eQTL mapping in single cells: New data, new models

Wei Zhou, Ph.D.

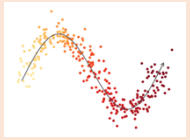
Center for Genomic Medicine, Massachusetts General Hospital
Stanley Center of Psychiatric Research, Broad Institute

The 2026 International Statistical Genetics Workshop

Single-cell RNA-seq (scRNA-seq) captures gene expression variations across cells

Cells vary in different aspects (cell context):

- Cell type
- Cell state
  - ➢ under stimulations/treatments
- Cell cycle phase
  - ➢ Position in division cycle (G1, S, G2, M)
- Developmental stage
- Metabolic state

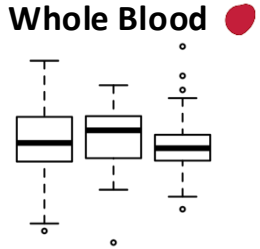


scRNA-seq data

			Gene1	Gene2	Gene3	Gene4	Gene5
Donor 1	Cell 1		0	2	4	0	3
	Cell 2		0	1	4	0	2
	Cell 3		0	2	4	1	3
Donor 2	Cell 4		0	1	3	0	1
	Cell 5		0	5	2	0	3
	Cell 6		0	5	2	1	4
	Cell 7		1	4	2	0	3
	Cell 8		0	2	3	0	5
Donor 3	Cell 9		0	6	0	0	8
	Cell 10		0	1	2	1	5
Donor 4	Cell 11		0	1	1	0	2
	Cell 12		0	6	5	1	7
	Cell 13		0	7	6	0	8
Donor 5	Cell 14		0	3	4	0	3
	Cell 15		0	3	5	1	3

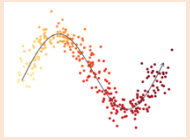
Read count reflecting gene expression level in each cell

Single-cell RNA-seq (scRNA-seq) captures gene expression variations across cells



Cells vary in different aspects (cell context):

- Cell type
- Cell state
  - ➢ under stimulations/treatments
- Cell cycle phase
  - ➢ Position in division cycle (G1, S, G2, M)
- Developmental stage
- Metabolic state



scRNA-seq data

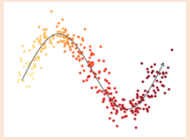
			Gene1	Gene2	Gene3	Gene4	Gene5
Donor 1	Cell 1		0	2	4	0	3
	Cell 2		0	1	4	0	2
	Cell 3		0	2	4	1	3
Donor 2	Cell 4		0	1	3	0	1
	Cell 5		0	5	2	0	3
	Cell 6		0	5	2	1	4
	Cell 7		1	4	2	0	3
	Cell 8		0	2	3	0	5
Donor 3	Cell 9		0	6	0	0	8
	Cell 10		0	1	2	1	5
Donor 4	Cell 11		0	1	1	0	2
	Cell 12		0	6	5	1	7
	Cell 13		0	7	6	0	8
Donor 5	Cell 14		0	3	4	0	3
	Cell 15		0	3	5	1	3

Read count reflecting gene expression level in each cell

Single-cell RNA-seq (scRNA-seq) captures gene expression variations across cells

Cells vary in different aspects (cell context):

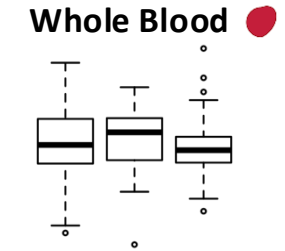
- **Cell type**
- Cell state
  - ➢ under stimulations/treatments
- Cell cycle phase
  - ➢ Position in division cycle (G1, S, G2, M)
- Developmental stage
- Metabolic state



scRNA-seq data

			Gene1	Gene2	Gene3	Gene4	Gene5
Donor 1	Cell 1		0	2	4	0	3
	Cell 2		0	1	4	0	2
	Cell 3		0	2	4	1	3
Donor 2	Cell 4		0	1	3	0	1
	Cell 5		0	5	2	0	3
	Cell 6		0	5	2	1	4
	Cell 7		1	4	2	0	3
	Cell 8		0	2	3	0	5
Donor 3	Cell 9		0	6	0	0	8
	Cell 10		0	1	2	1	5
Donor 4	Cell 11		0	1	1	0	2
	Cell 12		0	6	5	1	7
	Cell 13		0	7	6	0	8
Donor 5	Cell 14		0	3	4	0	3
	Cell 15		0	3	5	1	3

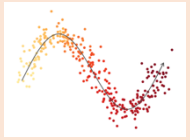
Read count reflecting gene expression level in each cell



Single-cell eQTL mapping reveals genetic effects hidden in bulk

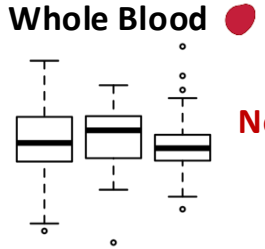
Cells vary in different aspects (cell context):

- **Cell type**
- Cell state
  - ➢ under stimulations/treatments
- Cell cycle phase
  - ➢ Position in division cycle (G1, S, G2, M)
- Developmental stage
- Metabolic state



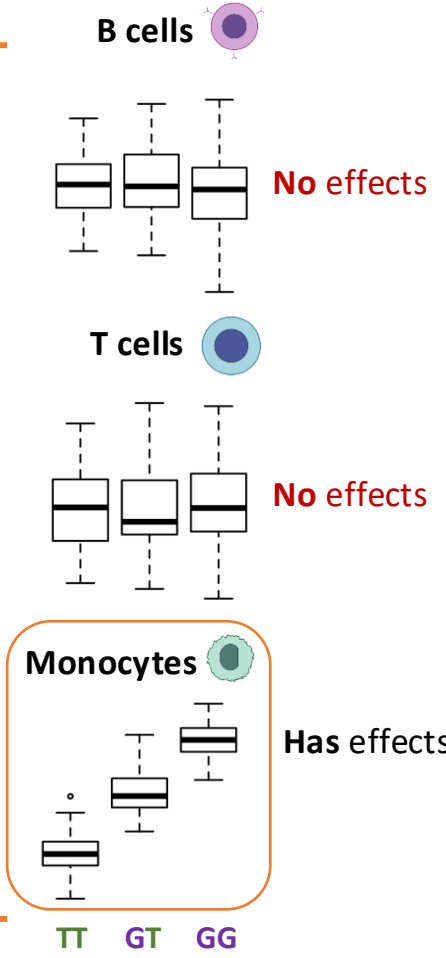
Bulk eQTL mapping uses averaged expression across all cells

-> masks cell-specific effects



Single-cell eQTL mapping uses the expression from individual cells

-> Uncovers cell-specific effects



scRNA-seq data

			Gene1	Gene2	Gene3	Gene4	Gene5
Donor 1	Cell 1		0	2	4	0	3
	Cell 2		0	1	4	0	2
	Cell 3		0	2	4	1	3
Donor 2	Cell 4		0	1	3	0	1
	Cell 5		0	5	2	0	3
	Cell 6		0	5	2	1	4
	Cell 7		1	4	2	0	3
	Cell 8		0	2	3	0	5
Donor 3	Cell 9		0	6	0	0	8
	Cell 10		0	1	2	1	5
Donor 4	Cell 11		0	1	1	0	2
	Cell 12		0	6	5	1	7
	Cell 13		0	7	6	0	8
Donor 5	Cell 14		0	3	4	0	3
	Cell 15		0	3	5	1	3

Read count reflecting gene expression level in each cell

Single-cell data are typically aggregated to pseudobulk for eQTL mapping

		Gene1	Gene2	Gene3	Gene4	Gene5
Donor 1	Cell 1	0	2	4	0	3
	Cell 2	0	1	4	0	2
	Cell 3	0	2	4	1	3
Donor 2	Cell 4	0	1	3	0	1
	Cell 5	0	5	2	0	3
	Cell 6	0	5	2	1	4
	Cell 7	1	4	2	0	3
	Cell 8	0	2	3	0	5
Donor 3	Cell 9	0	6	0	0	8
	Cell 10	0	1	2	1	5
Donor 4	Cell 11	0	1	1	0	2
	Cell 12	0	6	5	1	7
	Cell 13	0	7	6	0	8
Donor 5	Cell 14	0	3	4	0	3
	Cell 15	0	3	5	1	3

Pseudobulk

Mean
Sum
Median

	Gene1	Gene2	Gene3	Gene4	Gene5
Donor 1					
Donor 2					
Donor 3					
Donor 4					
Donor 5					

- Apply the standard bulk linear model:
  $Y = \beta_0 + \beta_1 G + \beta_2 C + \varepsilon$
- Computationally efficient

Pseudobulk approach loses cell-level resolution

- Does not model the distribution of gene expression across cells: one summary value may not accurately reflect the true expression level
- Does not model cell-level covariates: cell state, cycle phase, or activation level cannot be included
- Does not account for cell number imbalance: donors with very few cells contribute equally to donors with thousands

To go further, we need models that work directly on individual cells
- handle count data, account for cells sharing a donor, and use cell-level information

Poisson regression: the right likelihood for count data

$$Y_{ij} \sim Poi(\mu_{ij}); \quad \log(\mu_{ij}) = \beta_0 + \beta_1 G_i + \beta_2 C_{ij} + \text{offset}$$

μ_{ij}
expected counts
for cell j of donor i

Key term
β_1 G
genotype effect size
G = 0, 1, or 2 (allele count)

$\beta_2 C_{ij}$
covariates *(sex, age, PCs, PEER factors)*

offset
log(library size)
normalises sequencing depth

The log-link function

Instead of modelling counts directly, we model their logarithm. This keeps outcome values strictly positive and gives a multiplicative interpretation: each additional minor allele multiplies expected expression by e^{β_1}. The same logic applies to all covariates.

Interpretation of β_1

A positive β_1 means each minor allele increases expected counts multiplicatively. e^{β_1} is the fold-change in expression per allele, which is directly interpretable as an effect size.

The correlation problem: multiple cells from the same donor

- Cells from the same donor share genetics, environment, and batch. **They are not independent**
- Treating each cell as an independent observation, inflating the sample size and **producing false positives**
- Statistical test: the true N is closer to the number of donors, not total cells

Solution:
a mixed model with fixed effects capture the genotype signal
random effects accounts for the donor-level correlation

Poisson mixed model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 G_i + \beta_2 C_{ij} + \text{offset} + b_i$$

μ_{ij}
expected counts
for cell j of donor i

Key term
$\beta_1 G$
genotype effect size
G = 0, 1, or 2 (allele count)

$\beta_2 C_{ij}$
covariates *(sex, age, PCs, PEER factors)*

offset
log(library size)
normalises sequencing depth

bi ~ *N*(0, σΣ)
random donor effect
absorbs donor correlation

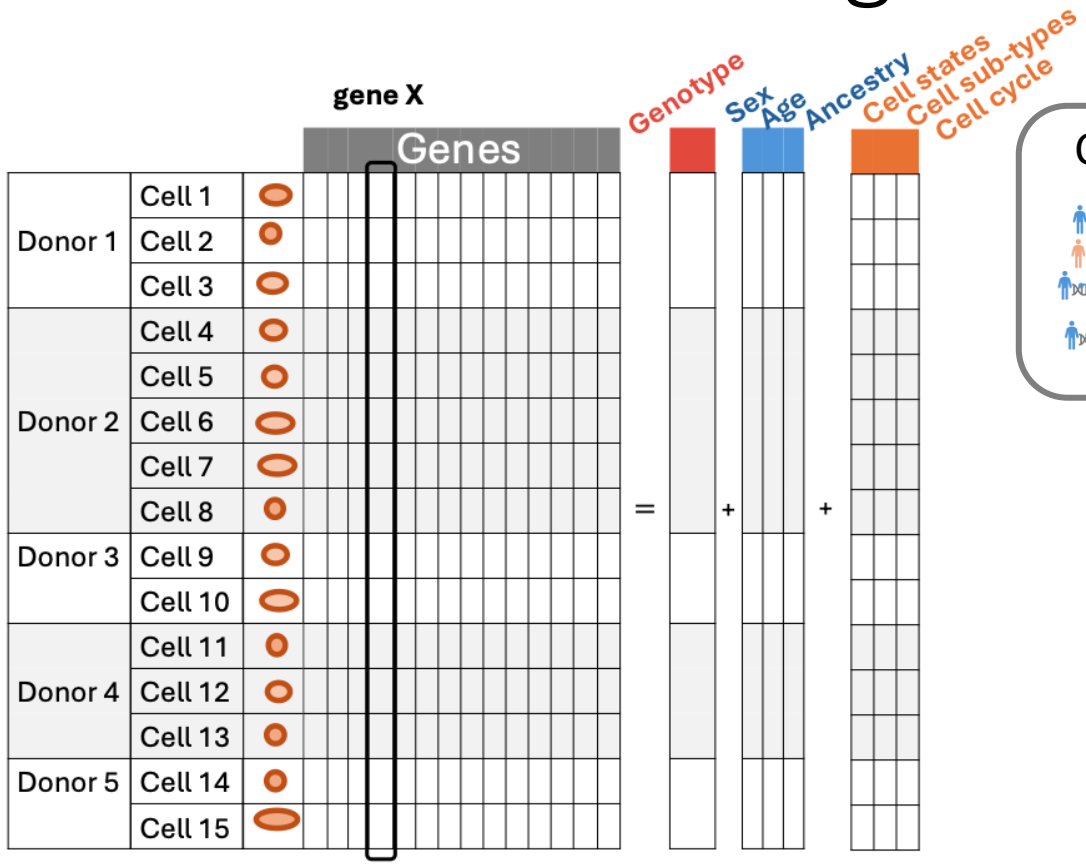
Σ =

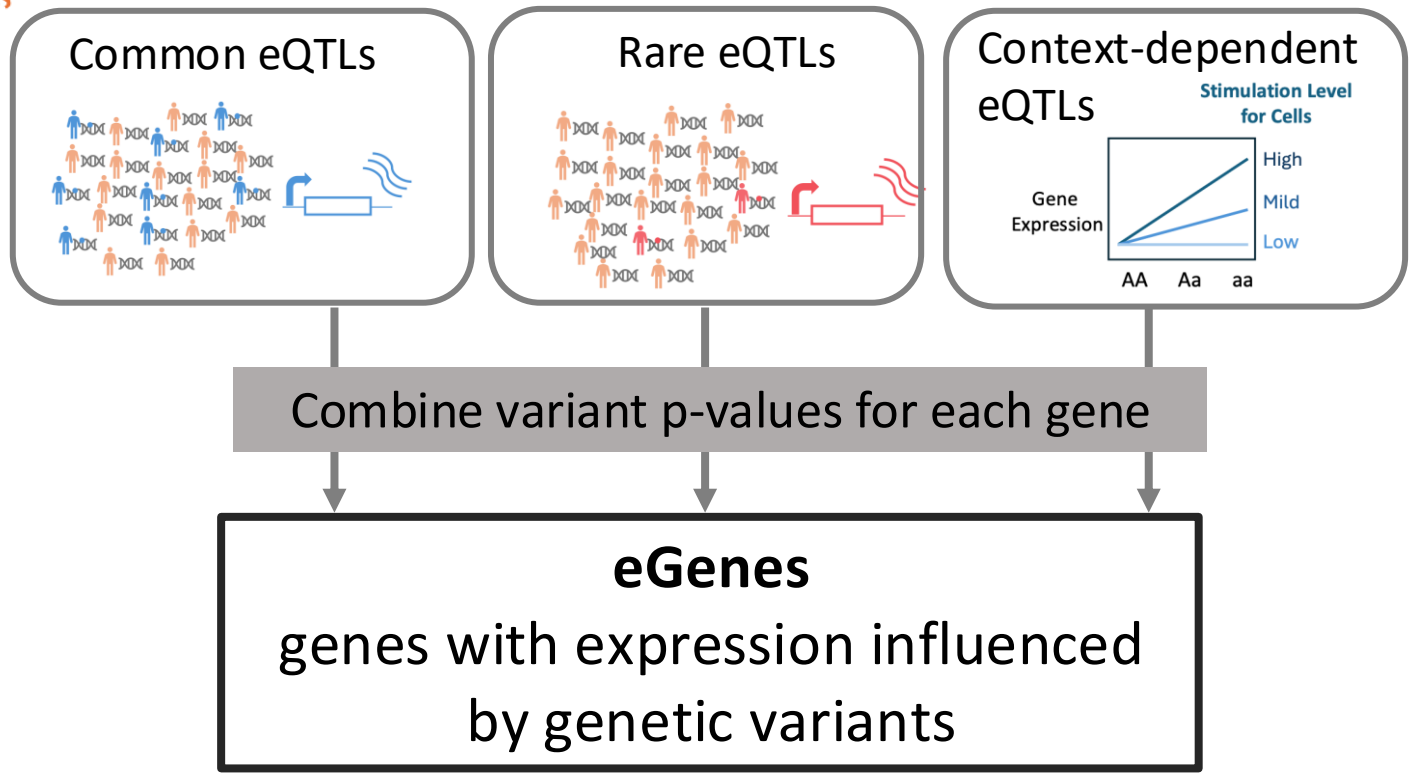
		D_1		D_2		D_3	
		c_1	c_2	c_1	c_2	c_1	c_2
D_1	c_1	1	1	0	0	0	0
	c_2	1	1	0	0	0	0
D_2	c_1	0	0	1	1	0	0
	c_2	0	0	1	1	0	0
D_3	c_1	0	0	0	0	1	1
	c_2	0	0	0	0	1	1

SAIGE-QTL: scalable and accurate tool for single-cell eQTL mapping

Challenges	Solutions
• Correlated cells	➢ **Mixed models with random effects**
• Sparse, discrete counts	➢ **Poisson linear regression**
• Millions of cells and variants	➢ **Computational and statistical optimization**
• Context-dependent effects	➢ **Interaction term (Genotype x Context)**
• Rare variants underpowered	➢ **Group rare variants for testing**

SAIGE-QTL: scalable and accurate tool for single-cell eQTL mapping





$$\text{read counts of gene X} \sim Poisson(\mu)$$

$$log(\mu) = SNP + covs\_donor + covs\_cell + random\ effect$$

Cell-cell correlation

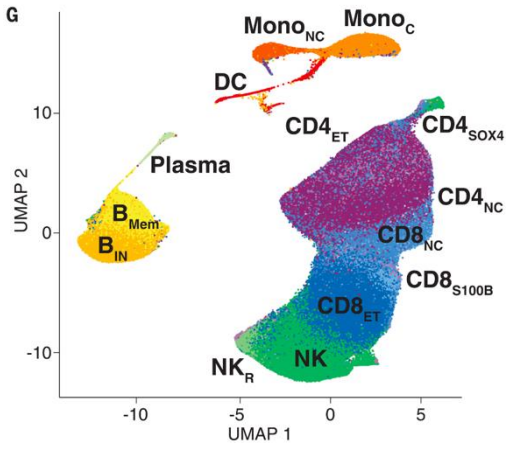
$$+ SNP \times context + context + random\ effect2$$

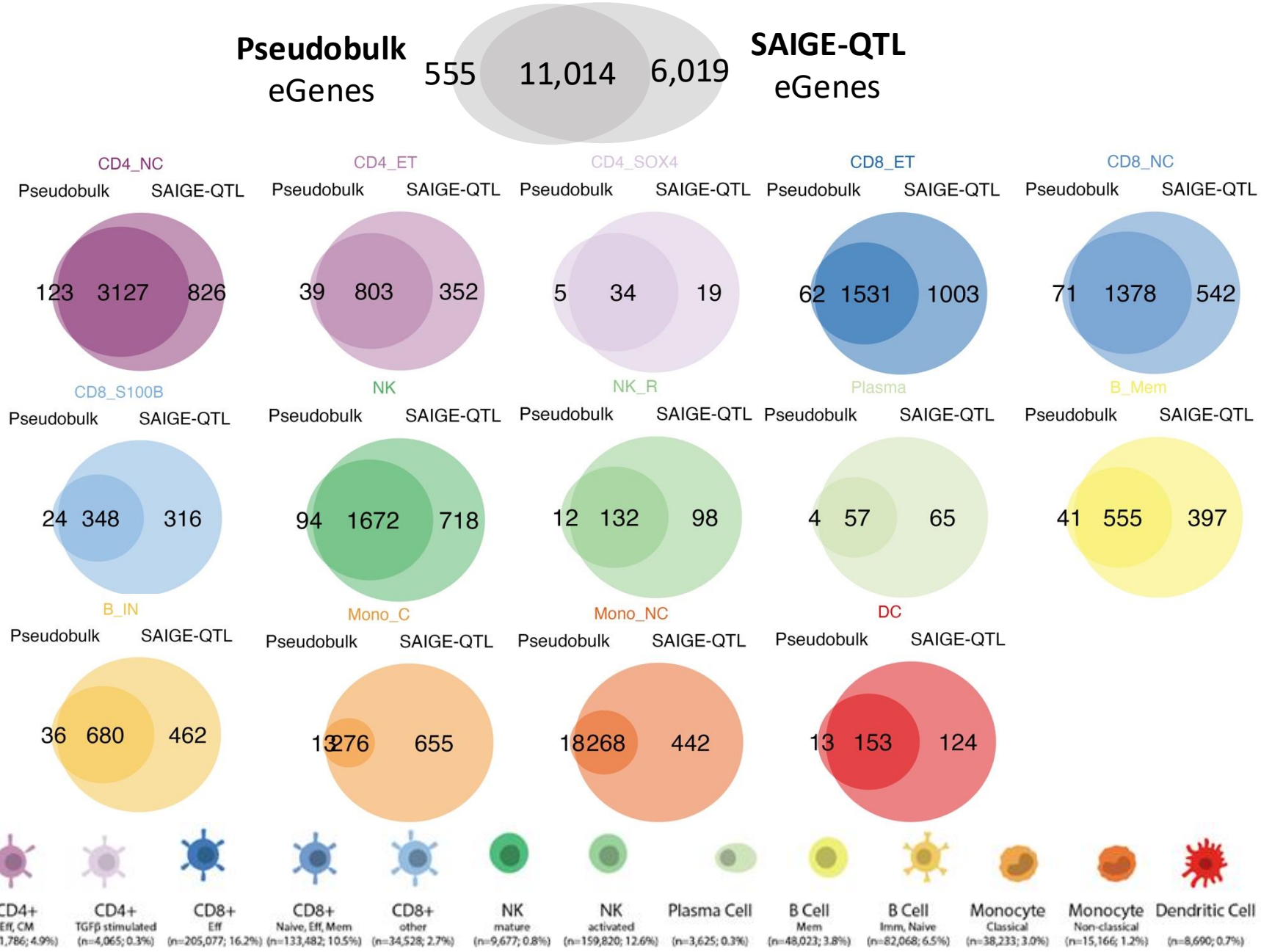
Testing eQTLs differ across one or multiple contexts, e.g. donor age, sex or cell state, stimuli

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Zhou *et al*, medRxiv, 2024

SAIGE-QTL identified ~48% more eGenes than pseudobulk





14 immune cell types:

OneK1K
Yazar et al., Science, 2022