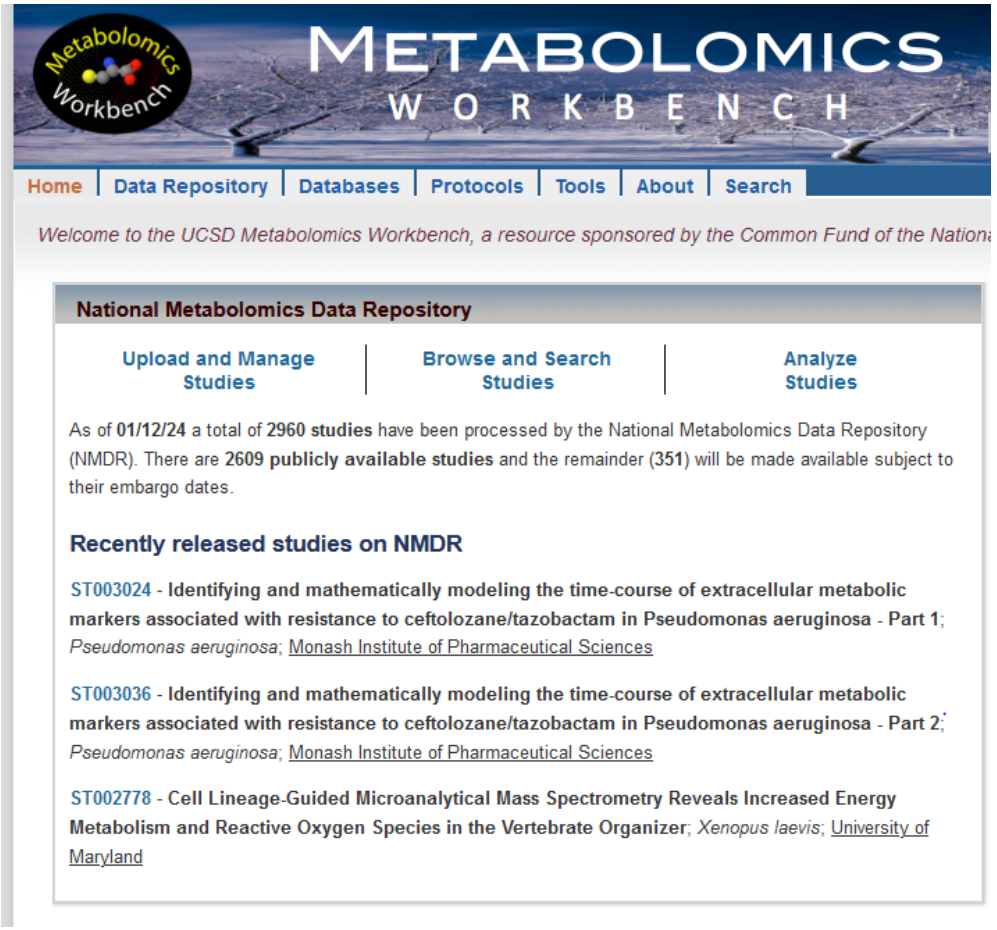
KEGG PATHWAY Database

- *"KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:"*
  1. *Metabolism*
  2. *Genetic Information Processing*
  3. *Environmental Information Processing*
  4. *Cellular Processes*
  5. *Organismal Systems*
  6. *Human Diseases*
  7. *Drug Development*

~320 terms



Metabolomics Workbench Metabolite Database



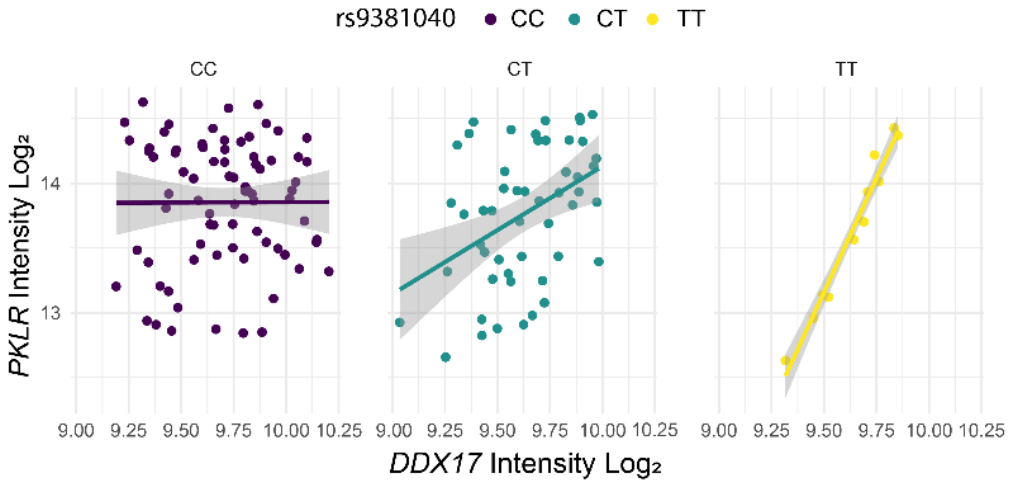
- Database of protein metabolite interactions

233 terms

Co-expression QTLs

Alzheimer's Disease associated with emerging and disrupted protein correlation patterns in Gyrus Temporalis Medialis through genetic variants and pathology

Gerard A. Bouland[1,2], Niccolò Tesi[1,3,4], Meng Zhang[1], Andrea B. Ganz[4], Marc Hulsman, Sven van der Lee, Marieke Graat, Annemieke Rozemuller, Martijn Huisman, Natasja van Schoor, Wiesje van der Flier, Jeroen Hoozemans, August B Smit[4], Marcel J.T. Reinders[1,2,6*], Henne Holstege[1,3,4*]

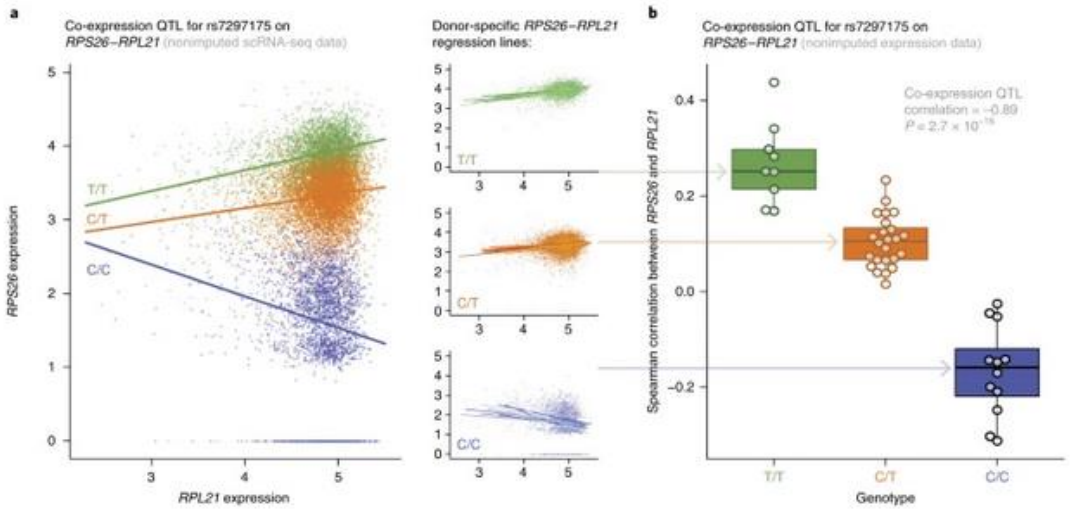


Letter | Published: 02 April 2018

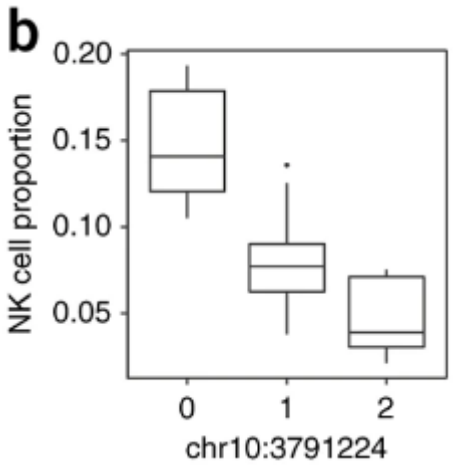
Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs

Monique G. P. van der Wijst, Harm Brugge, Dylan H. de Vries, Patrick Deelen, Morris A. Swertz, LifeLines Cohort Study, BIOS Consortium & Lude Franke

Nature Genetics **50**, 493–497 (2018) | Cite this article



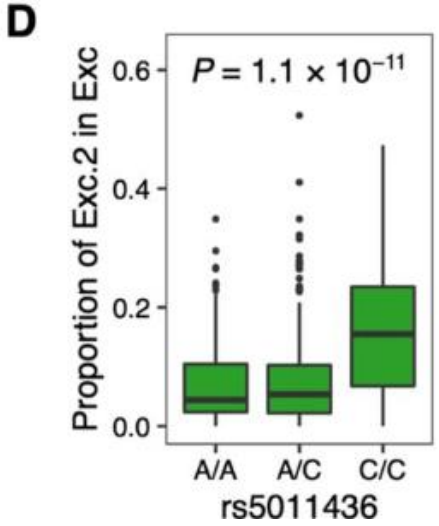
Cell state abundance quantitative trait loci (csaQTLs)



Published: 11 December 2017

Multiplexed droplet single-cell RNA-sequencing using natural genetic variation

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell & Chun Jimmie Ye

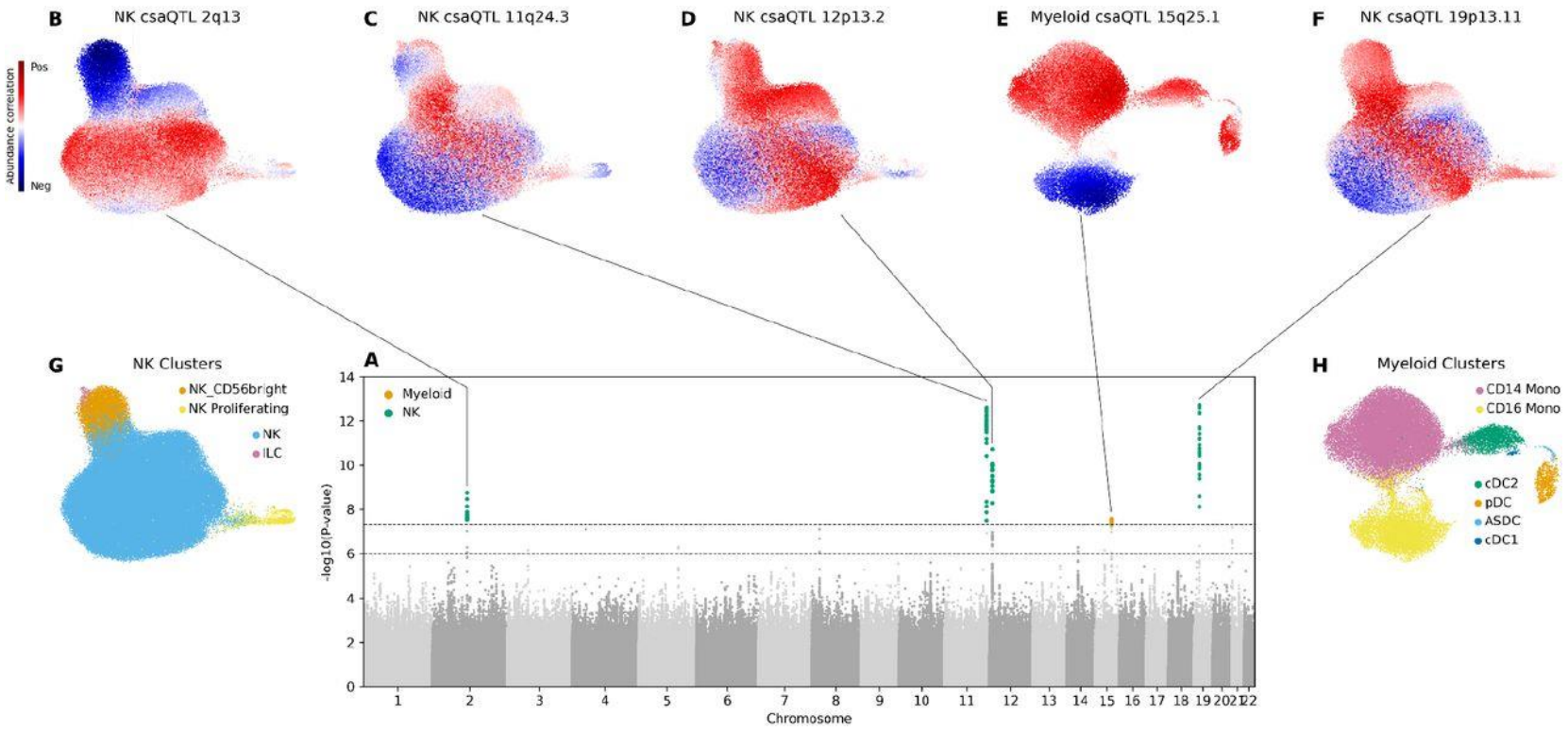


Cell-subtype specific effects of genetic variation in the aging and Alzheimer cortex

Masashi Fujita, Zongmei Gao, Lu Zeng, Cristin McCabe, Charles C. White, Bernard Ng, Gilad Sahar Green, Orit Rozenblatt-Rosen, Devan Phillips, Liat Amir-Zilberstein, Hyo Lee, Richard V. Pearse II, Atlas Khan, Badri N. Vardarajan, Krzysztof Kiryluk, Chun Jimmie Ye, Hans-Ulrich Klein, Gao Wang, Aviv Regev, Naomi Habib, Julie A. Schneider, Yanling Wang, Tracy Young-Pearse, Sara Mostafavi, David A. Bennett, Vilas Menon, Philip L. De Jager

doi: https://doi.org/10.1101/2022.11.07.515446

Cell state abundance quantitative trait loci (csaQTLs)



Identifying genetic variants that influence the abundance of cell states in single-cell data

Laurie Rumker, Saori Sakaue, Yakir Reshef, Joyce B. Kang, Seyhan Yazar, Jose Alquicira-Hernandez, Cristian Valencia, Kaitlyn A Lagattuta, Annelise Mah-Som, Aparna Nathan, Joseph E. Powell, Po-Ru Loh, Soumya Raychaudhuri

doi: https://doi.org/10.1101/2023.11.13.566919