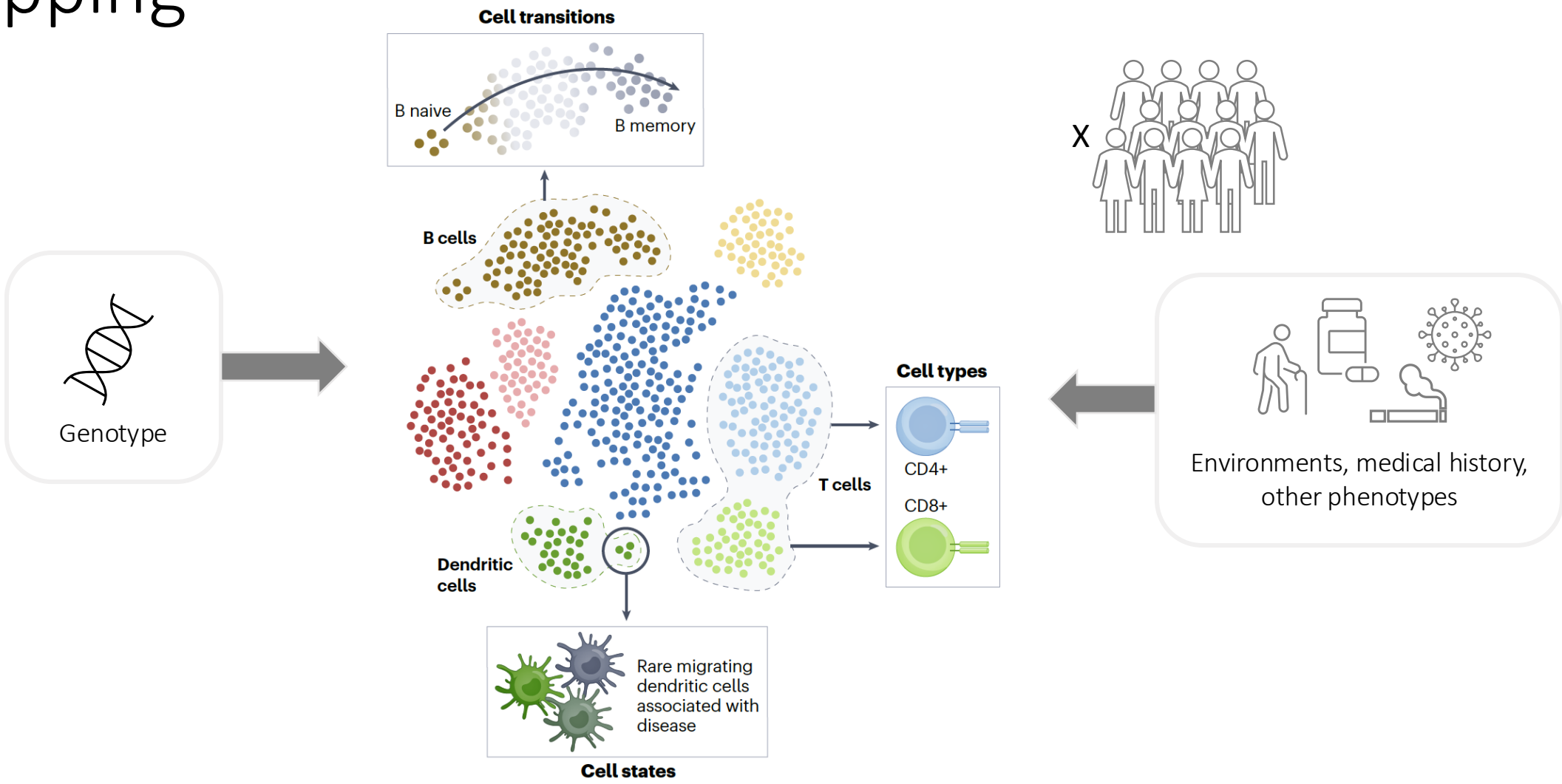
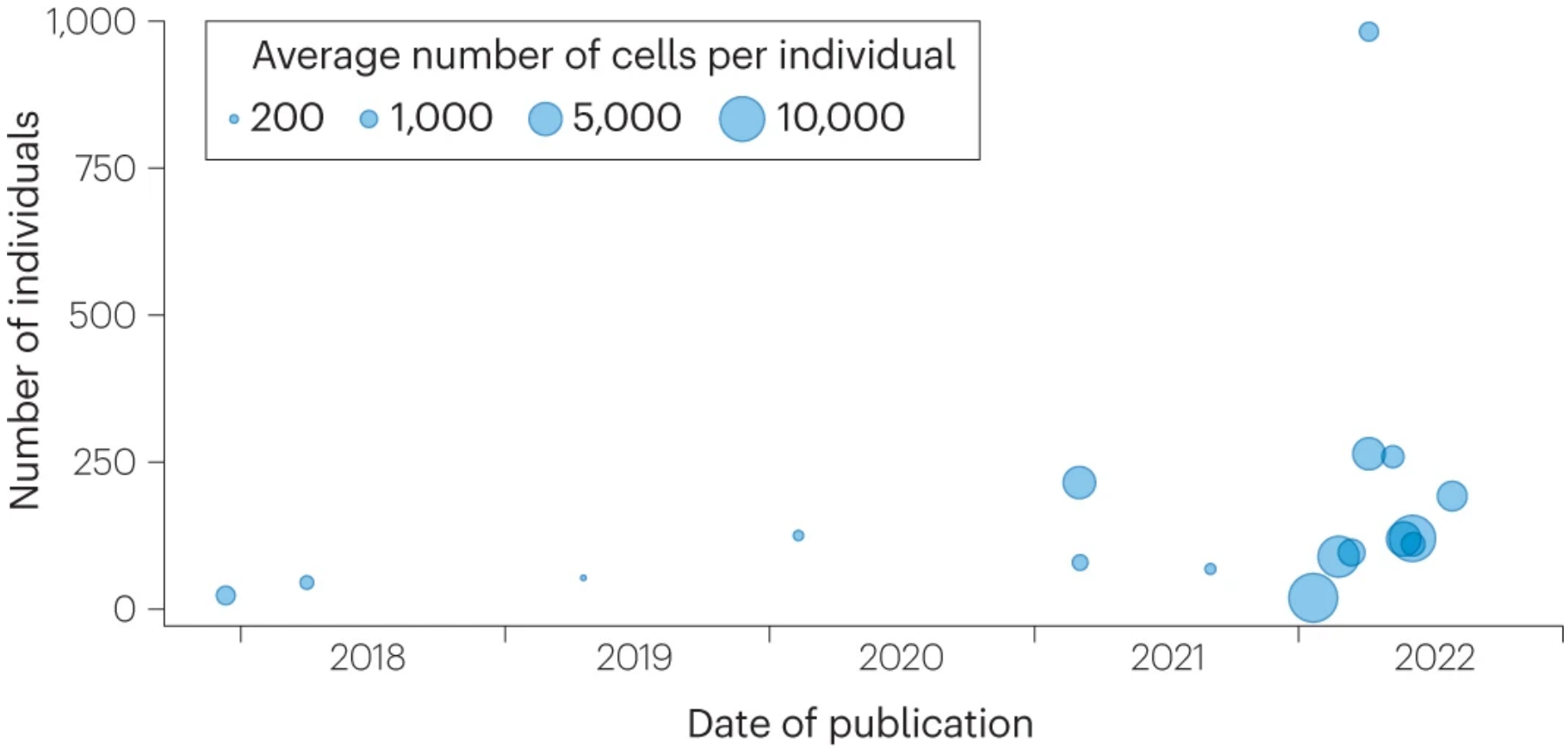
Population-scale single-cell RNA seq empowers eQTL mapping



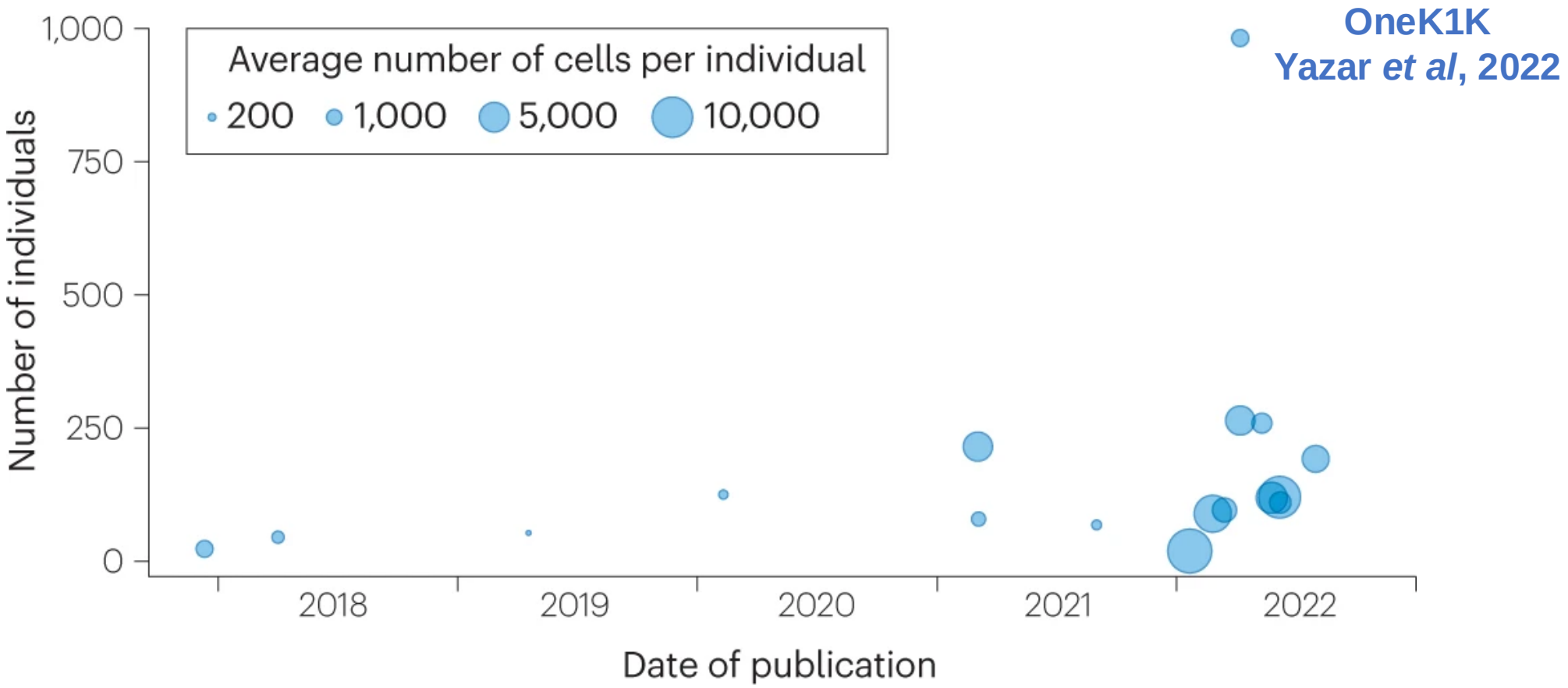
Cuomo *et al*, Nature Reviews Genetics 2023

Single-cell eQTL studies have gained prominence in the past 5 years



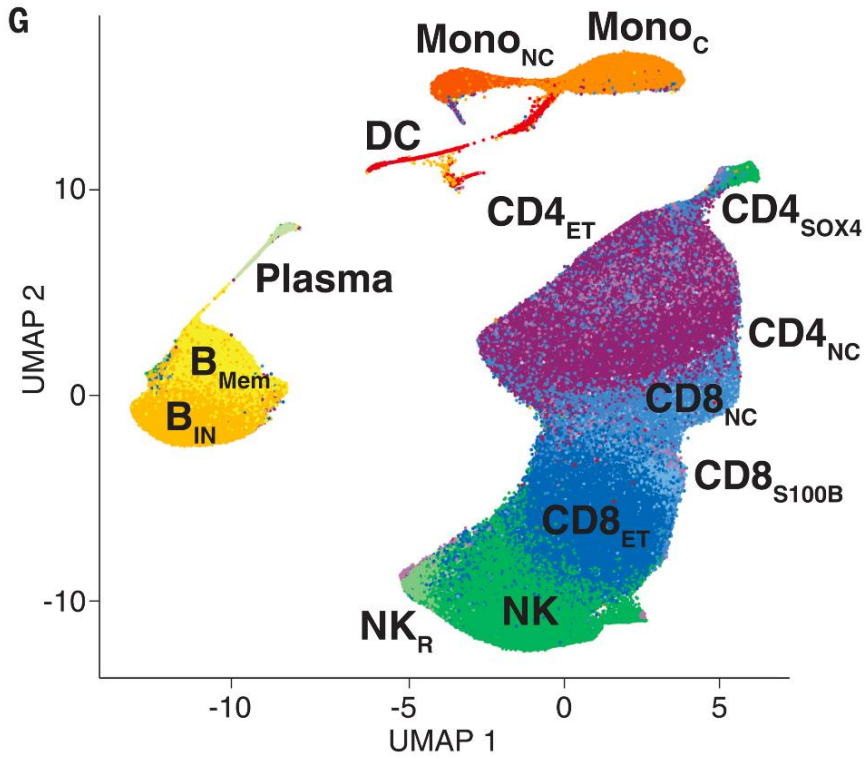
Cuomo *et al*, Nature Reviews Genetics 2023

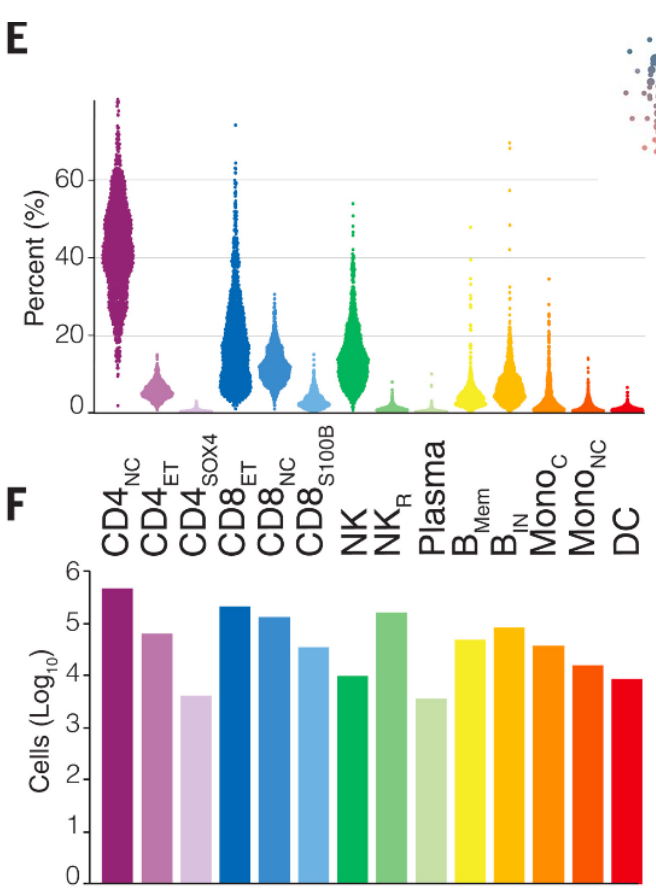
Single-cell eQTL studies have gained prominence in the past 5 years



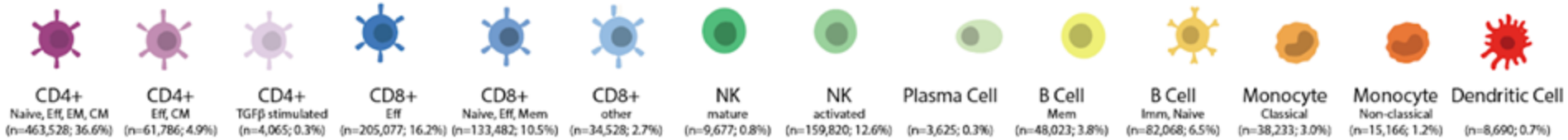
Cuomo *et al*, Nature Reviews Genetics 2023

OneK1K data: ~1,000 immune cells from each of ~1,000 individuals



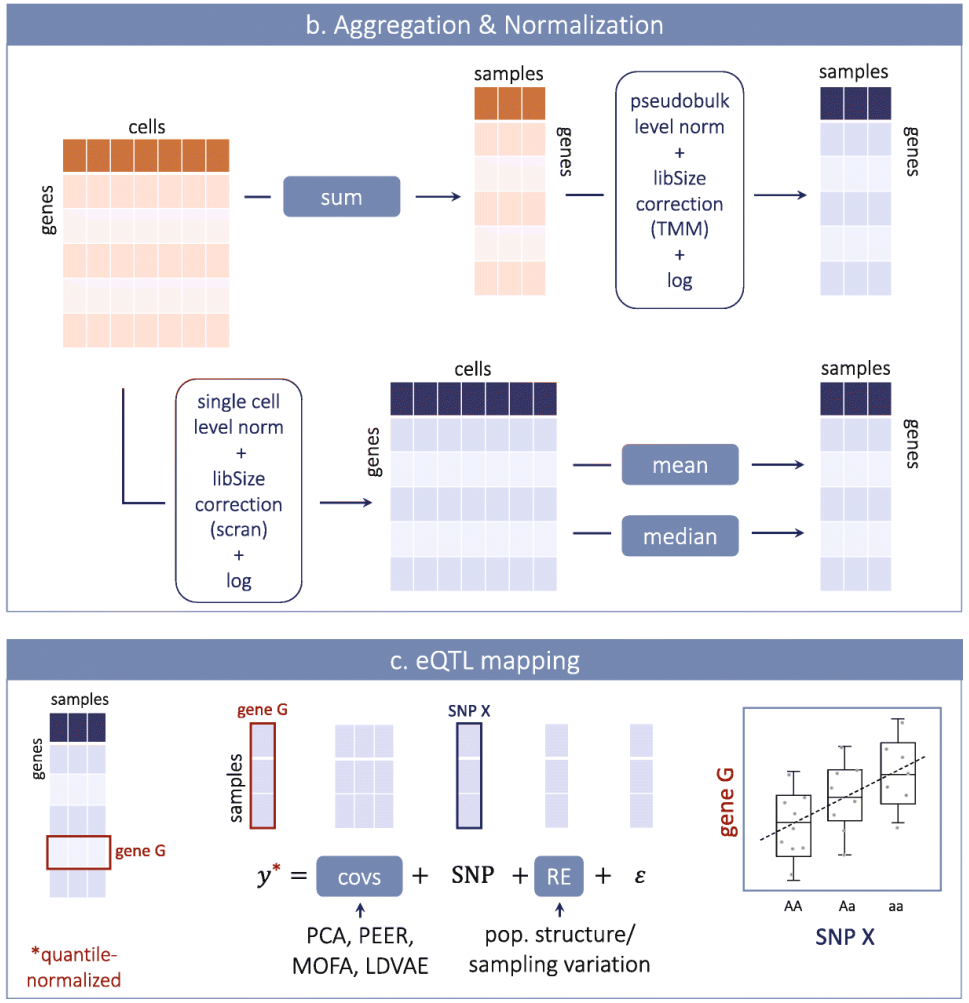


14 immune cell types



Yazar *et al*, 2022

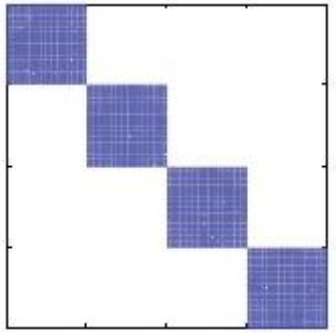
Typically, single-cell RNA-seq data are harmonized to **pseudo-bulk** for eQTL mapping



Cuomo *et al*, Genome Biology, 2021

Challenges of mapping eQTLs at single-cell resolution

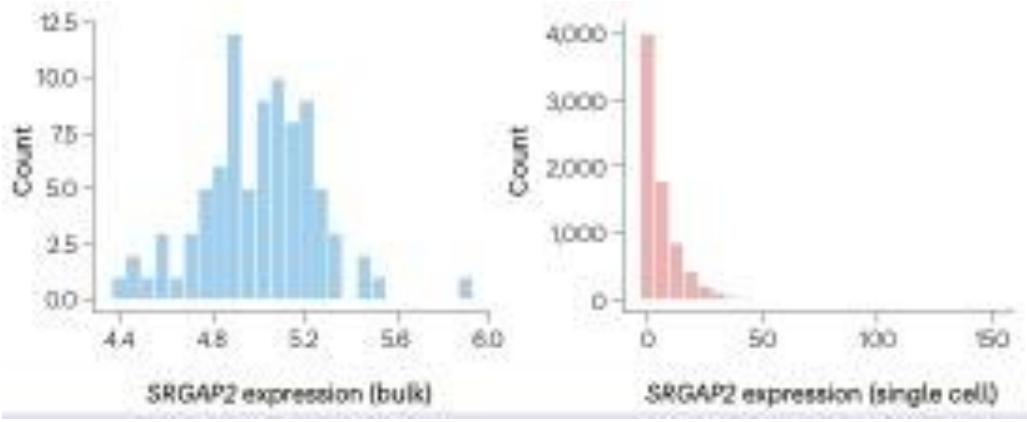
1. **Model repeat and complex data structure**, due to multiple cells per individual and relatedness between individuals



Challenges of mapping eQTLs at single-cell resolution

1. **Model repeat and complex data structure**, due to multiple cells per individual and relatedness between individuals

2. **Model discrete read counts**



Expression levels of the *SRGAP2* gene in induced pluripotent stem cells from the same ~100 individuals (Cuomo *et al*, Nature Reviews Genetics 2023)

Challenges of mapping eQTLs at single-cell resolution

1. **Model repeat and complex data structure**, due to multiple cells per individual and relatedness between individuals

2. **Model discrete read counts**

3. **Fast and scalable for large data**, test 20k genes, 200 cell types, millions of cells, millions of variants

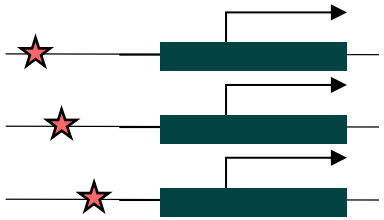
Challenges of mapping eQTLs at single-cell resolution

1. **Model repeat and complex data structure**, due to multiple cells per individual and relatedness between individuals

2. **Model discrete read counts**

3. **Fast and scalable for large data**, test 20k genes, 200 cell types, millions of cells, millions of variants

4. **Test rare variations.** Single-variant test is underpowered



Motivation:

NO scalable computational tools exist to handle all current challenges for eQTL mapping with sc-RNA seq data

Mapping sc-eQTLs shares similar challenges with conducting GWAS for human diseases in large-scale biobanks

1. Model repeat and complex data structure
   - sc-eQTL: multiple cells per individual and relatedness between individuals
   - GWAS: population substructure and relatedness between individuals
2. Model phenotypes that are not continuous and Normally distributed
   - sc-eQTL: discrete read counts
   - GWAS: binary disease status, cases << controls
3. Fast and scalable for large data
   - sc-eQTL: test 20k genes, 200 cell types, millions of cells, millions of variants
   - GWAS: millions of genetic markers for thousands of human disease phenotypes
4. Test rare variations. Single-variant test is underpowered

SAIGE/SAIGE-GENE: R package suite developed for testing common and rare variant associations in large-scale biobanks

nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0184-y

Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies

Wei Zhou[1,2], Jonas B. Nielsen[3], Lars G. Fritsche[2,4,5], Rounak Dey[2,5], Maiken E. Gabrielsen[4], Brooke N. Wolford[1,2], Jonathon LeFaive[2,5], Peter VandeHaar[2,5], Sarah A. Gagliano[2,5], Aliya Gifford[6], Lisa A. Bastarache[6], Wei-Qi Wei[6], Joshua C. Denny[6,7], Maoxuan Lin[3], Kristian Hveem[4,8], Hyun Min Kang[2,5], Goncalo R. Abecasis[2,5], Cristen J. Willer[1,3,9,10*] and Seunggeun Lee[2,5,10*]

TECHNICAL REPORT

https://doi.org/10.1038/s41588-020-0621-6

nature genetics

Check for updates

Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts

Wei Zhou[1,2,3,4,13] ✉, Zhangchen Zhao[1,5,13], Jonas B. Nielsen[6], Lars G. Fritsche[1,5], Jonathon LeFaive[1,5], Sarah A. Gagliano Taliun[1,5], Wenjian Bi[1,5], Maiken E. Gabrielsen[7], Mark J. Daly[2,3,4,8], Benjamin M. Neale[2,3,4], Kristian Hveem[7,9], Goncalo R. Abecasis[1,5], Cristen J. Willer[6,10,11] and Seunggeun Lee[1,5,12] ✉

BRIEF COMMUNICATION

https://doi.org/10.1038/s41588-022-01178-w

nature genetics

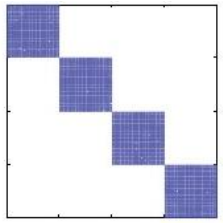
Check for updates

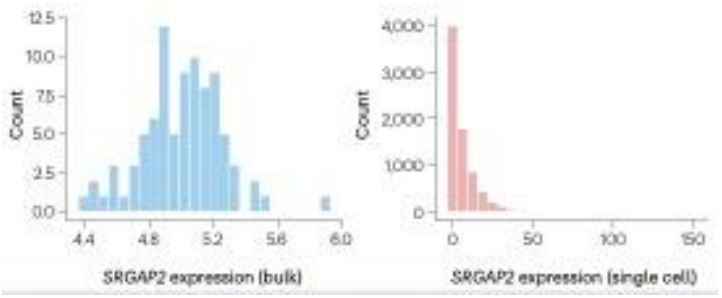
OPEN

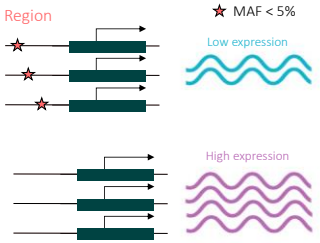
SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests

Wei Zhou[1,2,3,10] ✉, Wenjian Bi[4,5,6,10] ✉, Zhangchen Zhao[5,6,10], Kushal K. Dey[7], Karthik A. Jagadeesh[7], Konrad J. Karczewski[1,2,3], Mark J. Daly[1,2,3,8], Benjamin M. Neale[1,2,3] and Seunggeun Lee[9] ✉

SAIGE-QTL: Generalized Linear Poisson Mixed-model for sc-eQTL mapping







1. **Model repeat and complex data structure**, due to multiple cells per individual and relatedness between individuals

Mixed model with random effects

2. **Model discrete read counts**

Generalized Poisson regression

3. **Fast and scalable for large data**, test 20k genes, 200 cell types, millions of cells, millions of variants

Approximations and computational speedups

4. **Test rare variations.** Single-variant test is underpowered

Set-based tests

SAIGE-QTL: Generalized Linear Poisson Mixed-model for sc-eQTL mapping

$$\boldsymbol{y \sim Poisson(\mu)}$$

$$\boldsymbol{\mu = exp(\eta)}$$

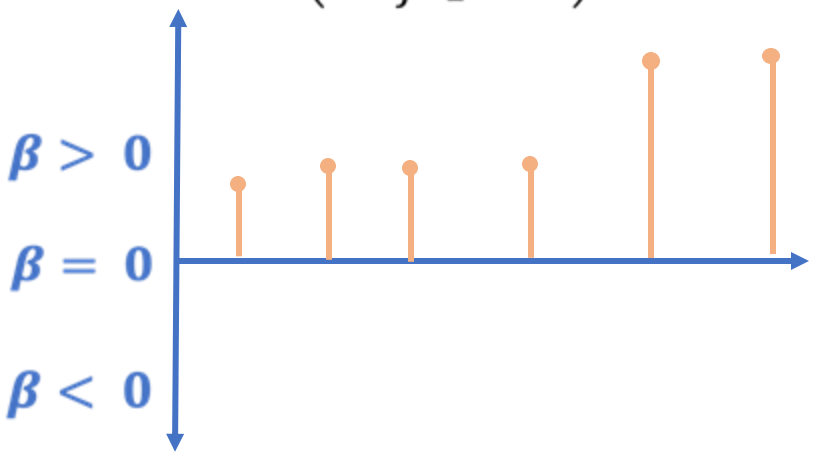
$$\boldsymbol{\eta = ZX_{ind}\alpha_1 + X_{cell}\alpha_2 + ZG\beta + Zb_{ind}}$$

$$\boldsymbol{b_{ind}} \sim N\left(0, \sum_{k=1}^{K} \tau_k \Psi_k\right)$$

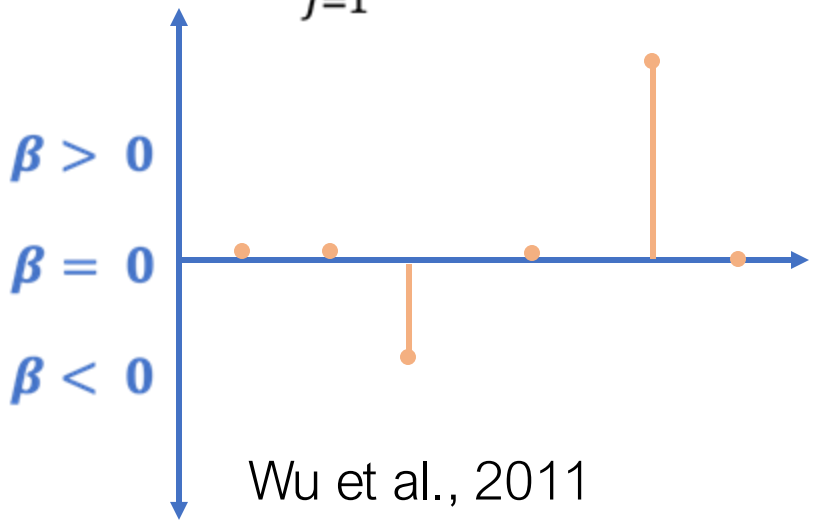
- $\boldsymbol{Z}$: the $N \times n$ design matrix indicating which cell is from which individual
- $\boldsymbol{X_{ind}}$: covariates for individuals, e.g, age, sex, ancestry PCs
- $\boldsymbol{X_{cell}}$: covariates for cells, e.g, pear factors
- $\boldsymbol{G}$: genotype
- $\boldsymbol{b_{ind}}$: $n \times 1$ vector for random effects
- $\boldsymbol{\Psi}$**:** Genetic relationship matrix for related samples; Identity matrix for unrelated samples

Set-based association tests: grouping rare variants to test

$$Q_{BURDEN} = \left(\sum_{j=1}^{q} w_j S_j\right)^2$$



$$Q_{SKAT} = \sum_{j=1}^{q} w_j^2 S_j^2$$



Wu et al., 2011

ACAT-V

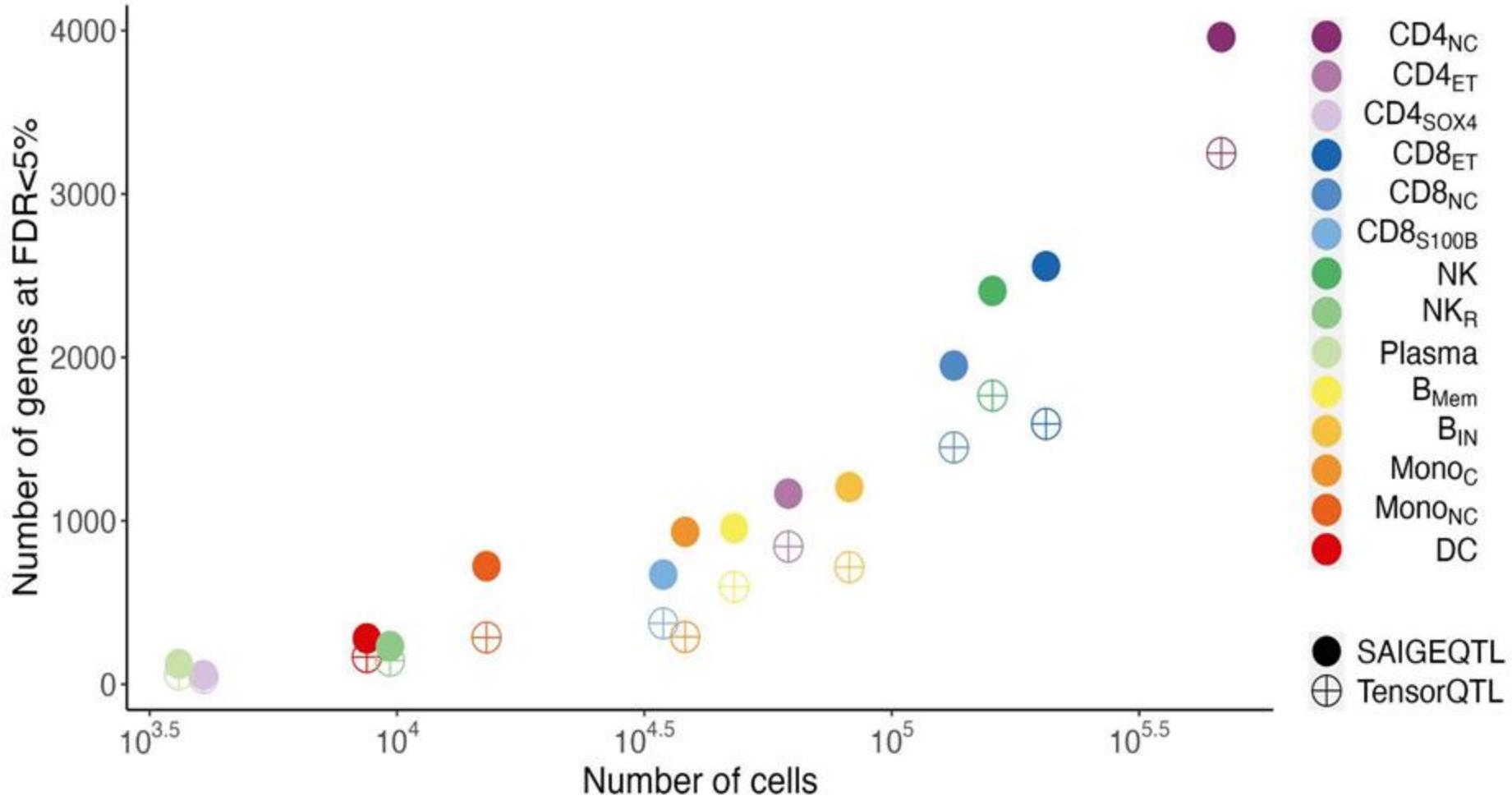
$$T_{ACAT-V} = \sum_{j=1}^{q} w_j^2 \tan\{(0.5 - p_i)\pi\}$$

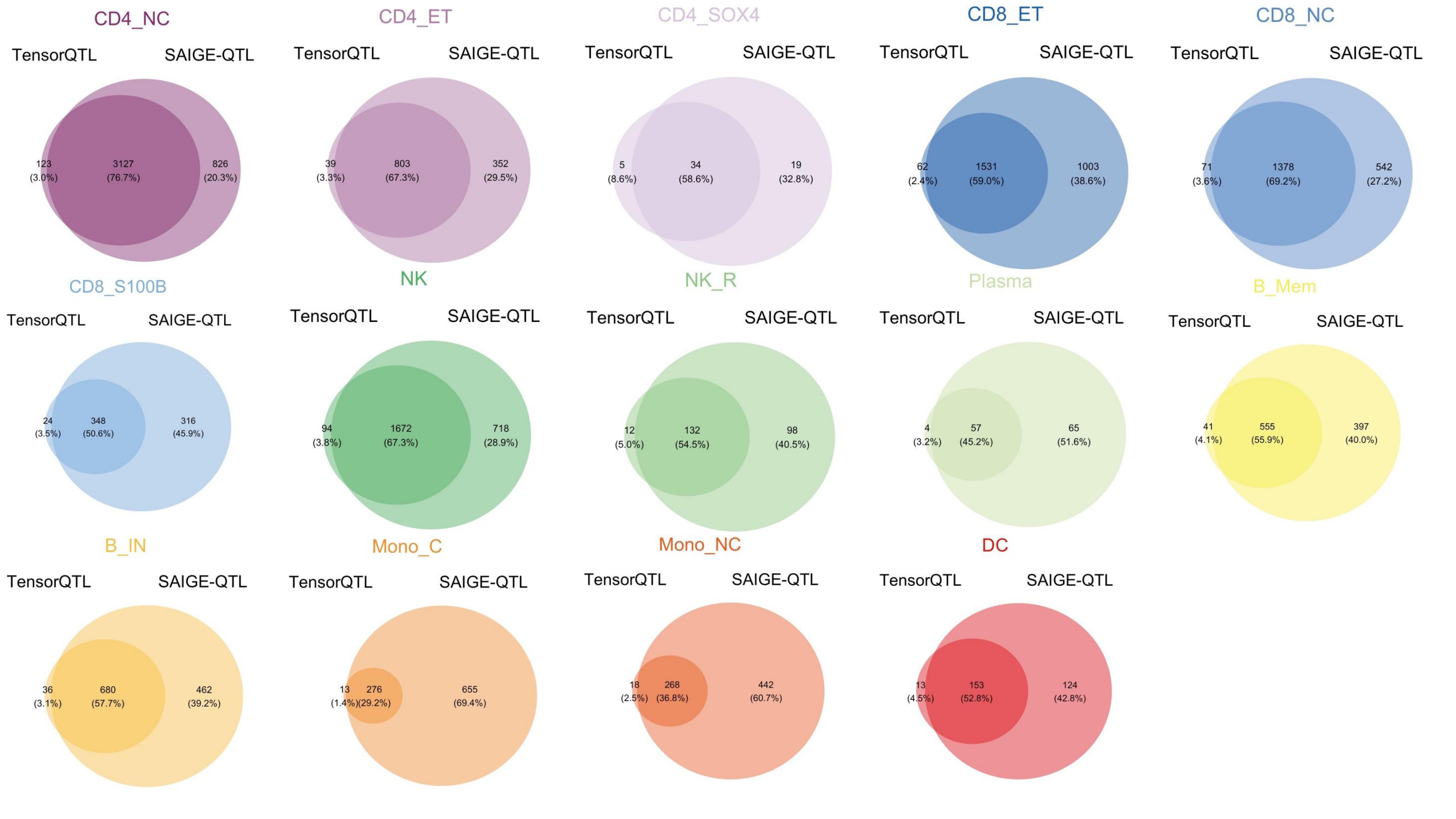
Liu et al., 2019

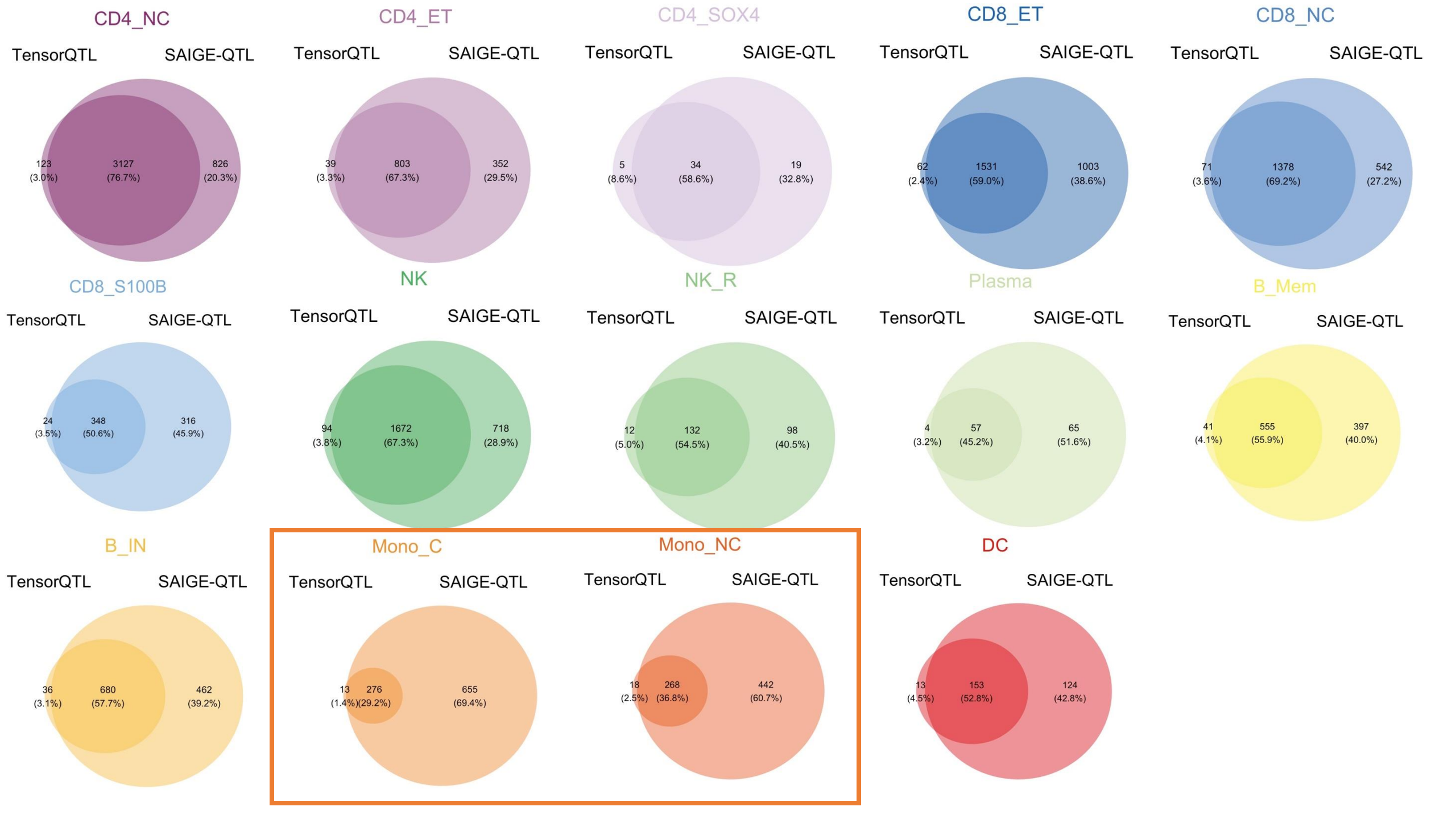
$S_j = \sum_{i=1}^{n} g_{ij}(y_i - \hat{\mu}_i)$ is the score statistic for the variant

Under the null $H_0: \beta_j = 0$

Number of eGenes increases as number of cells increases with both methods

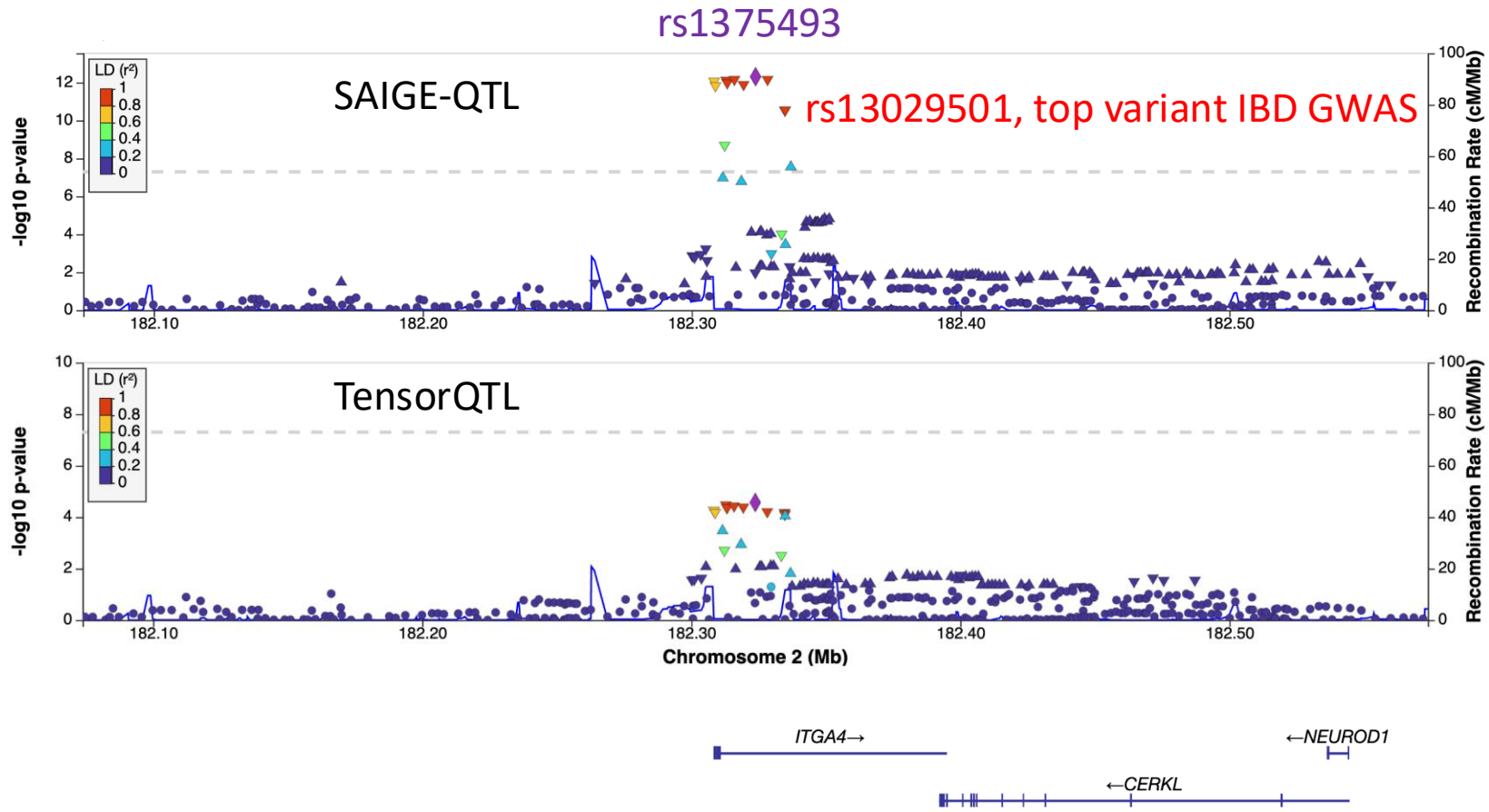



CD4_NC
TensorQTL
SAIGE-QTL
123
(3.0%)
3127
(76.7%)
826
(20.3%)
CD4_ET
TensorQTL
SAIGE-QTL
39
(3.3%)
803
(67.3%)
352
(29.5%)
CD4_SOX4
TensorQTL
SAIGE-QTL
5
(8.6%)
34
(58.6%)
19
(32.8%)
CD8_ET
TensorQTL
SAIGE-QTL
62
(2.4%)
1531
(59.0%)
1003
(38.6%)
CD8_NC
TensorQTL
SAIGE-QTL
71
(3.6%)
1378
(69.2%)
542
(27.2%)
CD8_S100B
NK
NK_R
Plasma
B_Mem
TensorQTL
SAIGE-QTL
24
(3.5%)
348
(50.6%)
316
(45.9%)
TensorQTL
SAIGE-QTL
94
(3.8%)
1672
(67.3%)
718
(28.9%)
TensorQTL
SAIGE-QTL
12
(5.0%)
132
(54.5%)
98
(40.5%)
TensorQTL
SAIGE-QTL
4
(3.2%)
57
(45.2%)
65
(51.6%)
TensorQTL
SAIGE-QTL
41
(4.1%)
555
(55.9%)
397
(40.0%)
B_IN
Mono_C
Mono_NC
DC
TensorQTL
SAIGE-QTL
36
(3.1%)
680
(57.7%)
462
(39.2%)
TensorQTL
SAIGE-QTL
13
(1.4%)
276
(29.2%)
655
(69.4%)
TensorQTL
SAIGE-QTL
18
(2.5%)
268
(36.8%)
442
(60.7%)
TensorQTL
SAIGE-QTL
13
(4.5%)
153
(52.8%)
124
(42.8%)


CD4_NC
TensorQTL
SAIGE-QTL
123
(3.0%)
3127
(76.7%)
826
(20.3%)
CD4_ET
TensorQTL
SAIGE-QTL
39
(3.3%)
803
(67.3%)
352
(29.5%)
CD4_SOX4
TensorQTL
SAIGE-QTL
5
(8.6%)
34
(58.6%)
19
(32.8%)
CD8_ET
TensorQTL
SAIGE-QTL
62
(2.4%)
1531
(59.0%)
1003
(38.6%)
CD8_NC
TensorQTL
SAIGE-QTL
71
(3.6%)
1378
(69.2%)
542
(27.2%)
CD8_S100B
TensorQTL
SAIGE-QTL
24
(3.5%)
348
(50.6%)
316
(45.9%)
NK
TensorQTL
SAIGE-QTL
94
(3.8%)
1672
(67.3%)
718
(28.9%)
NK_R
TensorQTL
SAIGE-QTL
12
(5.0%)
132
(54.5%)
98
(40.5%)
Plasma
TensorQTL
SAIGE-QTL
4
(3.2%)
57
(45.2%)
65
(51.6%)
B_Mem
TensorQTL
SAIGE-QTL
41
(4.1%)
555
(55.9%)
397
(40.0%)
B_IN
TensorQTL
SAIGE-QTL
36
(3.1%)
680
(57.7%)
462
(39.2%)
Mono_C
TensorQTL
SAIGE-QTL
13
(1.4%)
276
(29.2%)
655
(69.4%)
Mono_NC
TensorQTL
SAIGE-QTL
18
(2.5%)
268
(36.8%)
442
(60.7%)
DC
TensorQTL
SAIGE-QTL
13
(4.5%)
153
(52.8%)
124
(42.8%)

Using the same data, SAIGE-QTL identified the eQTL for the IBD-associated *ITGA4* locus that were missed in pseudo-bulk mapping

Mono C



IBD: inflammatory bowel disease

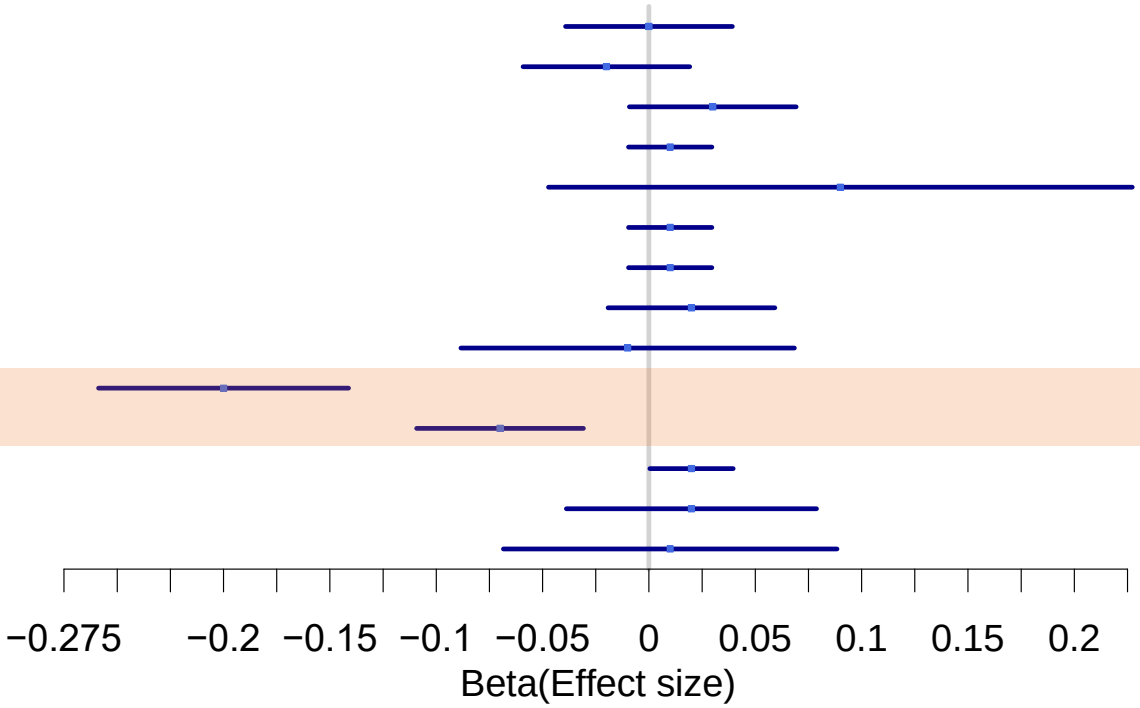
IBD GWAS: de Lange *et al*, Nature Genetics 2017

eQTL finempping: Kundu *et al,* Nature Genetics 2022

SAIGE-QTL identified the eQTL (only) in monocytes for the IBD-associated *ITGA4* locus (missed in pseudo-bulk mapping)

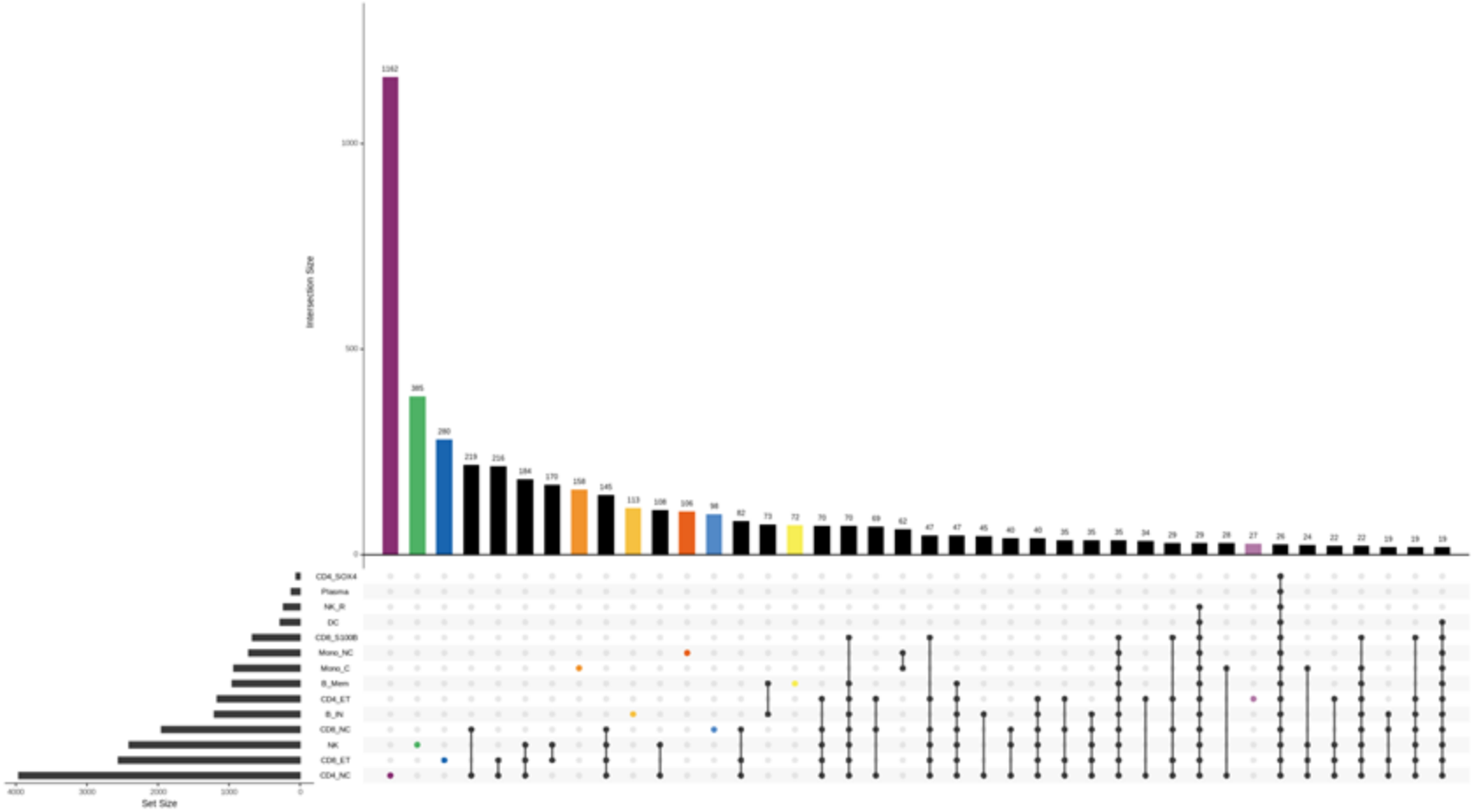
rs1375493

Cell type	Beta	SE	P−value
B_IN	0	0.02	9.2e−01
B_Mem	−0.02	0.02	5.3e−01
CD4_ET	0.03	0.02	6.7e−02
CD4_NC	0.01	0.01	4.7e−01
CD4_SOX4	0.09	0.07	2.0e−01
CD8_ET	0.01	0.01	5.4e−01
CD8_NC	0.01	0.01	5.4e−01
CD8_S100B	0.02	0.02	2.6e−01
DC	−0.01	0.04	8.3e−01
Mono_C	−0.2	0.03	4.9e−13
Mono_NC	−0.07	0.02	1.1e−03
NK	0.02	0.01	2.7e−01
NK_R	0.02	0.03	5.7e−01
Plasma	0.01	0.04	8.2e−01



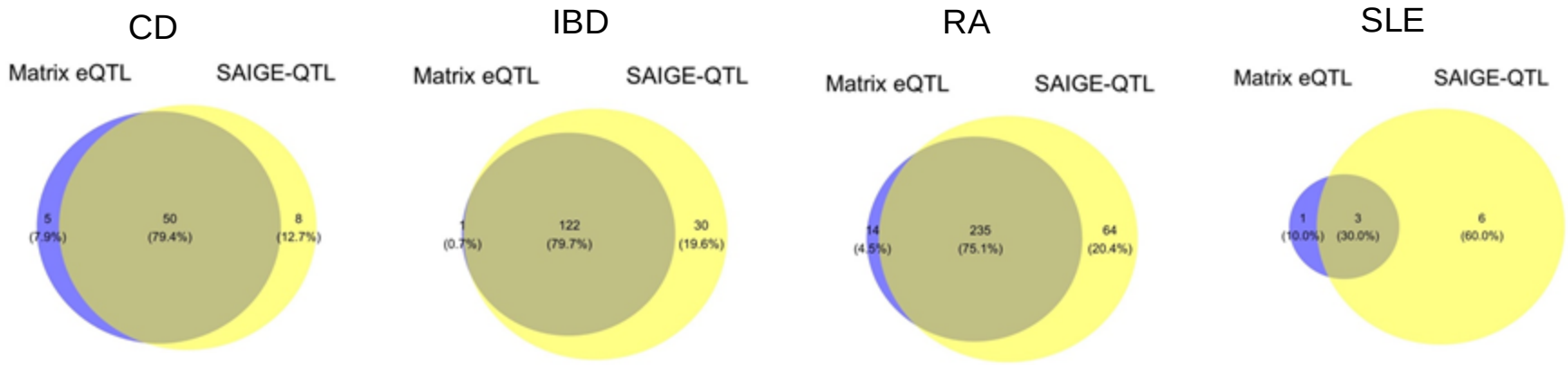
IBD: inflammatory bowel disease

Across all 14 cell types, SAIGE-QTL identified 17,218 eGenes across all cell types (FDR<5%, as described below and in Methods), representing 5,894 unique genes of which 2,447 (**42%**) were specific to one cell type only



Values >10 are included

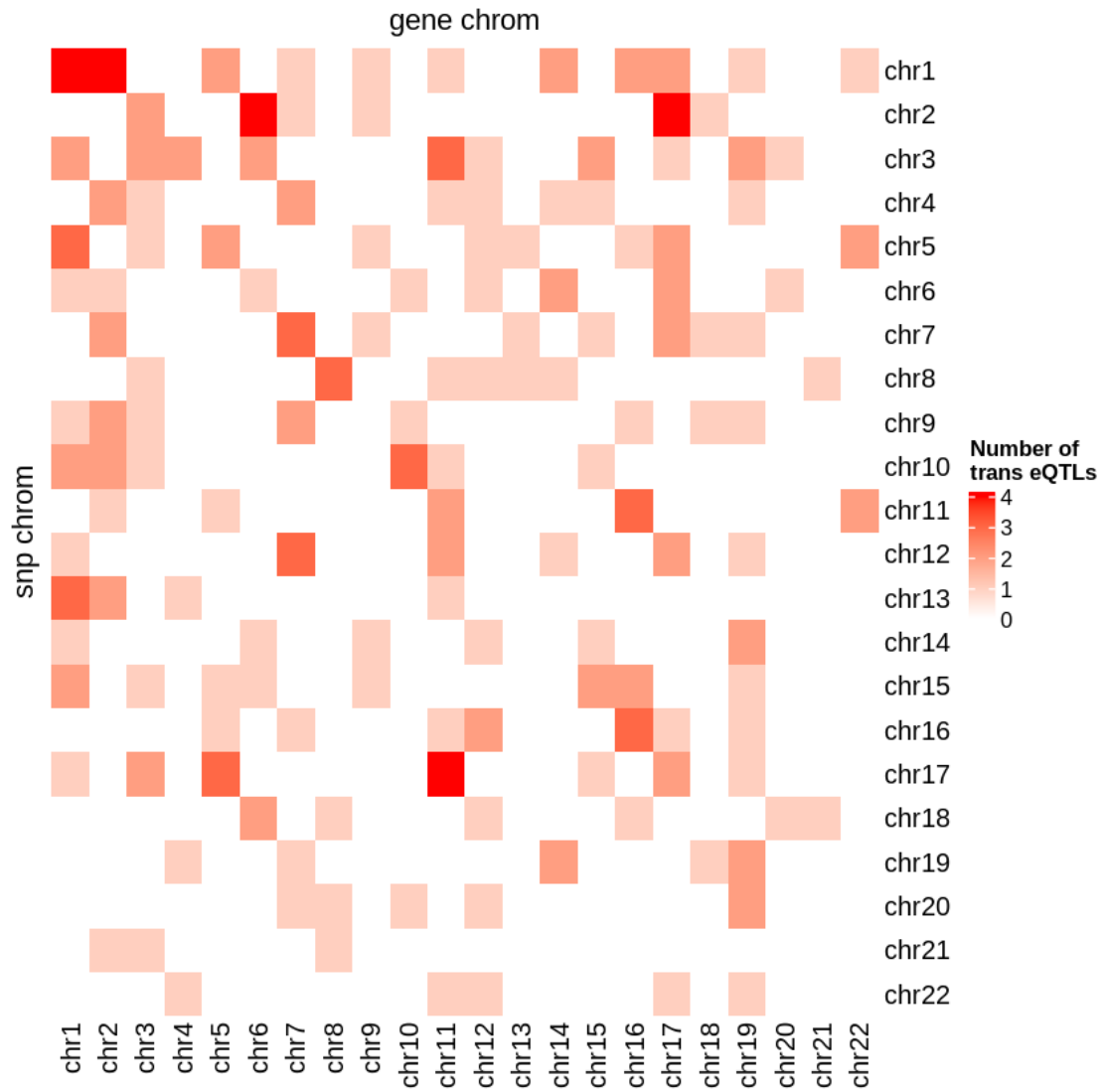
SAIGE-QTL identified additional immune disease-associated loci mediated by gene expression



- Among eGenes identified by SAIGE-QTL and TensorQTL across 14 cell types, gene-cell type pairs that are identified by the Summary-data–based Mendelian randomization (SMR, Zhu et al., 2016) framework to be associated with four autoimmune disease risks. CD: Crohn's disease, IBD: inflammatory bowel disease, RA: rheumatoid arthritis, SLE: systemic lupus erythematosus.

- e.g. 9q21 was shown to affect the risk of CD and IBD through changes in CARD9 expression in monocytes (MonoC)

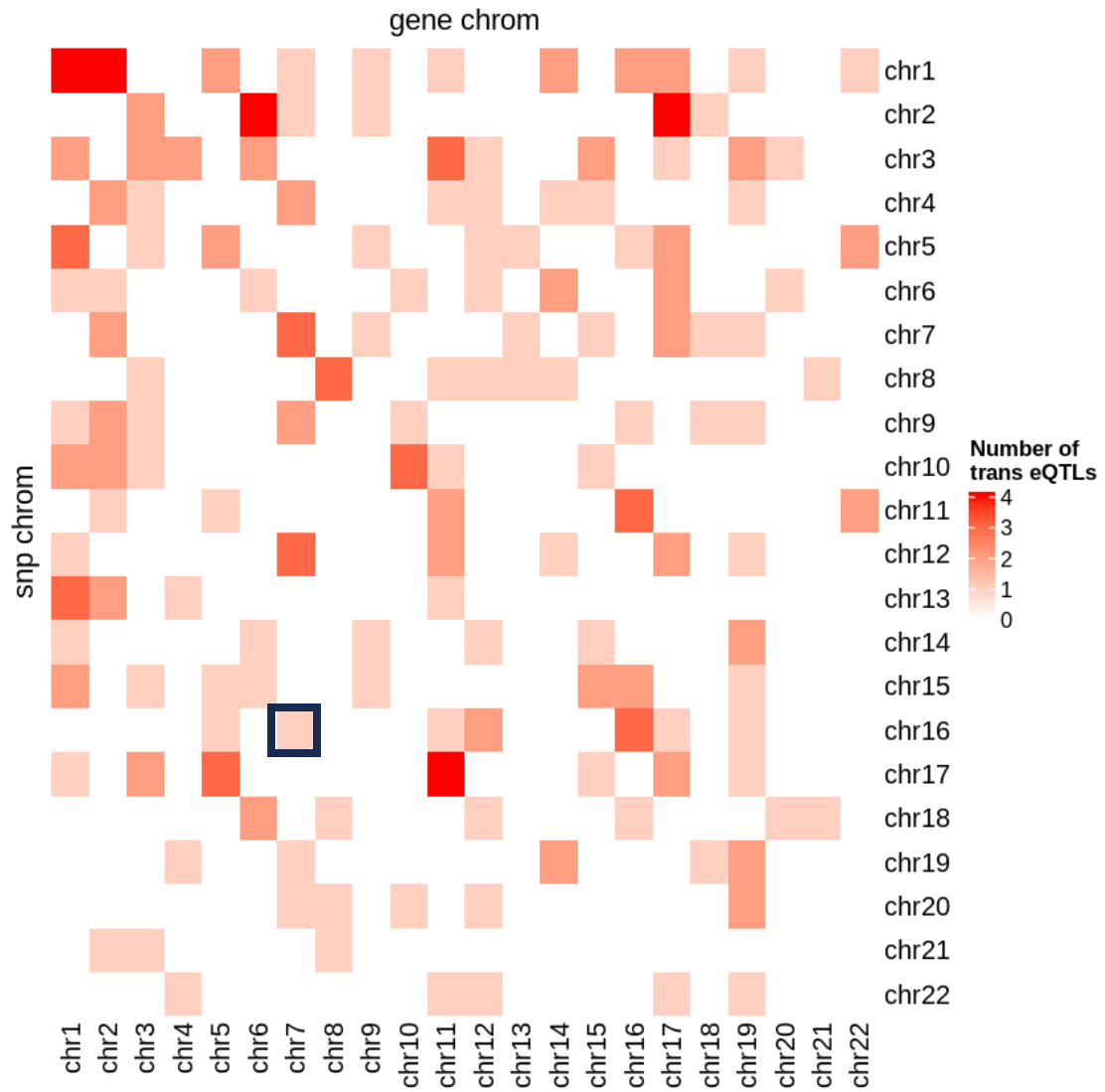
Trans-eQTL regions for all protein coding genes in Naive CD4^{+}T cells



Trans-region:

- MAF >= 10%
- Outside a 2Mb window around the gene

Trans-eQTL regions for all protein coding genes in Naive CD4^{+}T cells

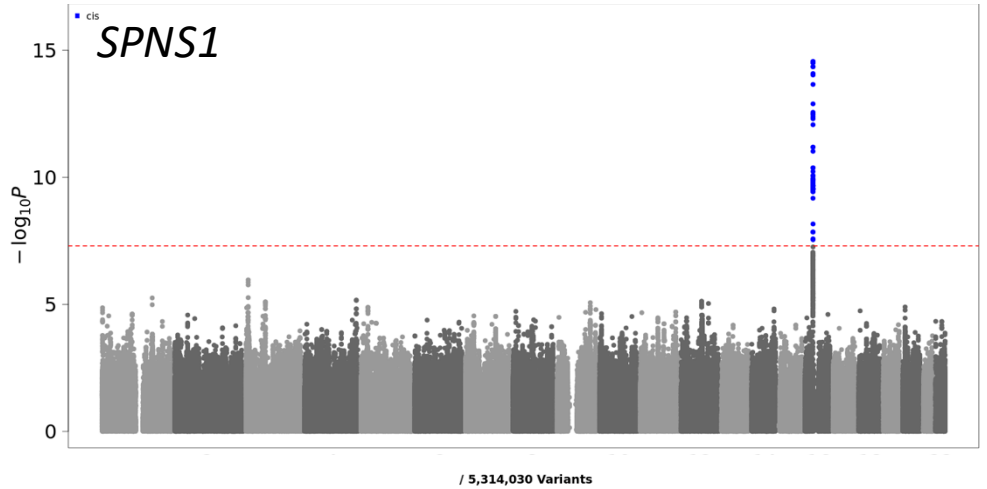


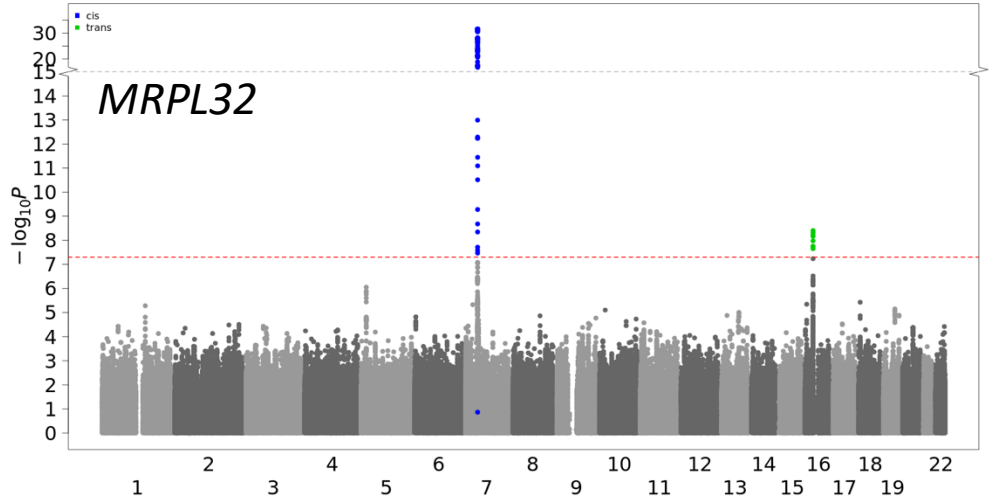
Trans-region:

- MAF >= 10%
- Outside a 2Mb window around the gene

SAIGE-QTL identified trans-eQTLs through genome-wide eQTL scan

CD4 NC





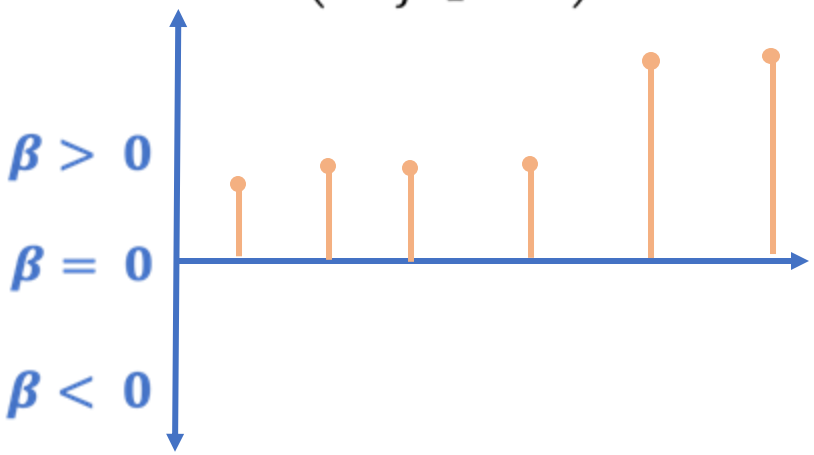
- cis
- trans

SPNS1 encodes SPNS lysolipid transporter 1, with potential role in maintaining **mitochondrial**-lysosomal homeostasis and preserving muscle mass and function (Zhang et al., 2023)

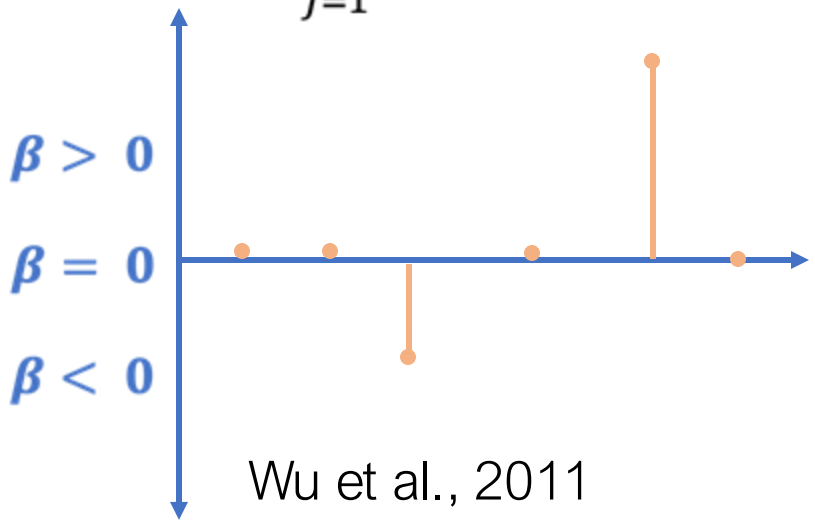
MRPL32 encodes Mitochondrial Ribosomal Protein L32. Among its related pathways are **Mitochondrial** translation and Metabolism of proteins (from GeneCards)

Set-based association tests: grouping rare variants to test

$$Q_{BURDEN} = \left(\sum_{j=1}^{q} w_j S_j\right)^2$$



$$Q_{SKAT} = \sum_{j=1}^{q} w_j^2 S_j^2$$



Wu et al., 2011

ACAT-V

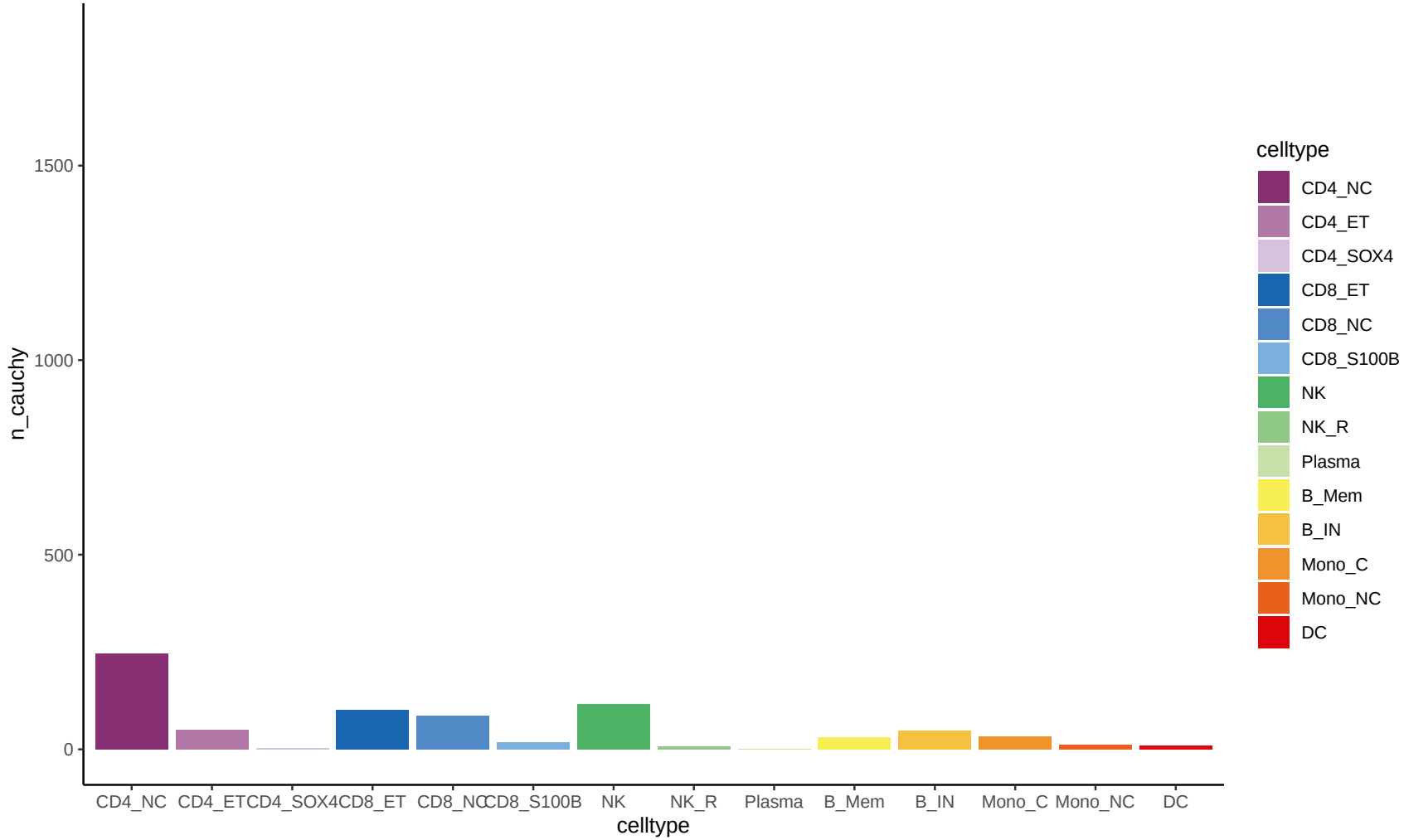
$$T_{ACAT-V} = \sum_{j=1}^{q} w_j^2 \tan\{(0.5 - p_i)\pi\}$$

Liu et al., 2019

$S_j = \sum_{i=1}^{n} g_{ij}(y_i - \hat{\mu}_i)$ is the score statistic for the variant

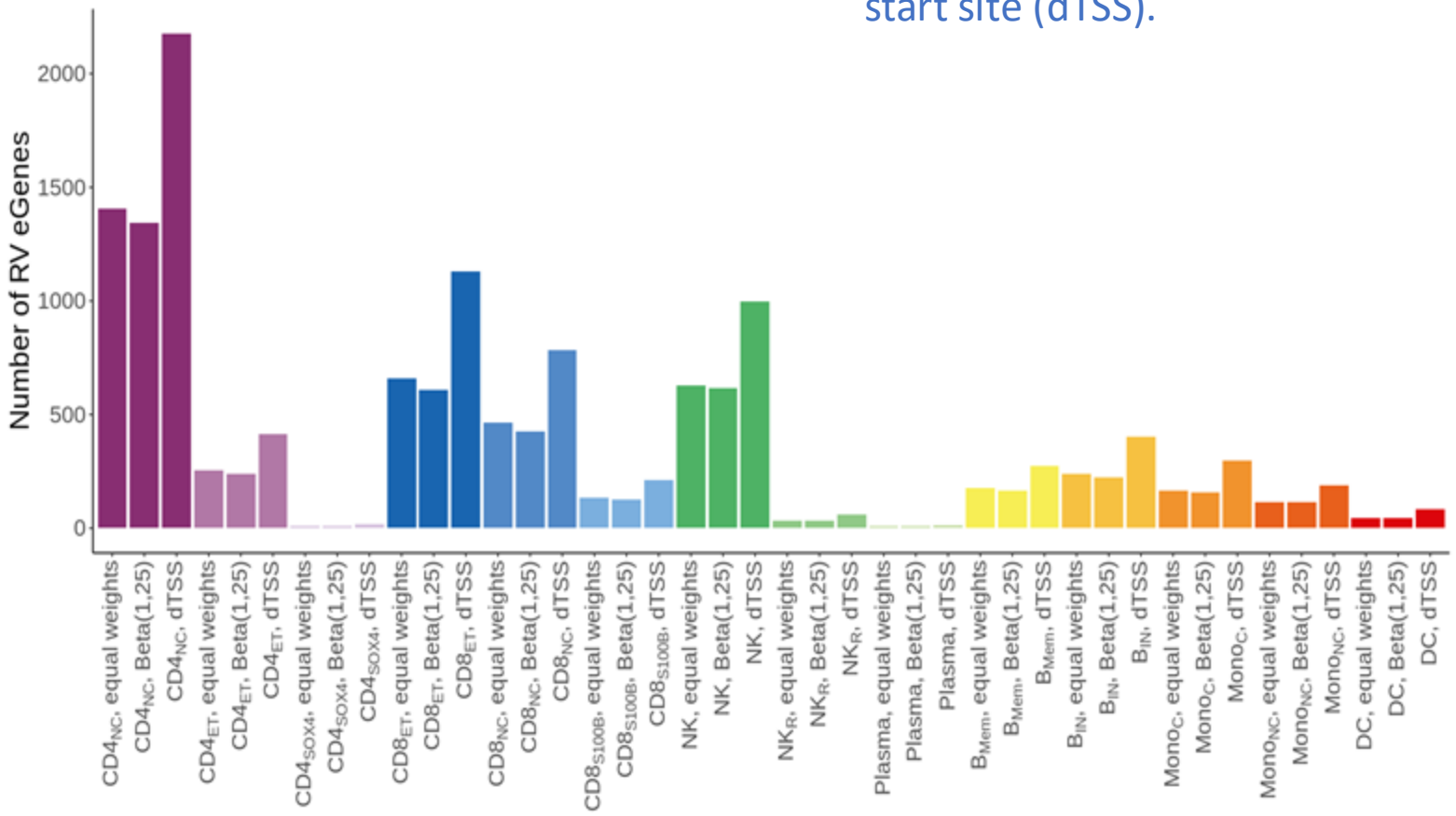
Under the null $H_0: \beta_j = 0$

5,541 eGenes (2,317 unique) eGenes have rare/less frequent signals (MAF <= 5%), 483 (21%) are independent from common eQTLs in the same genes



Different weighting strategies in the set-based tests across all cell types

- Equal weights
- Beta(MAF, 1,25) – upweight rarer variants
- based on the distance of each variant from the transcription start site (dTSS).



Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 376, NO. 6589 > SINGLE-CELL RNA-SEQ REVEALS CELL TYPE–SPECIFIC MOLECULAR AND GENETIC ASSOCIATIONS TO LUPUS

RESEARCH ARTICLE | IMMUNOGENOMICS

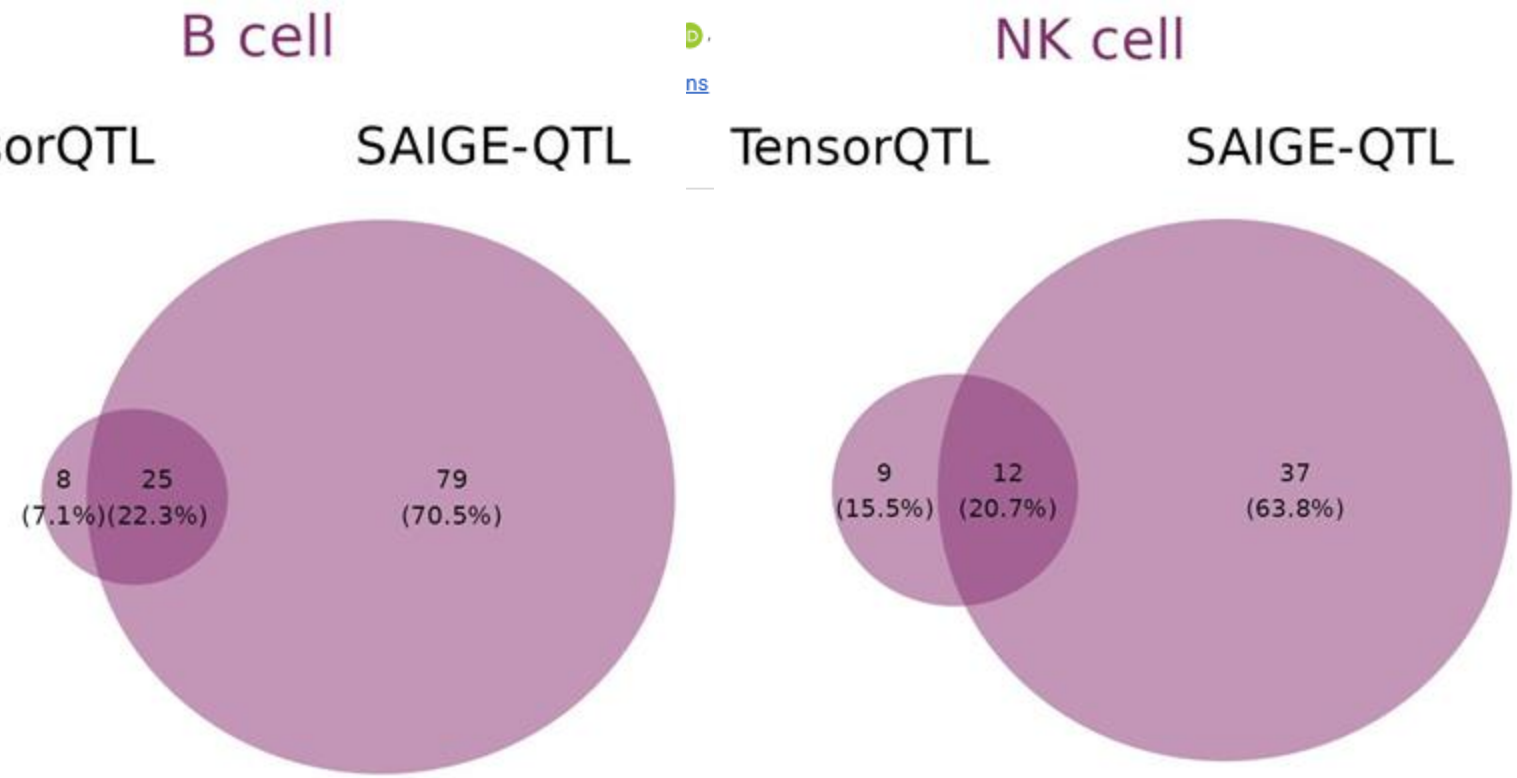
Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus

RICHARD K. PEREZ, M. GR

ANTON OGORODNIKOV, RAYMI

SCIENCE · 8 Apr 2022 · Vol

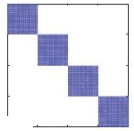
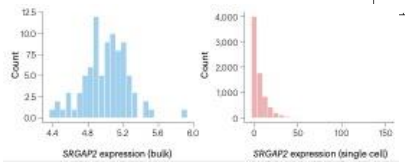
ns

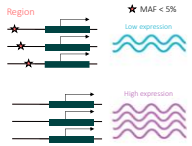


Linfeng Hu

95 out of 104 genes are eGenes in OneK1K (B_IN+B_Mem)

SAIGE-QTL: Generalized Linear Poisson Mixed-model for sc-eQTL mapping

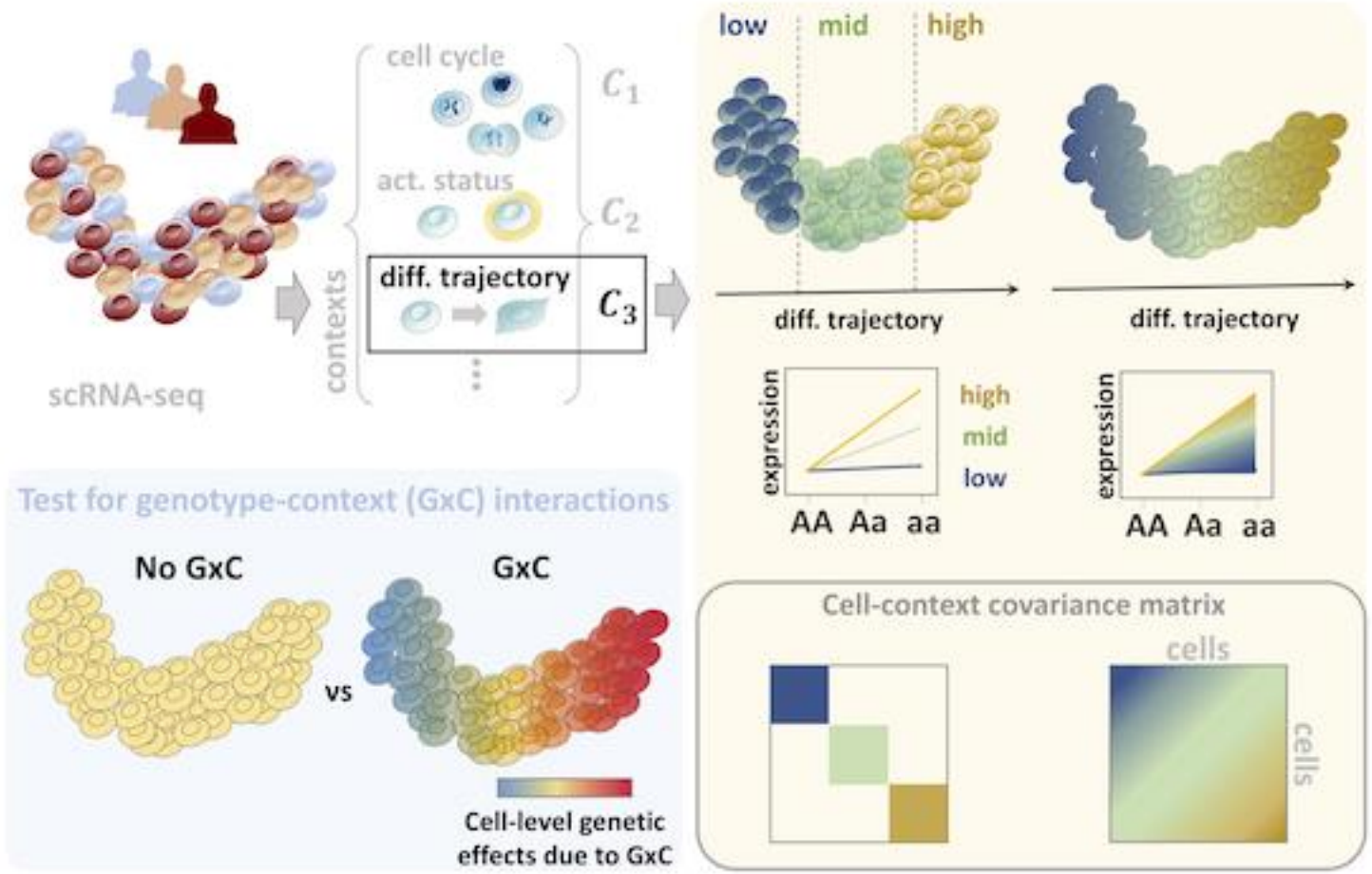
1. Models repeat and complex data structure 
2. Models discrete read counts 

3. Fast and scalable for large data 
4. Tests common and rare variations 

Code and Data Availability

SAIGE-QTL is implemented as an open-source R method available at:
https://github.com/weizhou0/qtl
Tutorial:
https://weizhou0.github.io/SAIGE-QTL-doc/
Code to reproduce all analyses and figures included here can be found at:
https://github.com/annacuomo/SAIGE_QTL_analyses
Summary statistics for the eQTLs in this study can be found on Zenodo:
https://zenodo.org/records/10811106
Pre-print
https://www.medrxiv.org/content/10.1101/2024.05.15.24307317v1.full.pdf

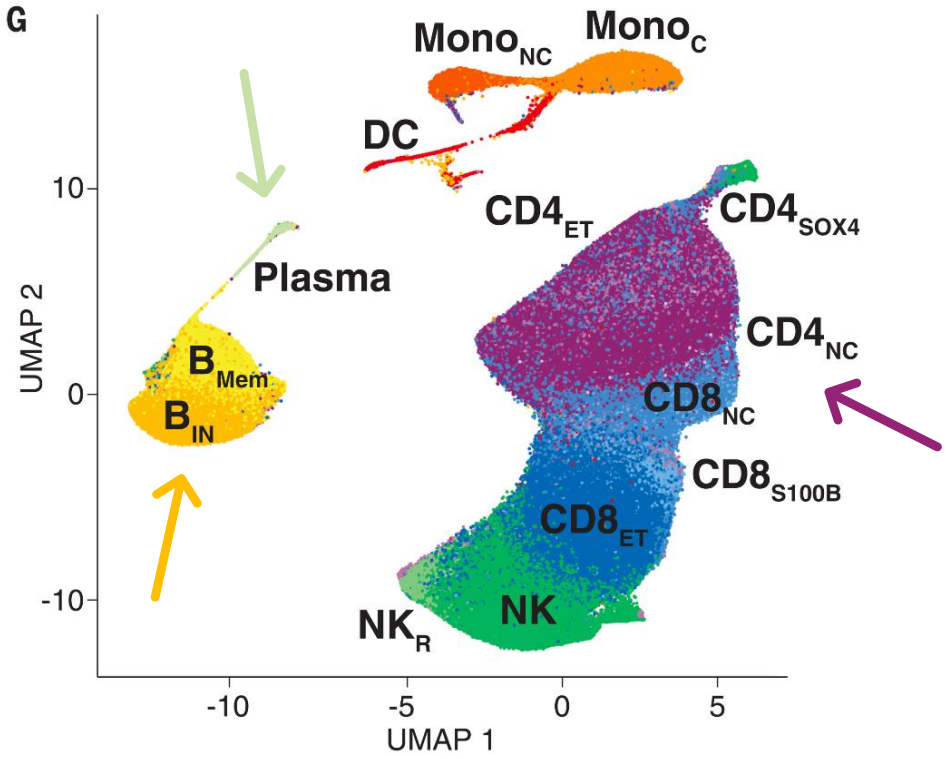
scRNA-seq data allow for identifying cell-level dynamic eQTLs

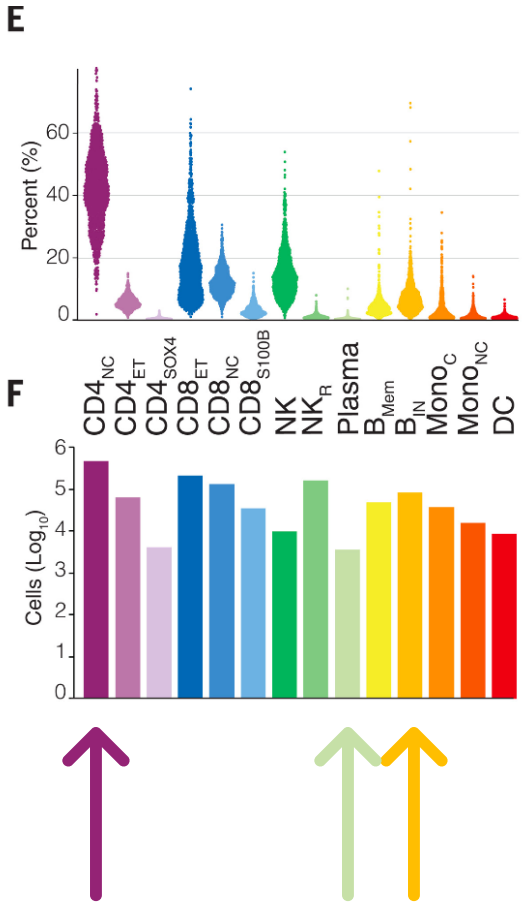


- CellRegMap. Cuomo et al., Mol Syst Biol (2022)
- GASPACHO. Kumasaka et al., Nature Genetics (2023)

More slides

OneK1K data, focus on three representative cell types





Focus on three cell types:

1. Most abundant: **CD4 NC** (n=463,528)
2. Middle: **Naïve B cells** (n=82,068)
3. Least abundant: **Plasma cells** (n=3,625)

Computation cost in OneK1K

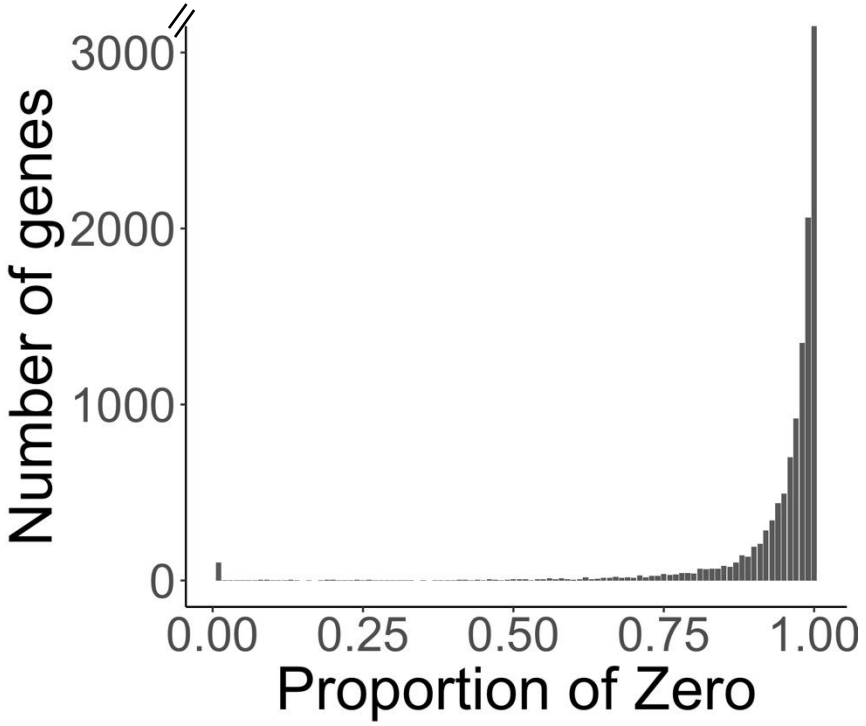
			Step 1 for 20,000 genes		Step 2 for 20,000 genes	
cell type	Individuals(n)	cells (N)	Time (CPU hours)	Memory (Gb)	Time (CPU hours) for cis (+- 1Mb)	Time (CPU hours) for genome-wide scan for 8M markers
Plasma	795	3,625	24.99	0.23	46	604
CD4_SOX4	857	4,065	25.78	0.23	47	
DC	968	8,690	35.88	0.24	46	
NK_R	969	9,677	37.56	0.25	50	
Mono_NC	969	15,166	49.72	0.26	47	
CD8_S100B	981	34,528	94.14	0.34	46	
Mono_C	969	38,233	106.79	0.35	49	
B_Mem	982	48,023	145.98	0.38	43	
CD4_ET	982	61,786	157.67	0.45	48	
B_IN	982	82,068	290.98	0.53	47	608
CD8_NC	982	133,482	337.27	0.73	50	
NK	982	159,820	425.84	0.80	54	
CD8_ET	982	205,077	596.32	0.97	50	
CD4_NC	982	463,528	1424.79	1.88	70	846

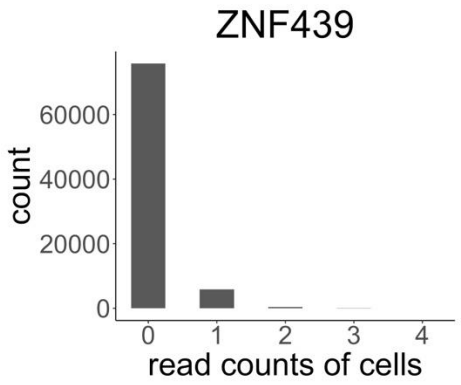
Compared to the R function glmer used in *Nathan et al.* (2022) paper

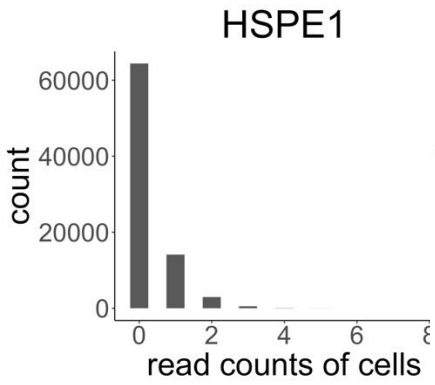
- 145 times faster for *cis*-eQTL mapping, and even more for *trans* mapping
- Allows for rare-variant association tests
- Allows for accounting for sample relatedness via the genetic relationship matrix
- Allows for accounting for cell-cell correlation with user-specified covariance matrices

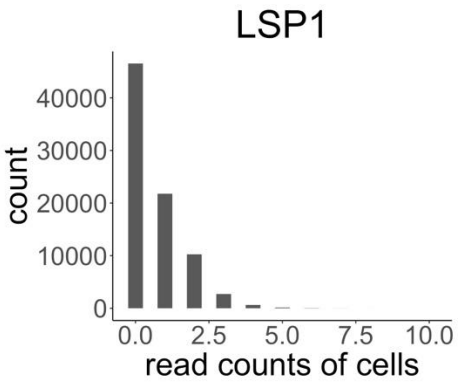
Genes can with different sparsity levels (proportion of zero read counts)

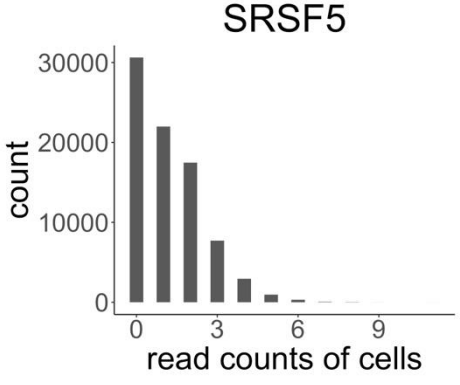
- Example: Naïve B cells

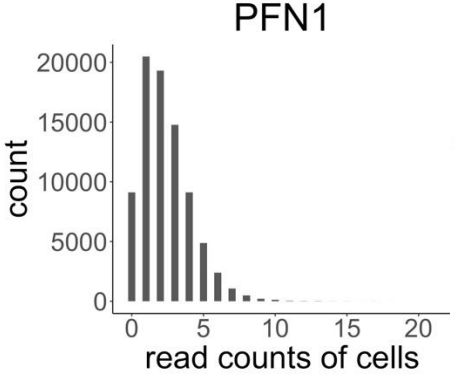


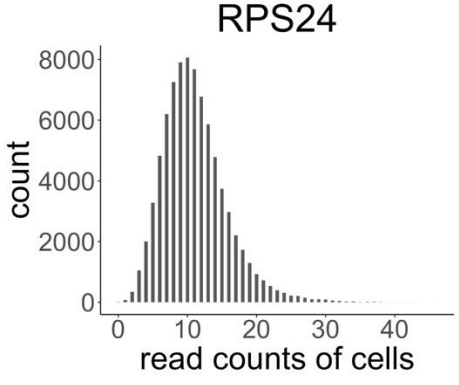






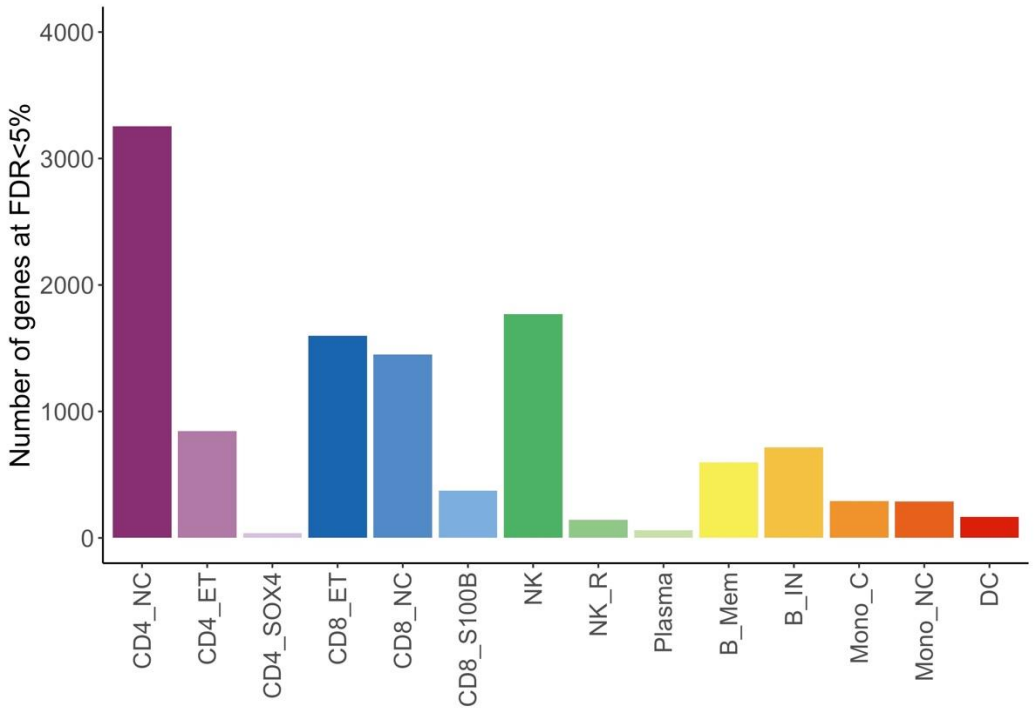






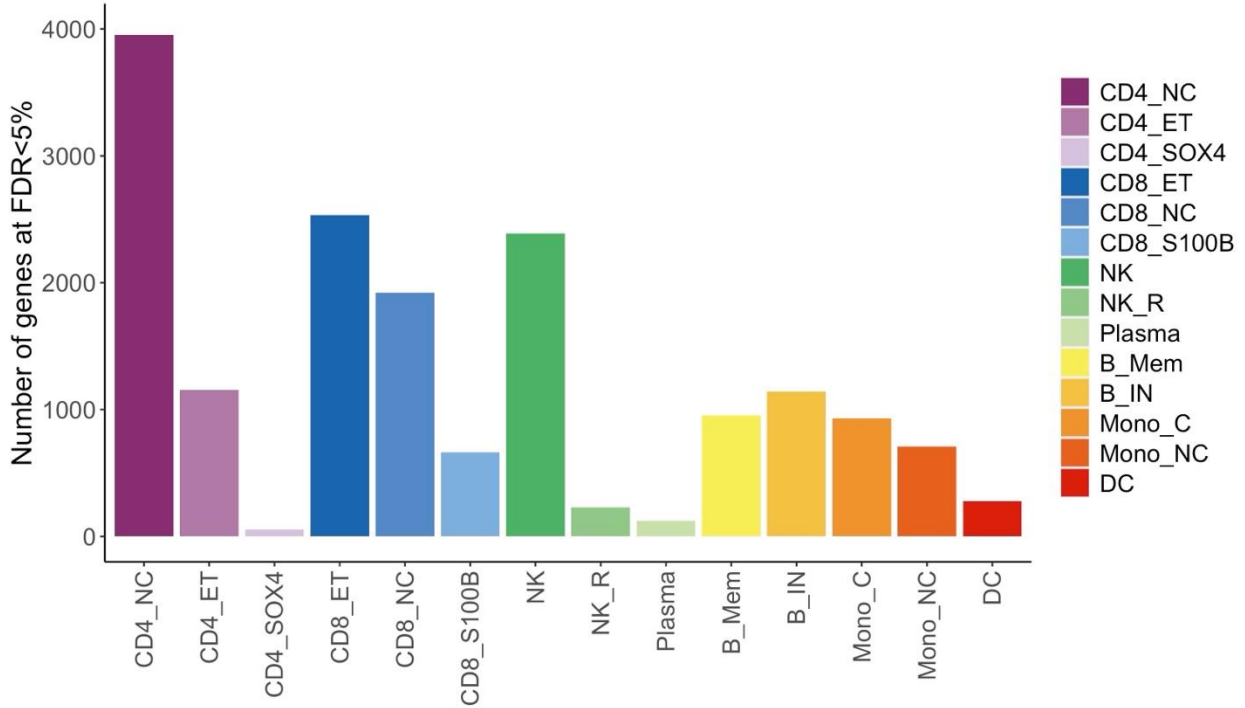
SAIGE-QTL identified ~48% more eGenes than previous pseudo-bulk analysis using TensorQTL (FDR<5%)

TensorQTL (linear regression)



Total: 11,585

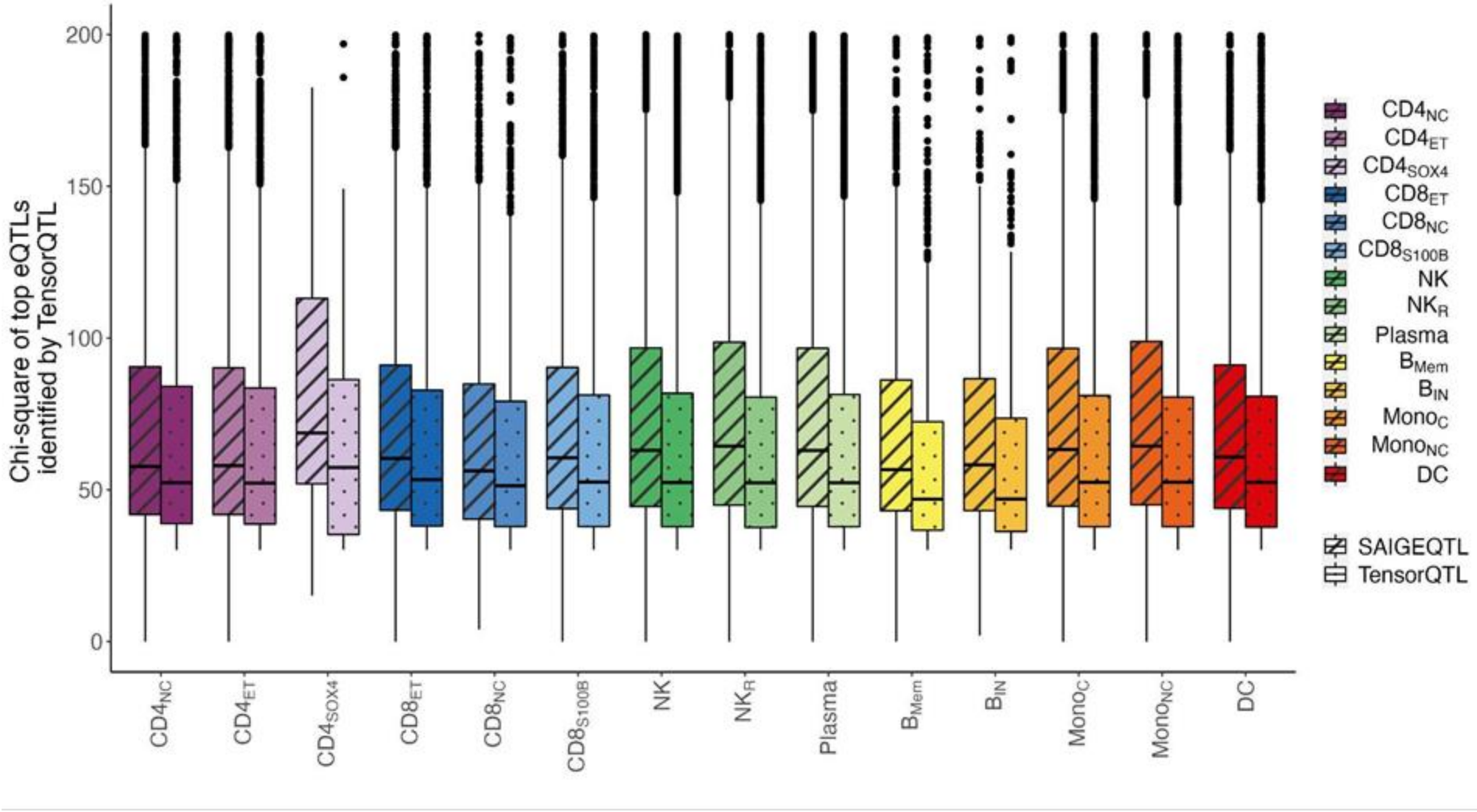
SAIGE-QTL (Poisson mixed model)



Total: **17,033**

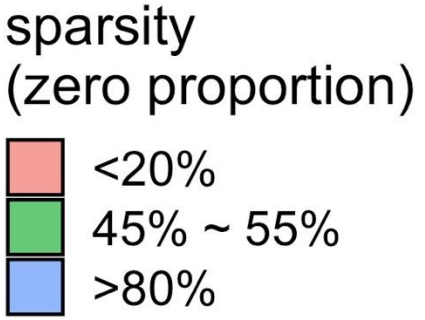
* Variants with MAF > 5% were tested
* Two analyses have highly correlated effect size estimates of cis-eQTLs (Pearson's $R^2 = 0.98$) in naive B cells

SAIGE-QTL has improved power with higher chi-square statistics



Simulation studies to evaluate the calibration of SAIGE-QTL based on OneK1K data

- Permuted genotypes across donors for genetic variants with MAF > 5% on Chr 1 for 20 times
- Randomly selected 50 genes from each sparsity group from three cell types
- Run single-variant assoc tests for Chr 1 using SAIGE-QTL for each gene



SAIGE-QTL has well calibrated false positive rates