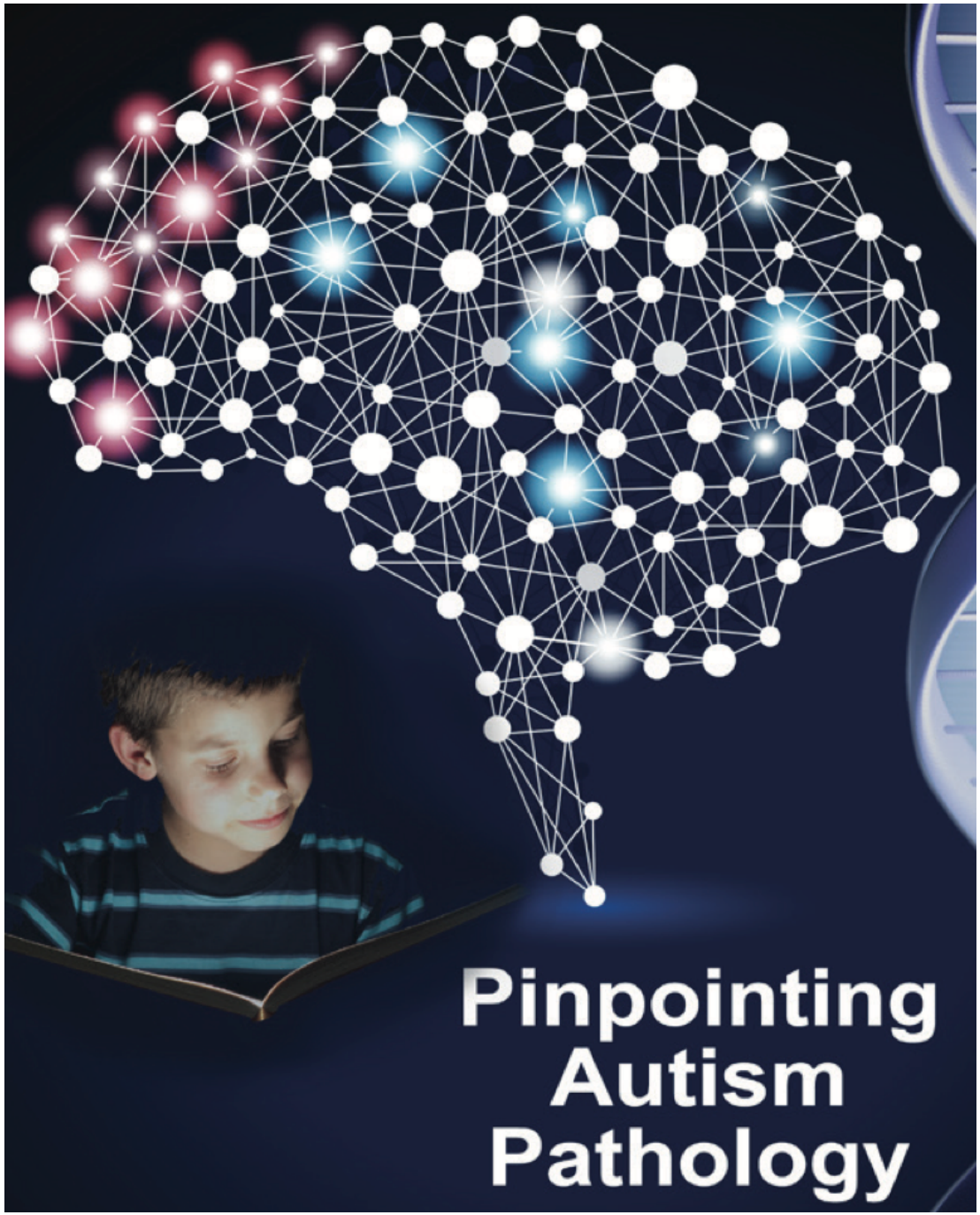
Learning from the Transcriptome: analysis of single cell and bulk RNA sequencing data

Kathryn Roeder

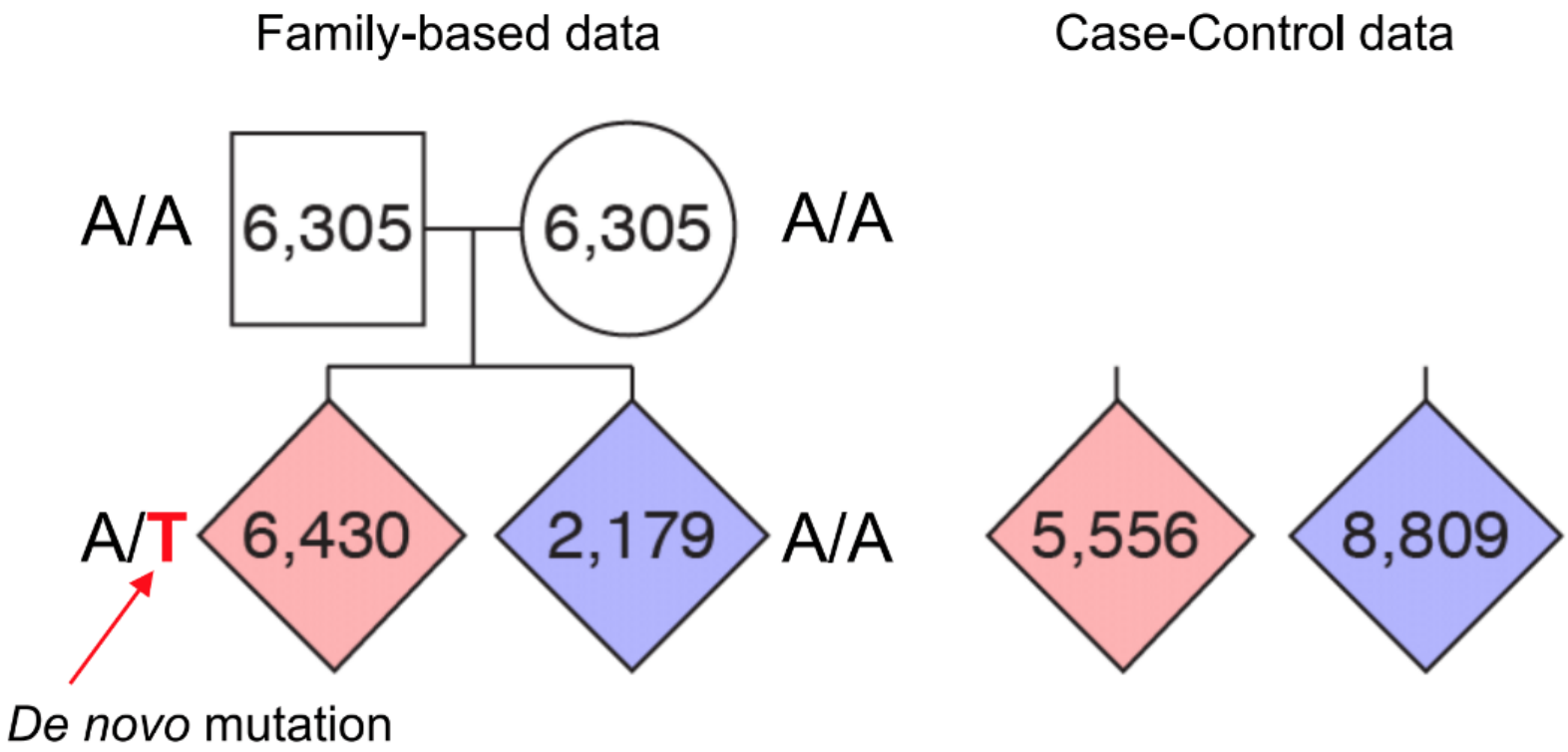
Department of Statistics and Data Science
Carnegie Mellon University

Feb 2019

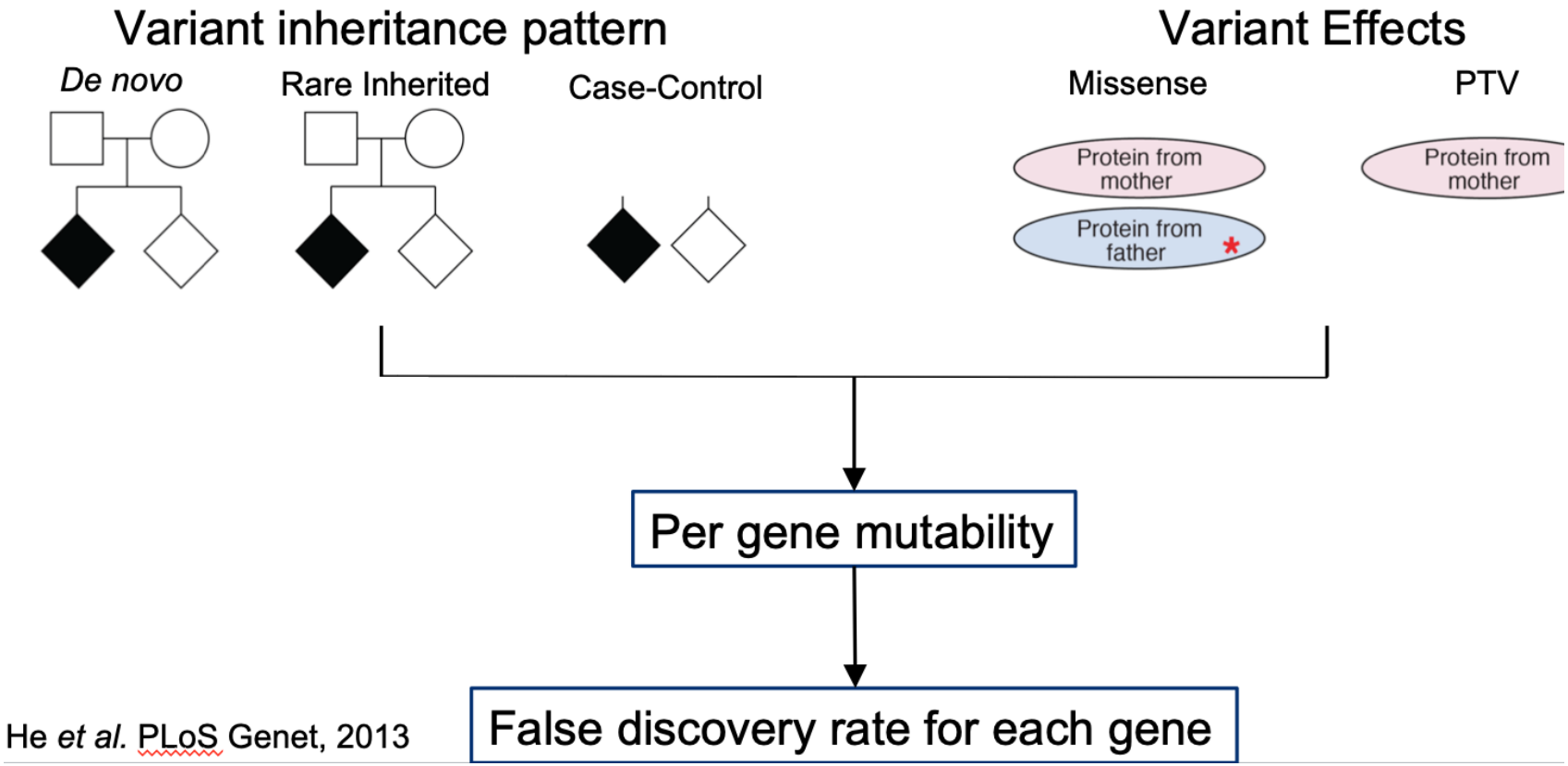

Pinpointing
Autism
Pathology

Austism Sequencing Consortium

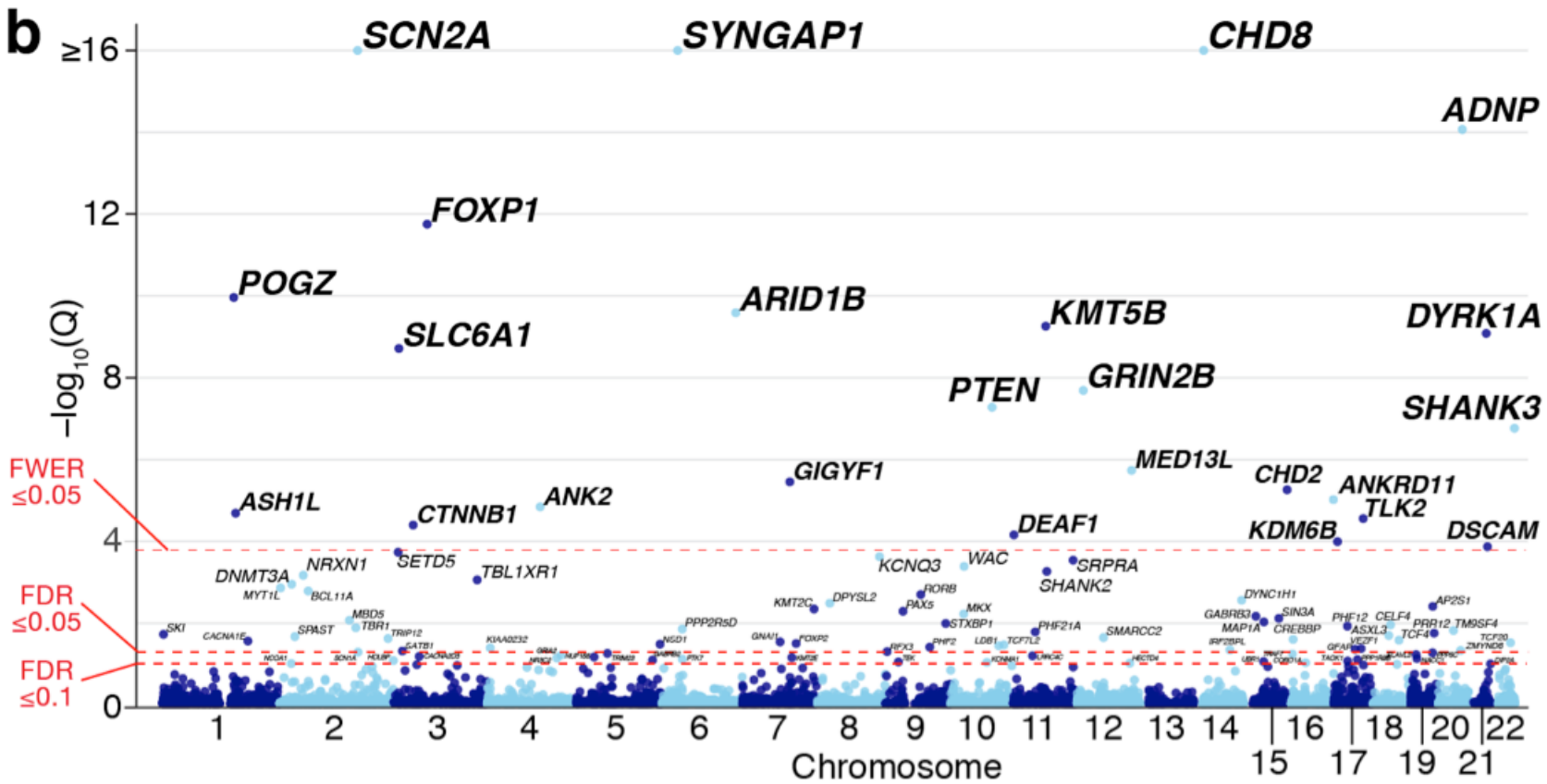
Whole exome data generated for 35,584 samples (11,986 ASD cases)



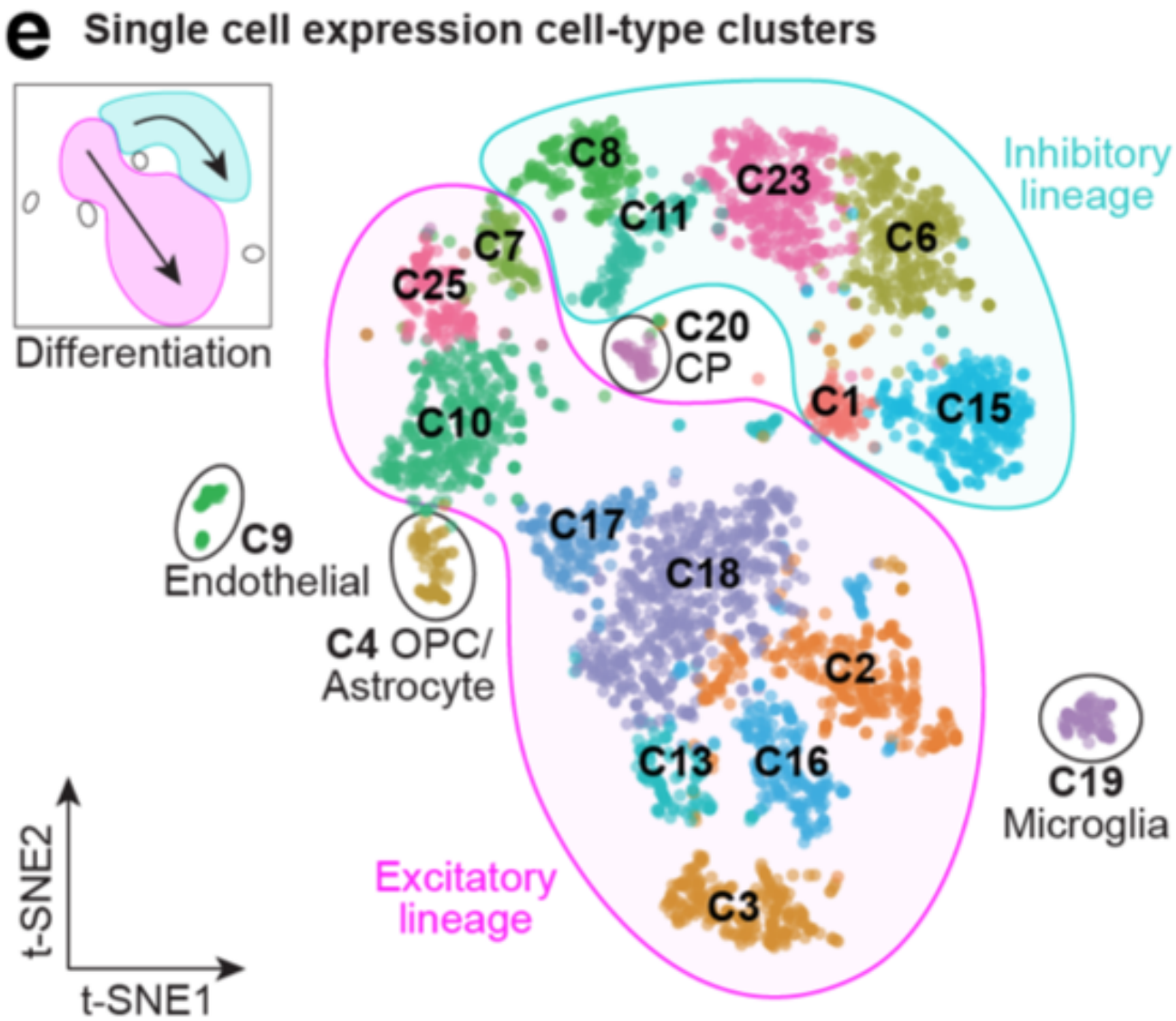
Austism Sequencing Consortium



Austism Sequencing Consortium



sc-RNAseq Human forebrain clusters: Nowakowski et al. 2017 Science



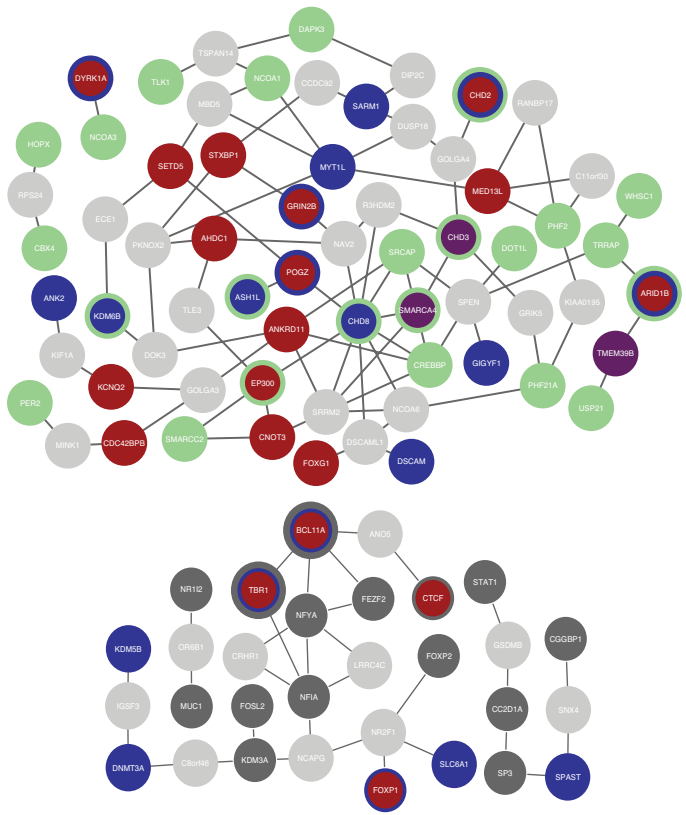
Overview

Genetics versus Genomics

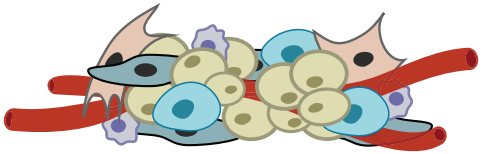
- Successful gene discovery
- What is the meaning?
- Evaluate transcription: cell type, gene-gene networks

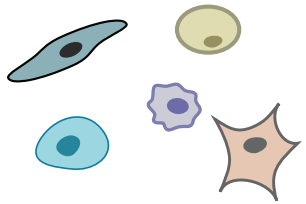
Two stories today

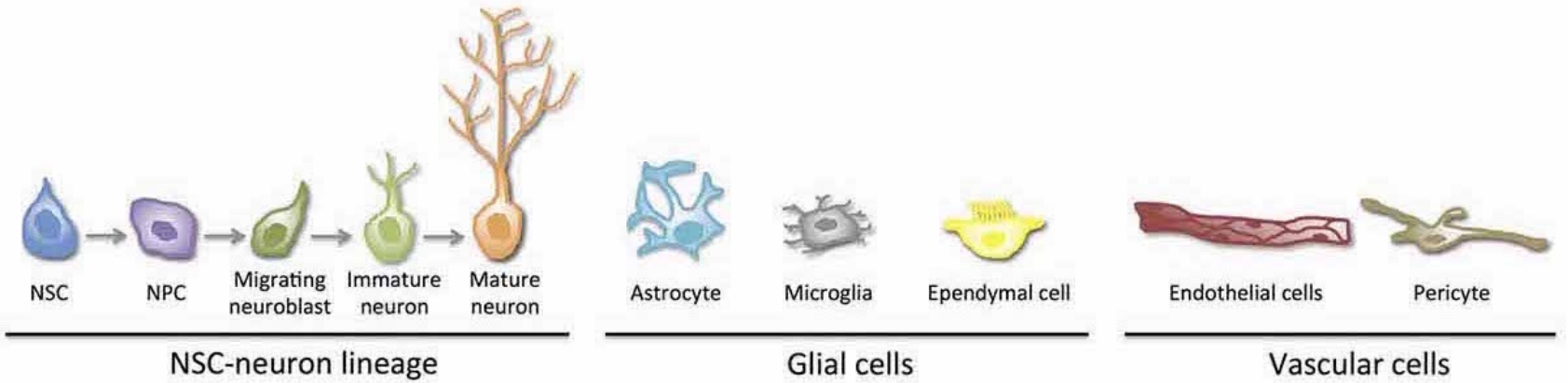
- Single Cell RNA-seq: estimating development
- Bulk RNA-seq: deconvolving multiple-samples



- Bulk RNA-seq
  - gene expression at the tissue level
  - mixture of various cell subpopulations
- Single cell RNA-seq
  - cellular gene expression levels
  - reveals cell-to-cell heterogeneity
  - high levels of technical noise





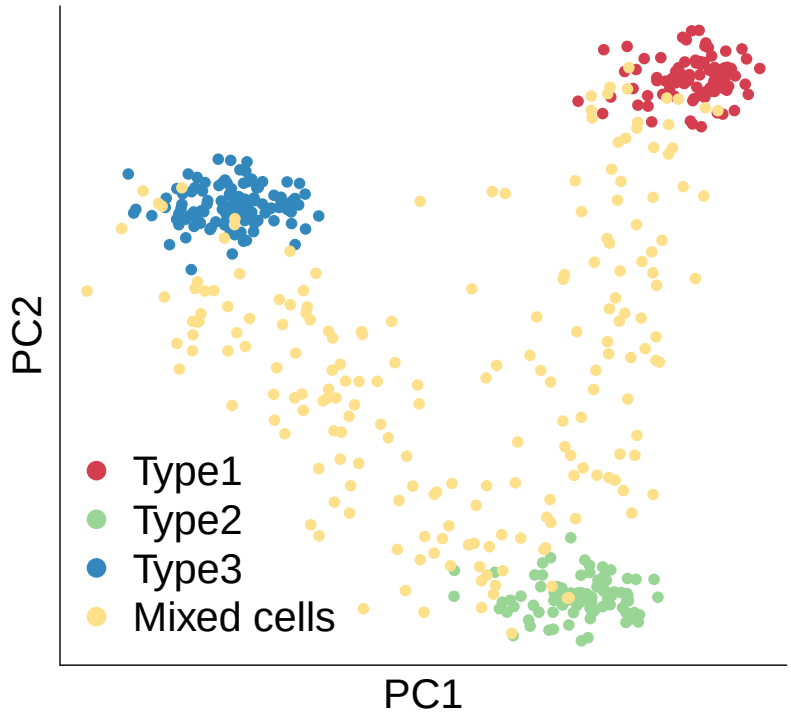


(various cell types in brain tissue)

Background

Single cell clustering

- Existing algorithms focus only on hard clustering
  - SC3, CIDR, Seurat ... *[Kiselev et al. (2017); Lin et al. (2017); Satija et al. (2015)]*
- Single cell data can be developing between cell types

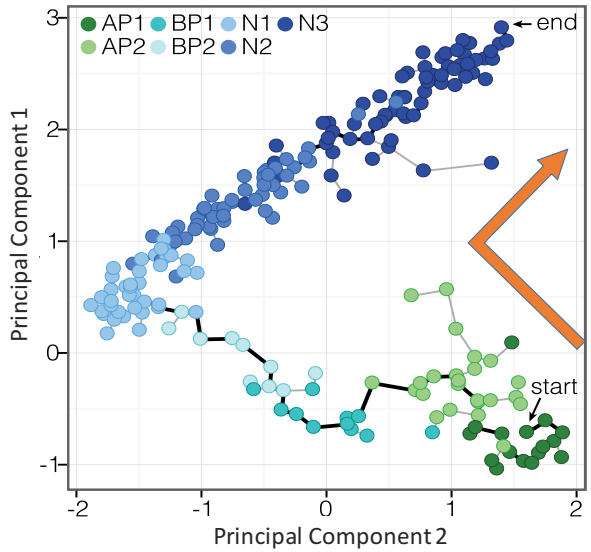


Application Results

Fetal brain cells, Camp et al.

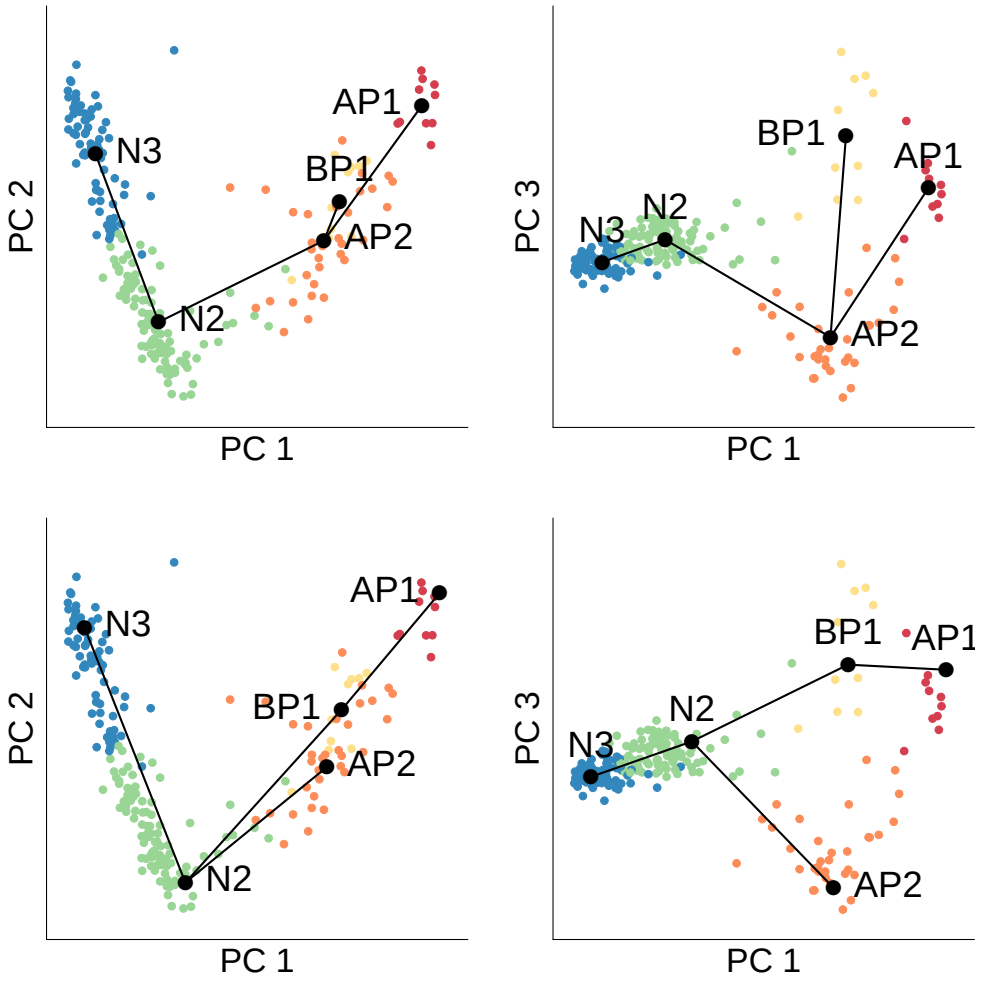
- 220 fetal brain cells
  - 12-13 post-conception weeks
  - $12,694 \rightarrow 430$ selected genes
  - 7 cell types
    - apical progenitor (AP1, AP2)
    - $\rightarrow$ basal progenitor (BP1, BP2)
    - $\rightarrow$ neuron (N1, N2, N3)

[Camp et al. (2015)]



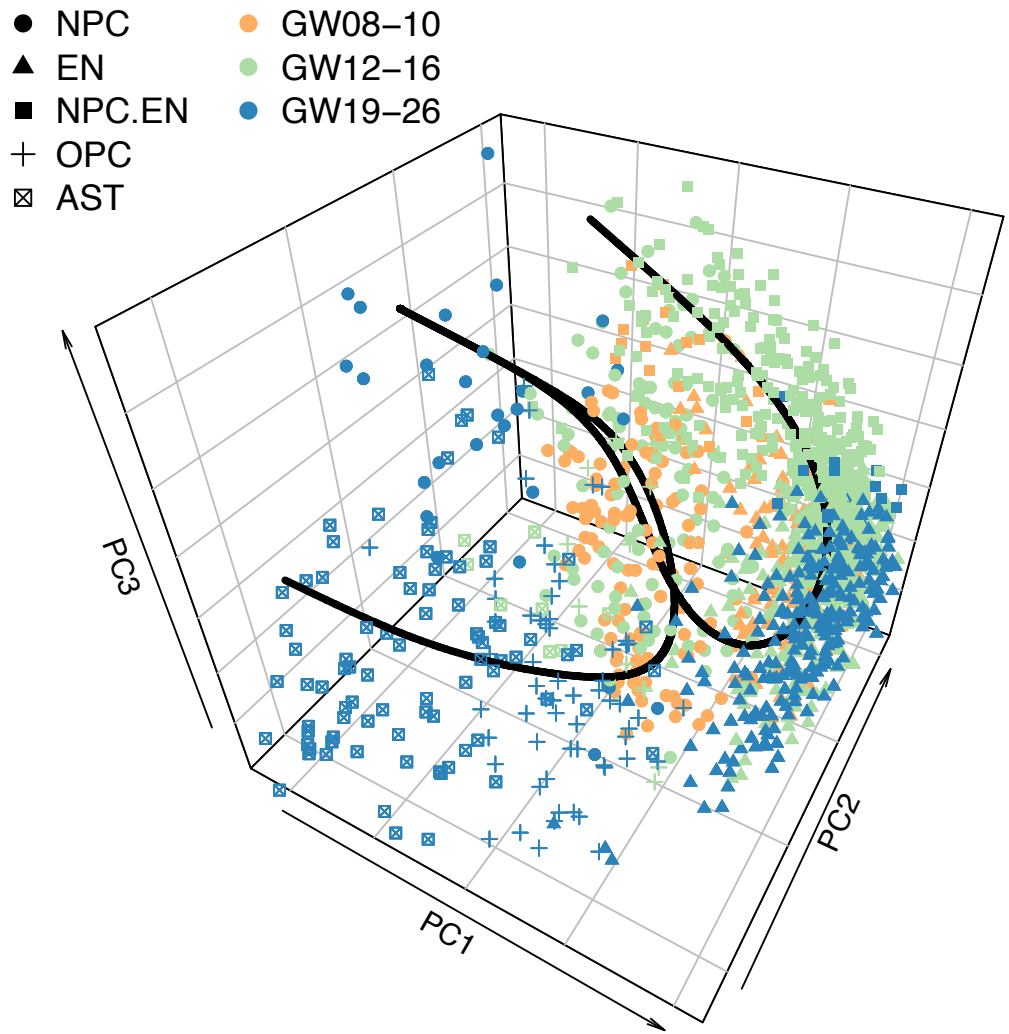
Application Results

Developmental Trajectories



Application Results

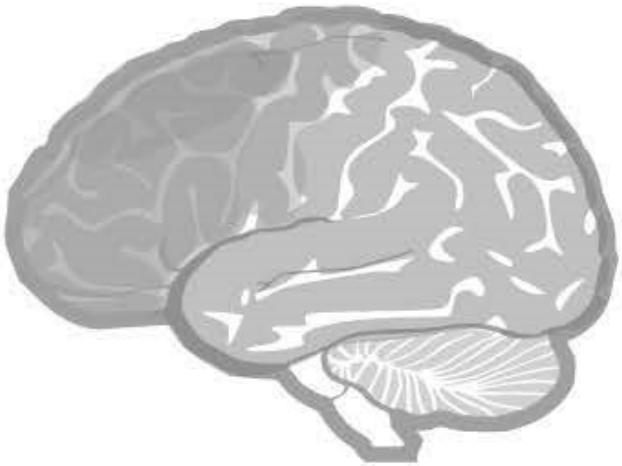
Developmental Trajectories



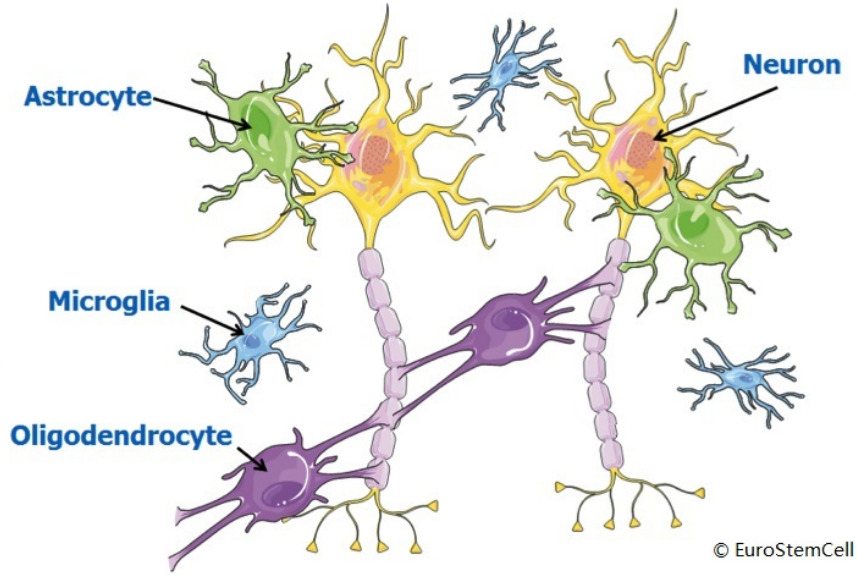
SOUP publication

Zhu, Lei, Klei, Devlin, Roeder, "Semisoft clustering of single-cell data", PNAS (2019)

What can we learn from bulk RNA-seq data?



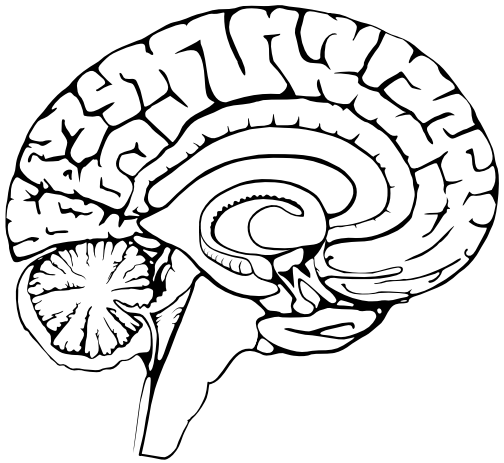
tissue expression data



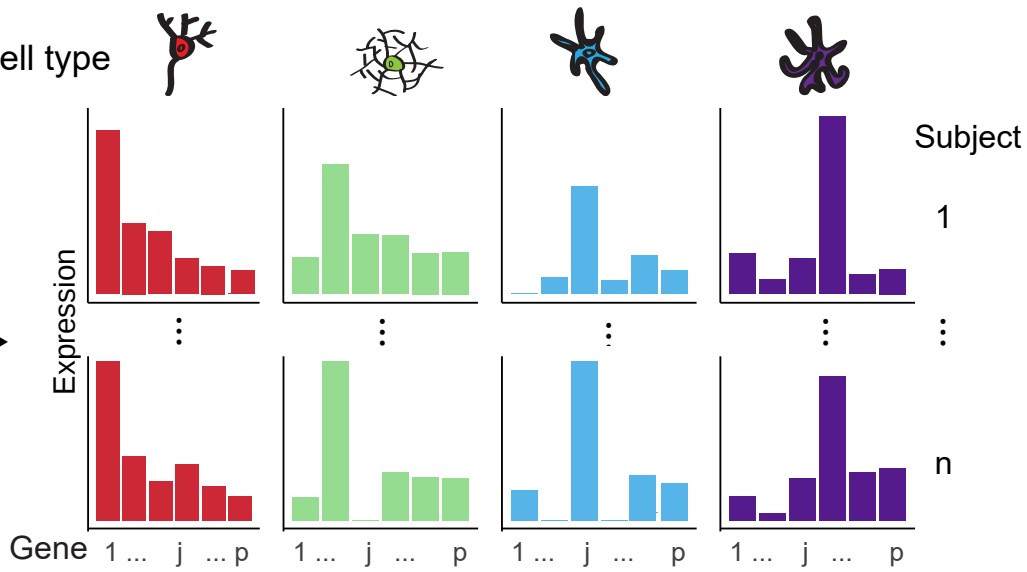
purified-cells data
single-cell

RNA-seq data

What can we learn from tissue expression data?

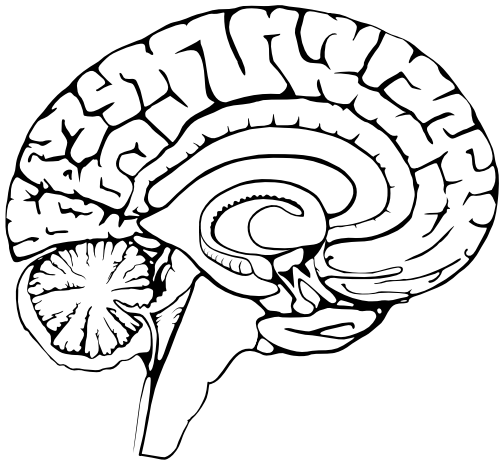


tissue expression

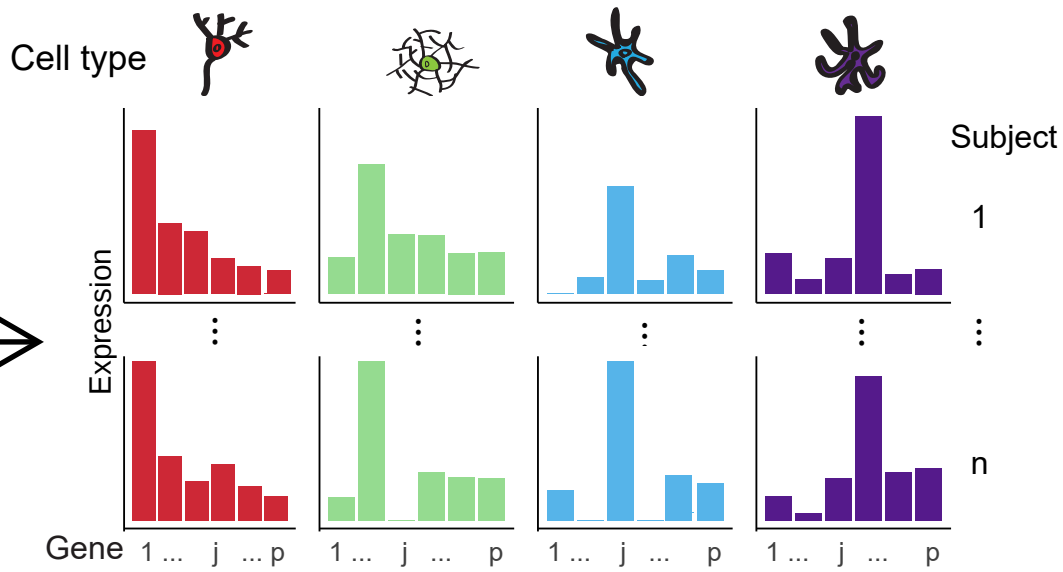


subject-level cell-type-specific expression

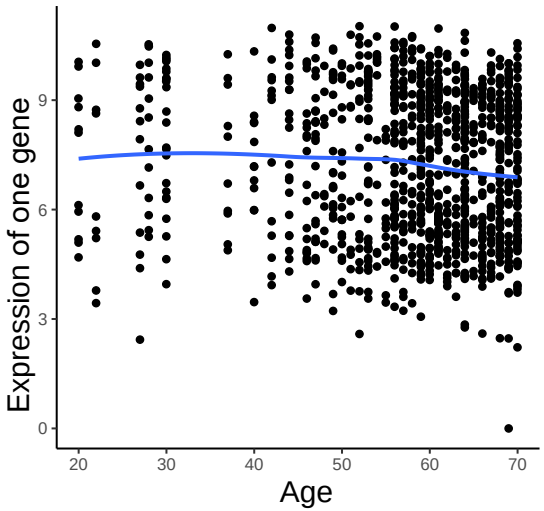
What can we learn from tissue expression data?



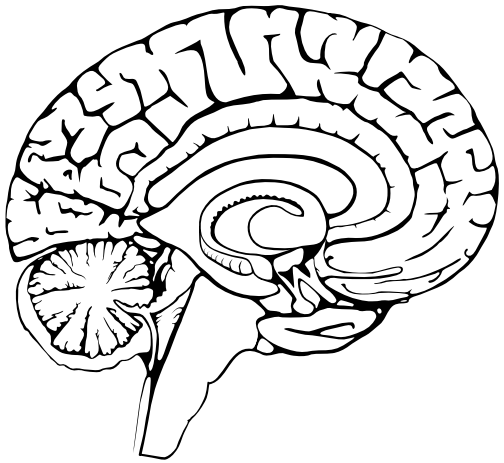
tissue expression

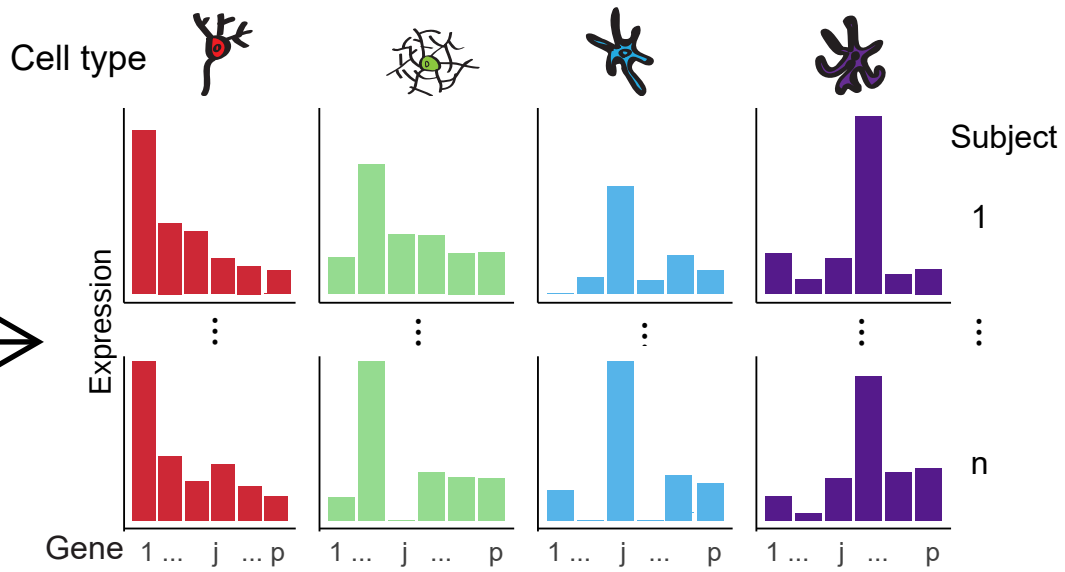


subject-level cell-type-specific expression



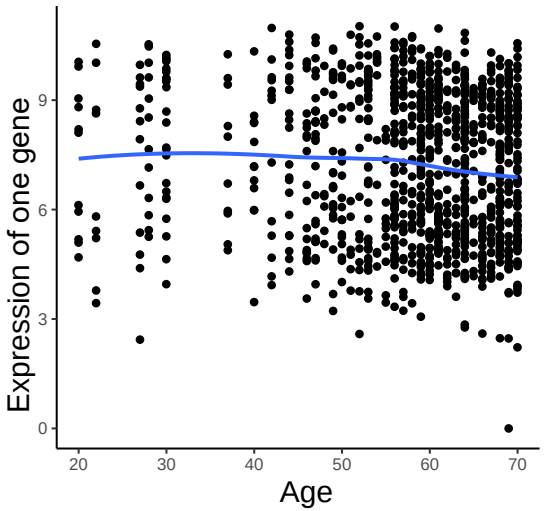
What can we learn from tissue expression data?

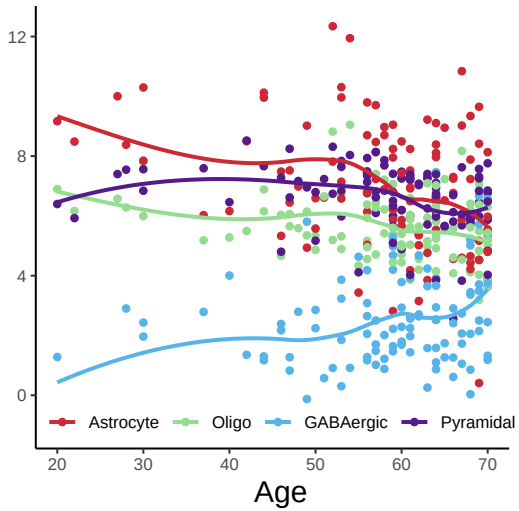




tissue expression

subject-level cell-type-specific expression





Gene expression deconvolution

- The deconvolution model is written as

$$\underset{(p\times n)}{X} \approx \underset{(p\times K)}{A}\,\underset{(K\times n)}{W},$$

  - X: **single-measure** tissue expression for p genes in n subjects,
  - A: average gene expression over subjects for K cell types,
  - W: mixing fractions of K cell types per subject (col.sum $= 1$).

Gene expression deconvolution

- The deconvolution model is written as

$$\underset{(p\times n)}{X} \approx \underset{(p\times K)}{A}\,\underset{(K\times n)}{W},$$

  - X: **single-measure** tissue expression for p genes in n subjects,
  - A: average gene expression over subjects for K cell types,
  - W: mixing fractions of K cell types per subject (col.sum $= 1$).

	subject				cell type			subject		
gene	6.9	7.1	6.7	$\approx$	gene	7.5	5.6	cell	0.7	0.8	0.6
	7.6	7.8	7.4			8.2	6.1	type	0.3	0.2	0.4
	5.2	5.2	5.1			5.3	4.9				
	tissue expression				**cell-type-specific expression**				**fraction**		

Gene expression deconvolution

- The deconvolution model is written as

$$\underset{(p\times n)}{X} \approx \underset{(p\times K)}{A}\ \underset{(K\times n)}{W},$$

  - X: **single-measure** tissue expression for p genes in n subjects,
  - A: average gene expression over subjects for K cell types,
  - W: mixing fractions of K cell types per subject (col.sum $= 1$).

	subject				cell type			subject		
gene	6.9	7.1	6.7	≈ gene	7.5	5.6	cell	0.7	0.8	0.6
	7.6	7.8	7.4		8.2	6.1	type	0.3	0.2	0.4
	5.2	5.2	5.1		5.3	4.9				
	tissue expression				**cell-type-specific expression**			**fraction**		

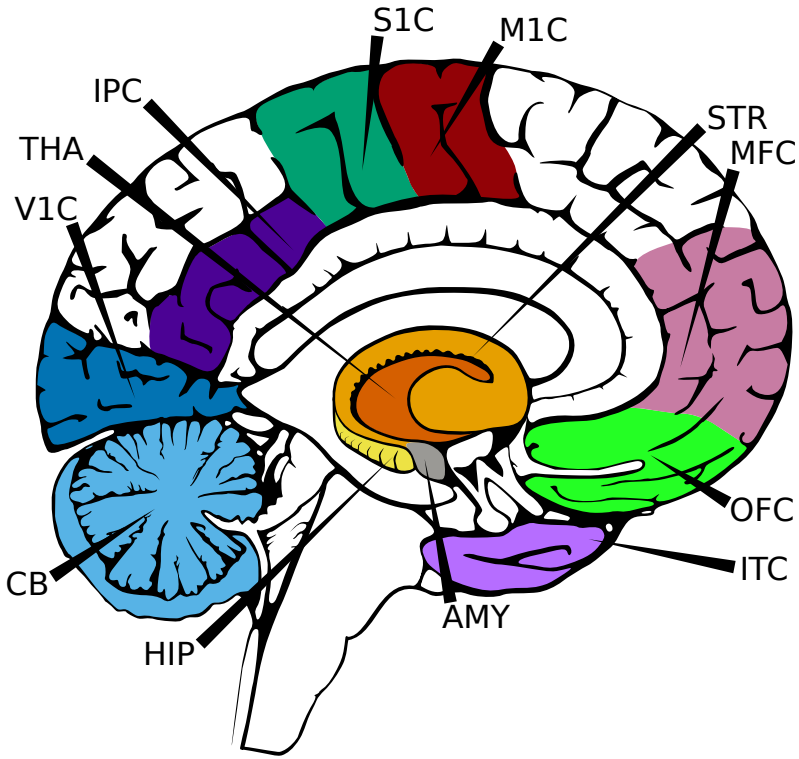
- Assumption:
  - A (cell-type-specific expression) is **constant across subjects**

Existing single-measure deconvolution algorithms

- Unsupervised deconvolution:
  - Estimating both A and W
    - non-negative matrix factorization (NMF)
- Semi-supervised deconvolution:
  - Given sparse structure of A, estimating A and W
    - semi-supervised NMF
    - quadratic programming
- **Supervised deconvolution**:
  - Given A, estimating W
    - least squares
    - Bayesian estimation
    - support vector regression
  - Given W, estimating A
    - least squares

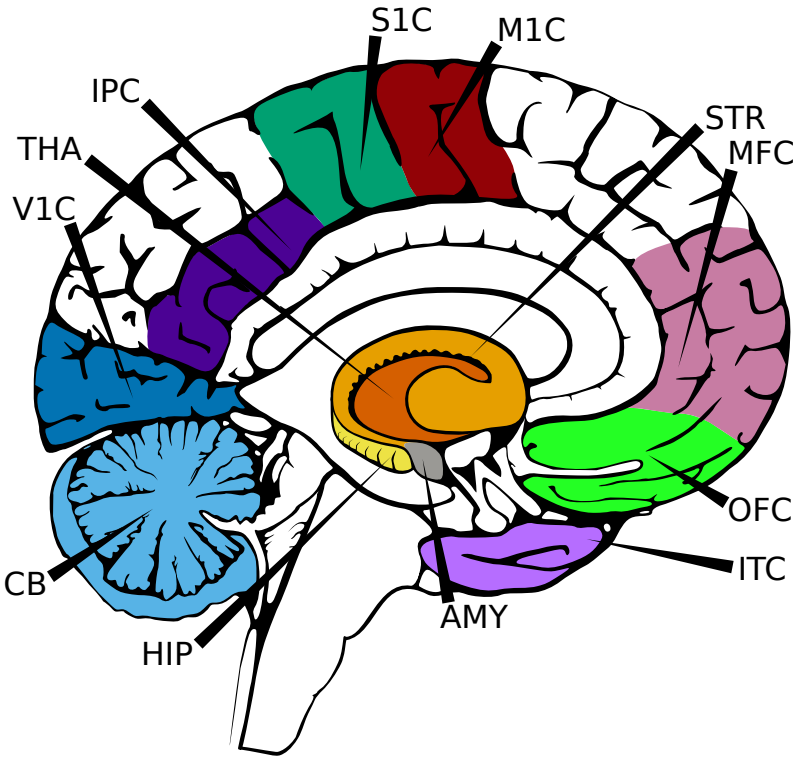
Multi-measure expression data

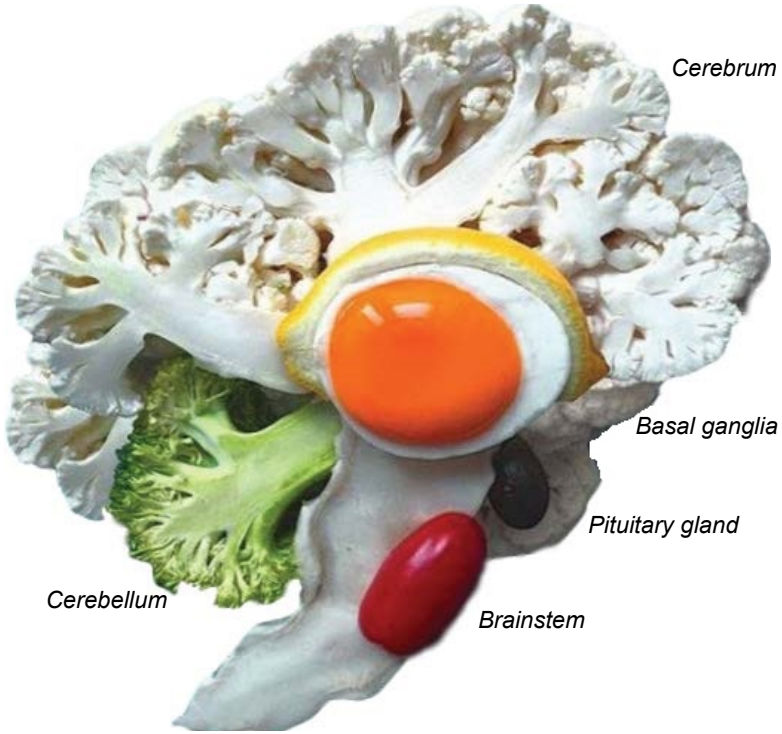
GTEx (Genotype-Tissue Expression) project: 13 brain regions/measures; 105 subjects
BrainSpan atlas of the developing human brain: 26 brain regions/measures; 33 subjects



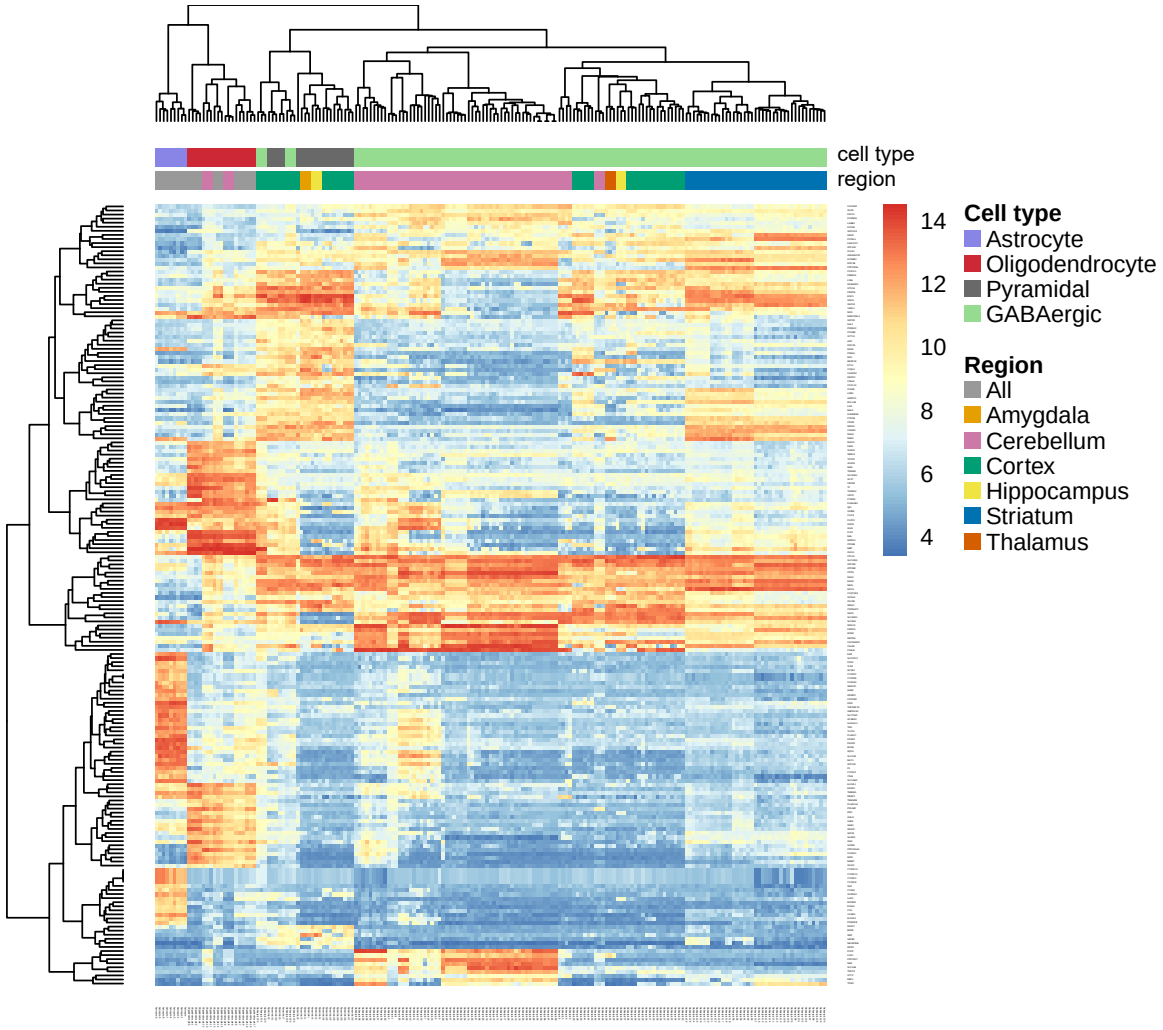
Multi-measure expression data

GTEx (Genotype-Tissue Expression) project: 13 brain regions/measures; 105 subjects
BrainSpan atlas of the developing human brain: 26 brain regions/measures; 33 subjects





Nueroexpresso: Variability by cell type and region



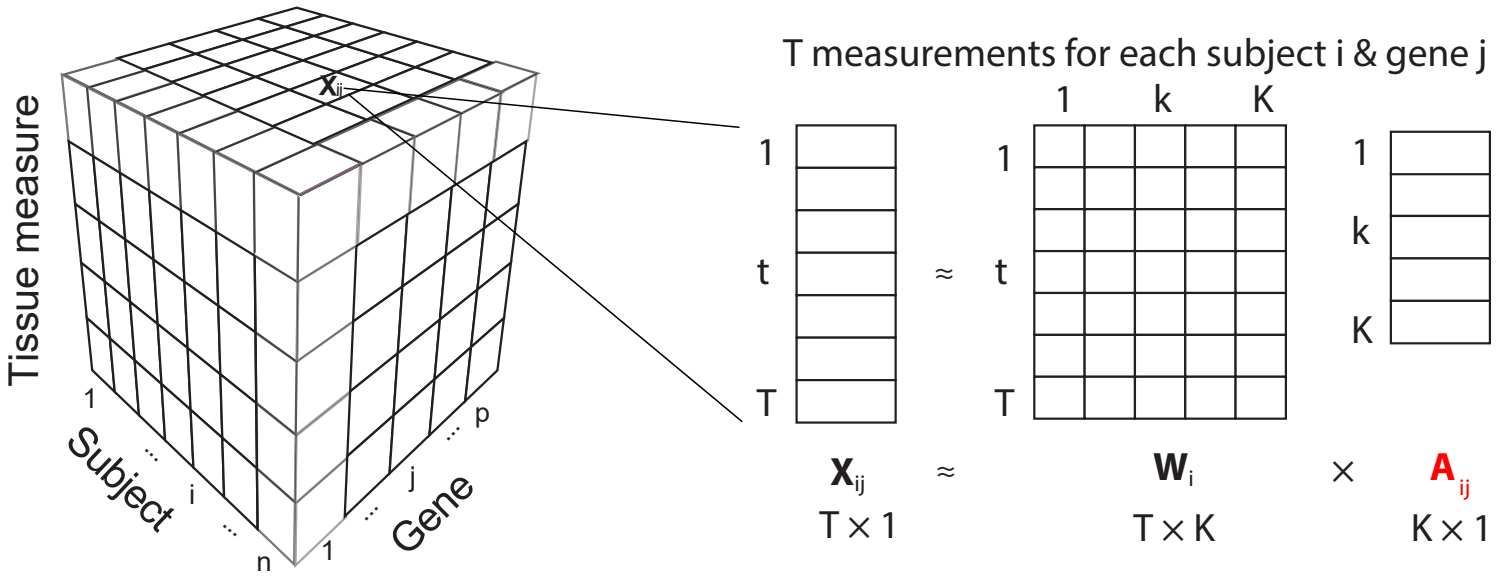
New idea: multi-measure deconvolution

Goal: estimate individual-level cell-type expression

Assumptions:

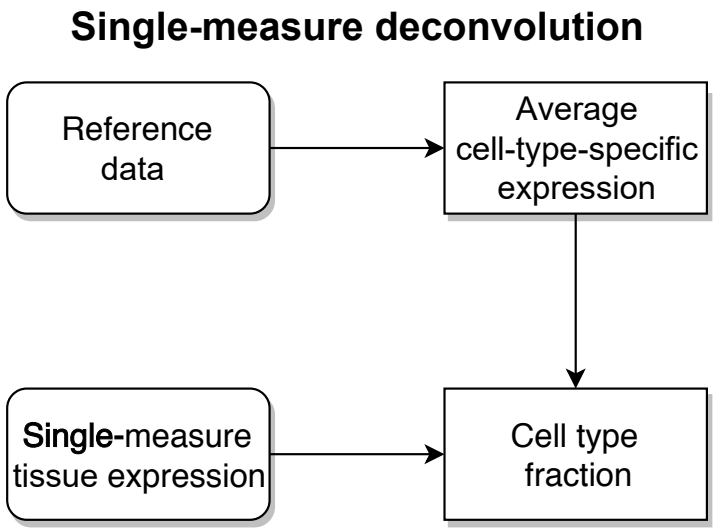
- Expected cell type expression is constant across measurements for an individual
  - Cells of a given type have a predictable expression pattern
  - Expression varies by individual because of genetic variation, developmental stage, disease status etc.
- Cell-type fraction varies by individual (i) and measurement (t)
  - Pre-estimate W_i: individual-level cell-type fraction, for each t using single cell data

New idea: multi-measure deconvolution (MIND)

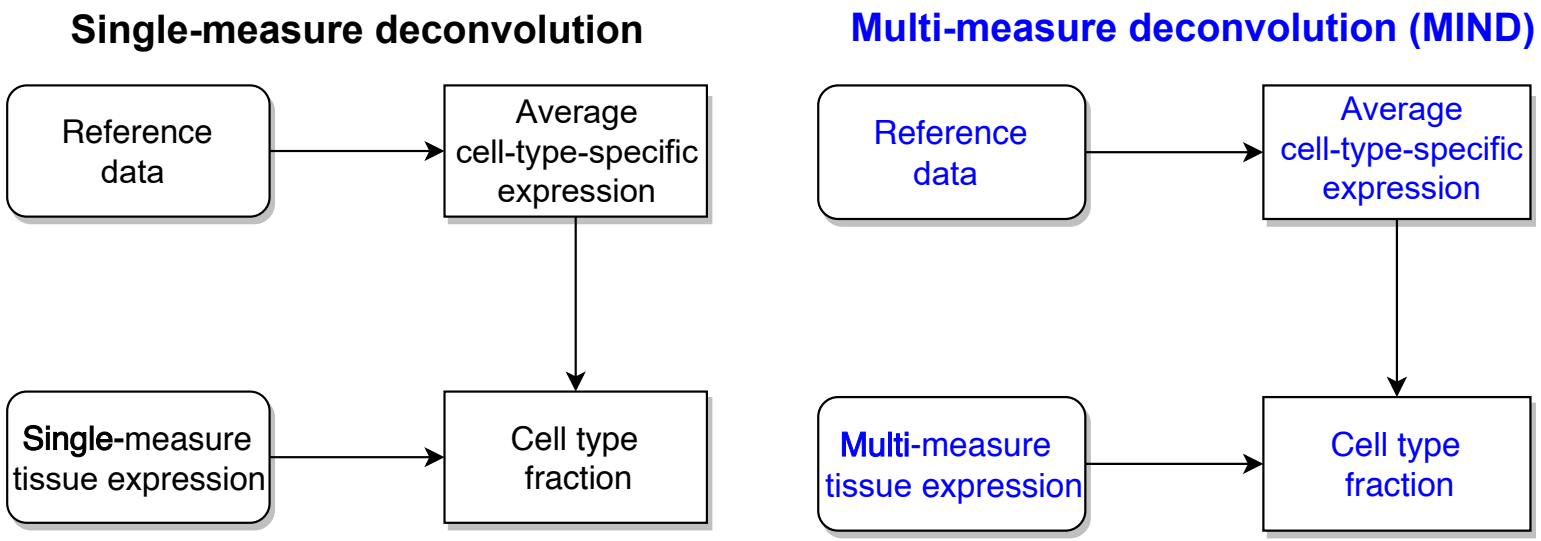


- $\boldsymbol{X}_{ij}$: tissue expression across multi-measures (observed)
- $\boldsymbol{W}_i$: pre-estimated cell type fractions (given)
- $\boldsymbol{A}_{ij}$: **subject-level** cell-type-specific gene expression (output)

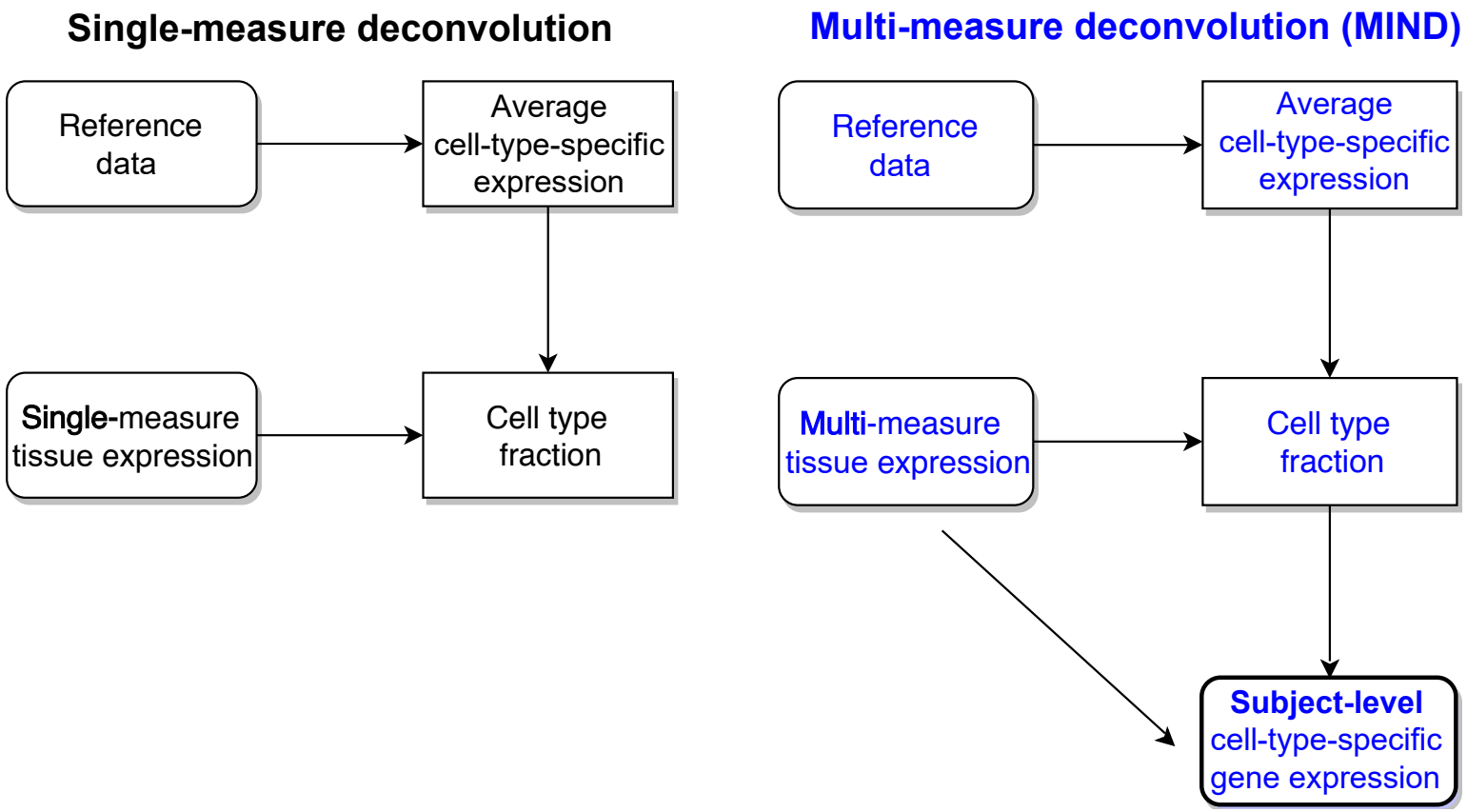
Single-measure vs. multi-measure deconvolution



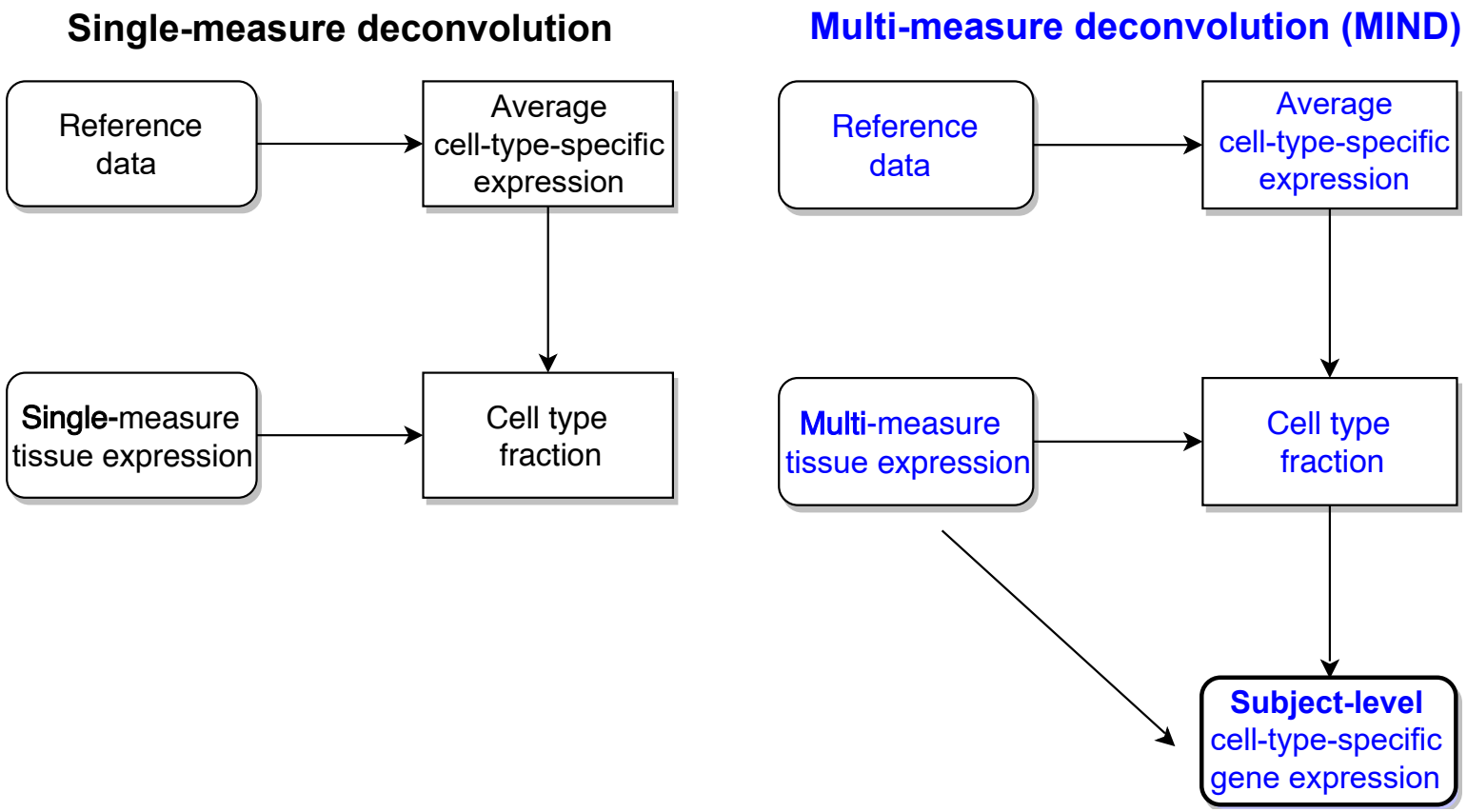
Single-measure vs. multi-measure deconvolution



Single-measure vs. multi-measure deconvolution



Single-measure vs. multi-measure deconvolution



Reference data with cell type information: scRNA-seq, NeuroExpresso
Multi-measure expression: GTEx, BrainSpan, ...

Three-level random-effects model for MIND

- Three-level random-effects model:

$$\begin{aligned} \underset{(T\times 1)}{\boldsymbol{X}_{ij}} &= \underset{(T\times K)(K\times 1)}{\boldsymbol{W}_i \ \boldsymbol{A}_{ij}} + \underset{(T\times 1)}{\boldsymbol{e}_{ij}} \ ; \\ \boldsymbol{A}_{ij} &\sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}_c\right), \\ \boldsymbol{e}_{ij} &\sim N\left(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_T\right). \end{aligned}$$

  - level 1: $T \approx 10$ measures
  - level 2: $p \approx 20,000$ genes (indexed by j)
  - level 3: $n \approx 100$ subjects (indexed by i)
  - input: $\boldsymbol{X}$ $(n \times p \times T)$, $\boldsymbol{W}$ $(n \times T \times K)$
  - output: $\boldsymbol{A}$ $(n \times p \times K)$

- We derived a computationally efficient EM algorithm:
  - Parameters are estimated via maximum likelihood;
  - All genes can be deconvolved together in minutes.

Estimation: random effects

Cell-type-specific expression ($\boldsymbol{A}_{ij}$, random effect) is estimated using an **empirical Bayes** method:

- Estimates of random effects: conditional mean of random effects given observed data and estimated parameter values

$$\hat{\boldsymbol{A}}_{ij} = \left[\boldsymbol{I} + \hat{\sigma}_e^2 \left(\hat{\boldsymbol{\Sigma}}_c \boldsymbol{W}_i' \boldsymbol{W}_i\right)^{-1}\right]^{-1} \left(\boldsymbol{W}_i' \boldsymbol{W}_i\right)^{-1} \boldsymbol{W}_i' \boldsymbol{X}_{ij}$$

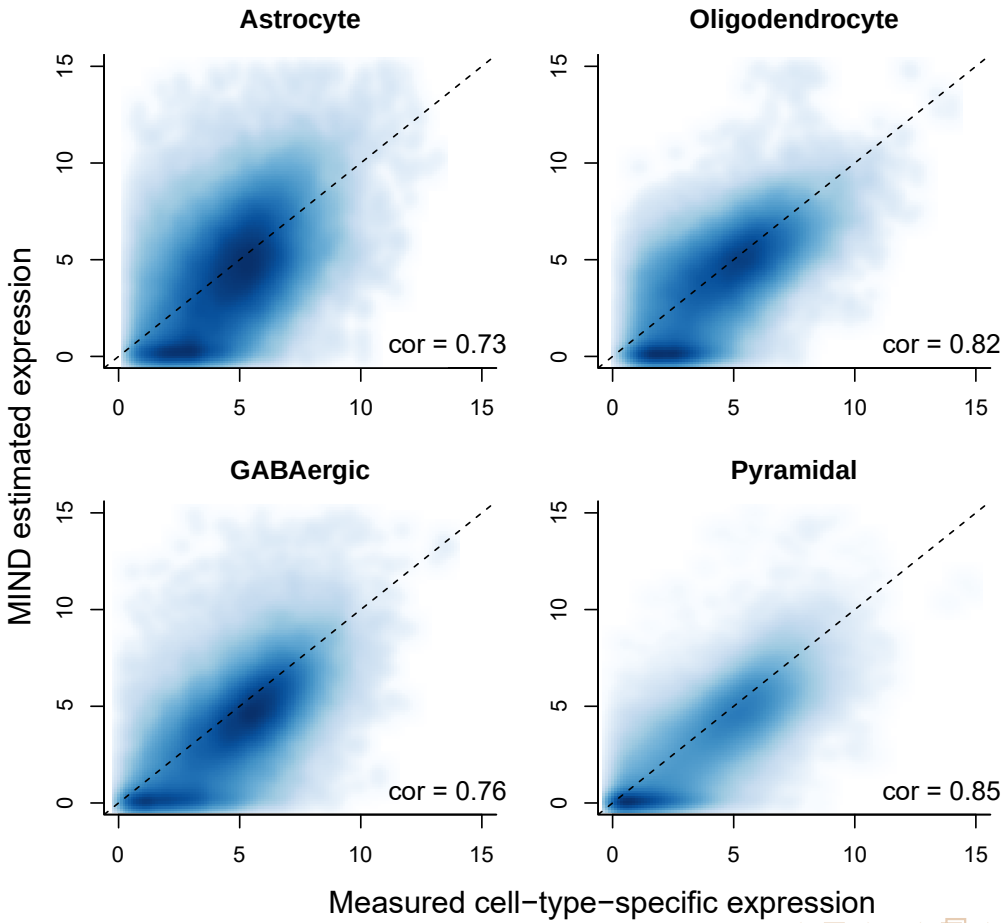
- Shrinkage to the origin (James-Stein estimator)
- Weight depends on variance components and $\boldsymbol{W}_i$
- More robust to outliers than least squares

Method evaluation: deconvolving GTEx brain data

- Measured cell-type-specific expression ($\boldsymbol{A}_{ij}$) from scRNA-seq (ground truth) for several subjects
- Estimated $\hat{\boldsymbol{A}}_{ij}$ by MIND for the same subjects

Method evaluation: deconvolving GTEx brain data

- Measured cell-type-specific expression ($\boldsymbol{A}_{ij}$) from scRNA-seq (ground truth) for several subjects
- Estimated $\hat{\boldsymbol{A}}_{ij}$ by MIND for the same subjects

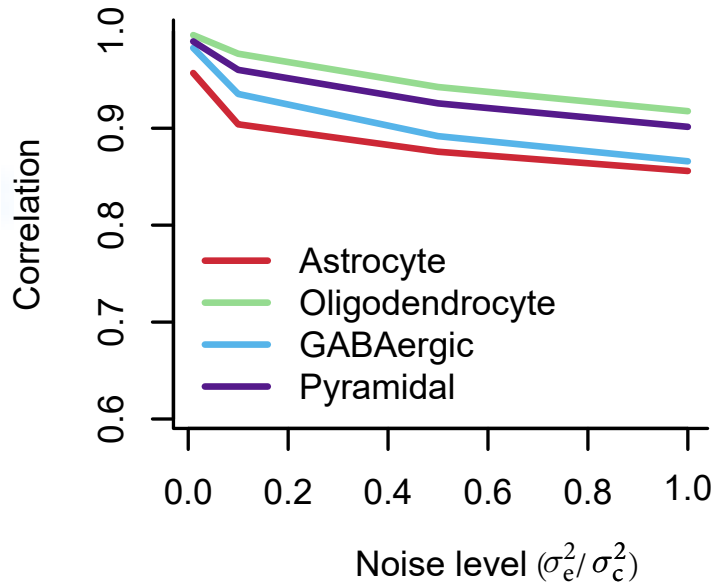


Method evaluation: simulation with real data

- Simulate tissue expression data ($\boldsymbol{X}_{ij}$) with
  - cell-type-specific expression ($\boldsymbol{A}_{ij}$) measured from scRNA-seq
  - cell type fraction ($\boldsymbol{W}_i$) estimated in GTEx
  - $\boldsymbol{e}_{ij}$ with variance $\sigma_e^2 \propto \sigma_c^2$ (variance of $\boldsymbol{A}_{ij}$)
- Calculate the correlation between deconvolved ($\hat{\boldsymbol{A}}_{ij}$) and true cell-type-specific expression ($\boldsymbol{A}_{ij}$)

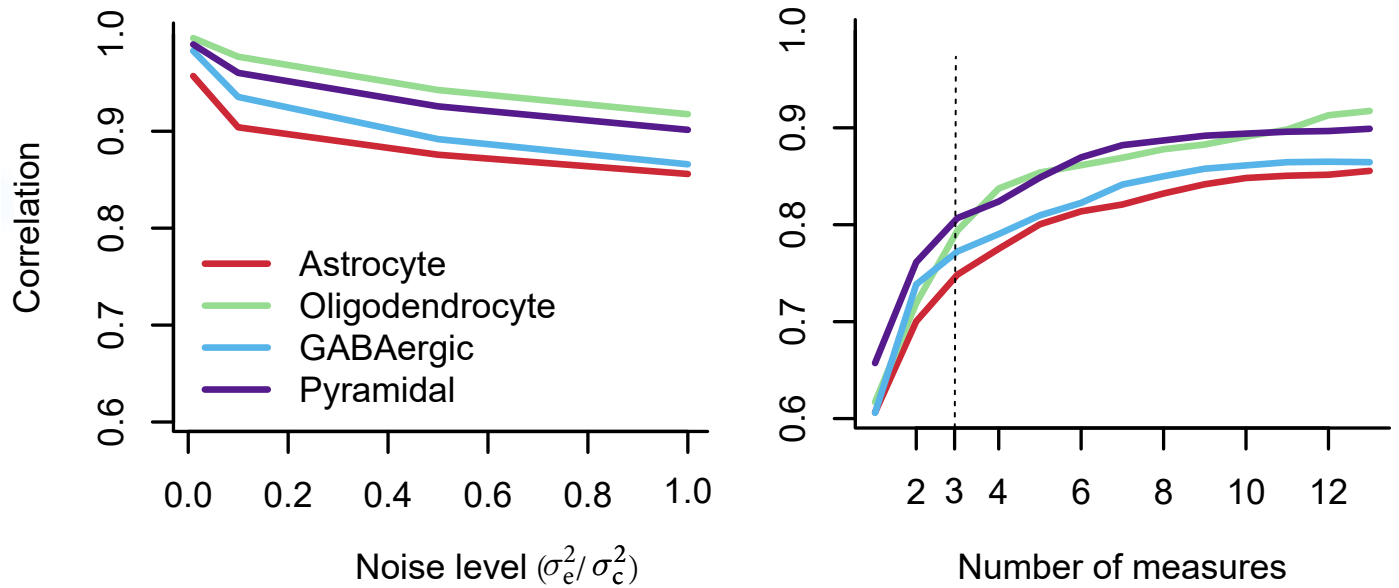
Method evaluation: simulation with real data

- Simulate tissue expression data ($\boldsymbol{X}_{ij}$) with
  - cell-type-specific expression ($\boldsymbol{A}_{ij}$) measured from scRNA-seq
  - cell type fraction ($\boldsymbol{W}_i$) estimated in GTEx
  - $\boldsymbol{e}_{ij}$ with variance $\sigma_e^2 \propto \sigma_c^2$ (variance of $\boldsymbol{A}_{ij}$)
- Calculate the correlation between deconvolved ($\hat{\boldsymbol{A}}_{ij}$) and true cell-type-specific expression ($\boldsymbol{A}_{ij}$)



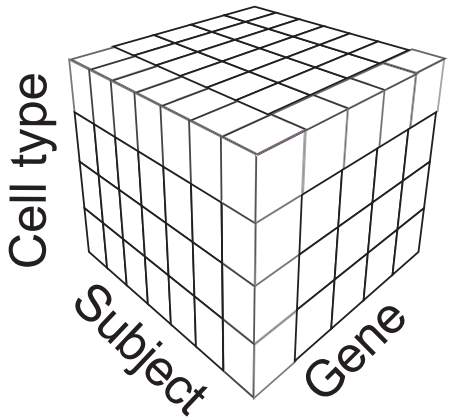
Method evaluation: simulation with real data

- Simulate tissue expression data ($\boldsymbol{X}_{ij}$) with
  - cell-type-specific expression ($\boldsymbol{A}_{ij}$) measured from scRNA-seq
  - cell type fraction ($\boldsymbol{W}_i$) estimated in GTEx
  - $\boldsymbol{e}_{ij}$ with variance $\sigma_e^2 \propto \sigma_c^2$ (variance of $\boldsymbol{A}_{ij}$)
- Calculate the correlation between deconvolved ($\hat{\boldsymbol{A}}_{ij}$) and true cell-type-specific expression ($\boldsymbol{A}_{ij}$)



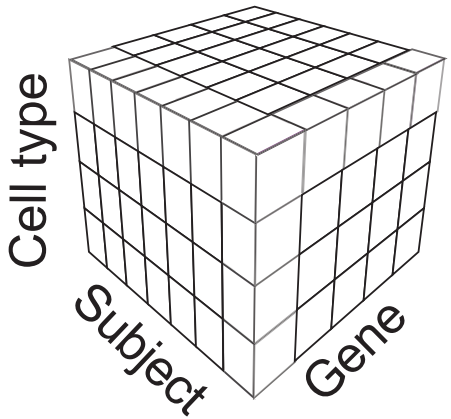
How can we use MIND?

Subject-level cell-type-specific expression



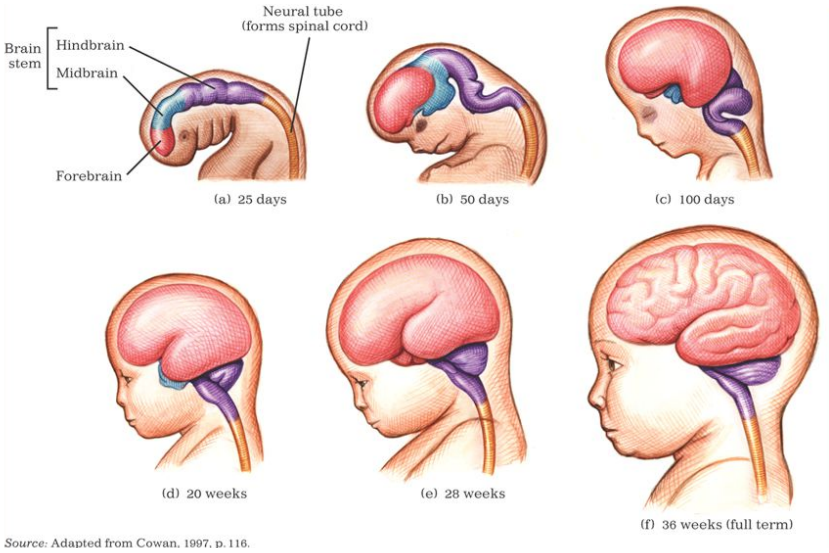
How can we use MIND?

Subject-level cell-type-specific expression can provide novel insights that are previously unavailable:



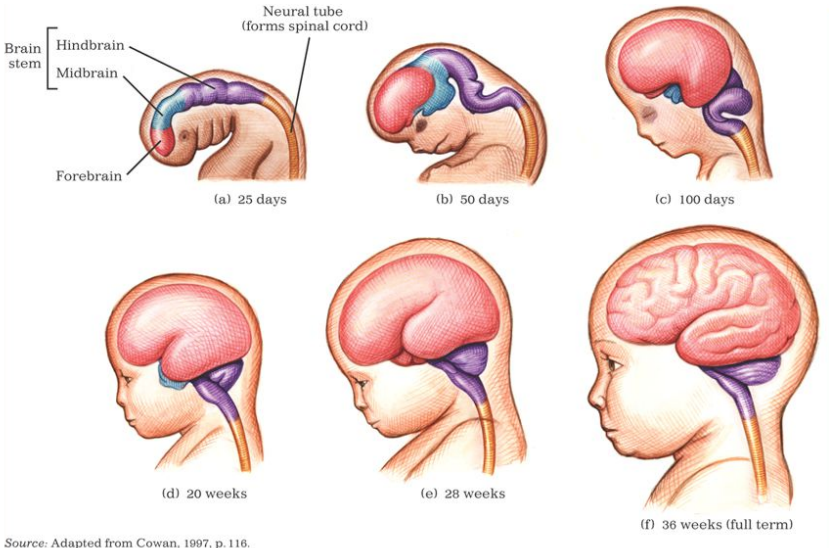
- versus key subject level covariates: case-control analysis
- versus gene lists for enrichment analysis
- versus genotype to discover eQTLs
- to obtain gene-gene correlation and networks

BrainSpan atlas of the developing human brain

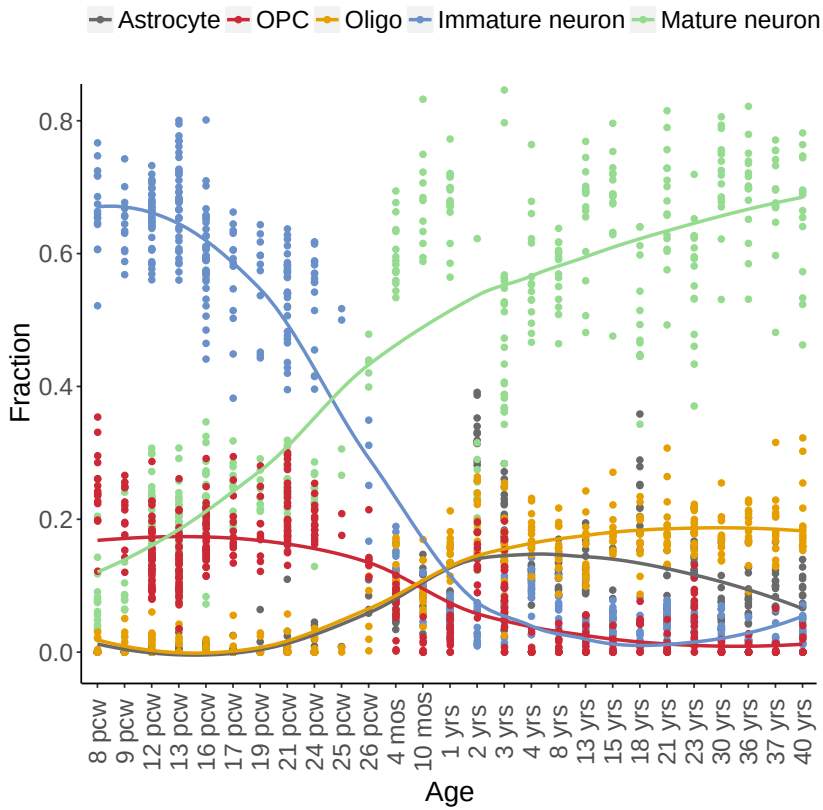


Source: Adapted from Cowan, 1997, p. 116.

BrainSpan atlas of the developing human brain



Source: Adapted from Cowan, 1997, p. 116.

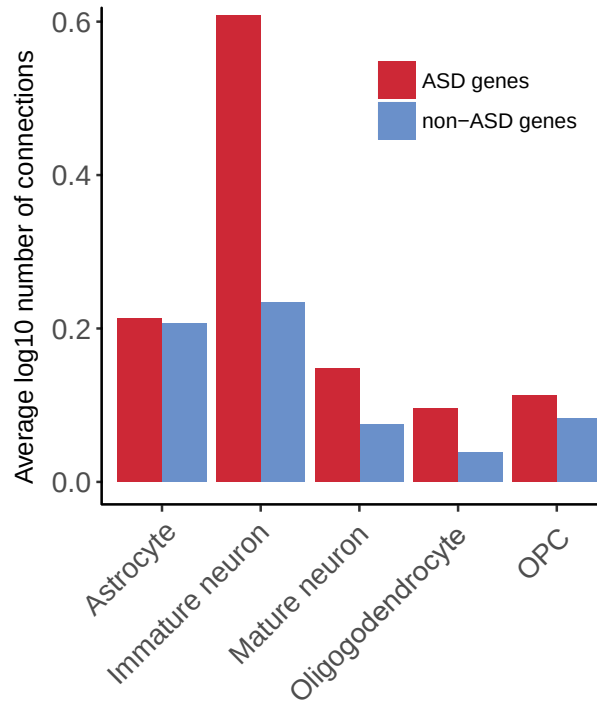


Case study: cell-type-specific co-expression network

- Gene expression correlation $\Rightarrow$ co-expression network
- Count number of connections per gene per cell type

Case study: cell-type-specific co-expression network

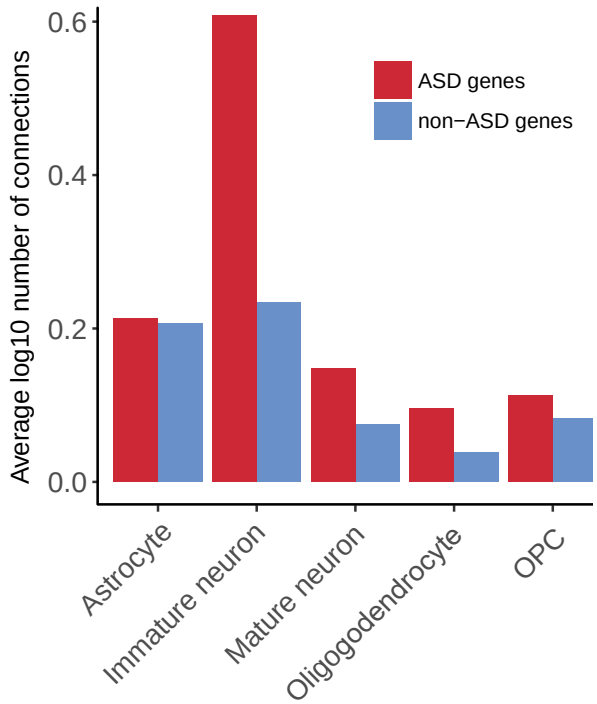
- Gene expression correlation $\Rightarrow$ co-expression network
- Count number of connections per gene per cell type
- ASD (autism spectrum disorder) genes have more connections than non-ASD genes in immature neurons



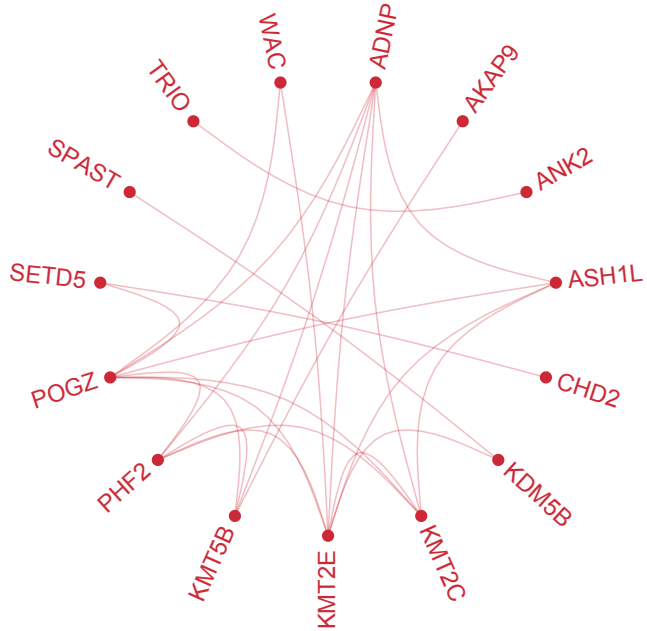
(Number of connections per gene)

Case study: cell-type-specific co-expression network

- Gene expression correlation $\Rightarrow$ co-expression network
- Count number of connections per gene per cell type
- ASD (autism spectrum disorder) genes have more connections than non-ASD genes in immature neurons

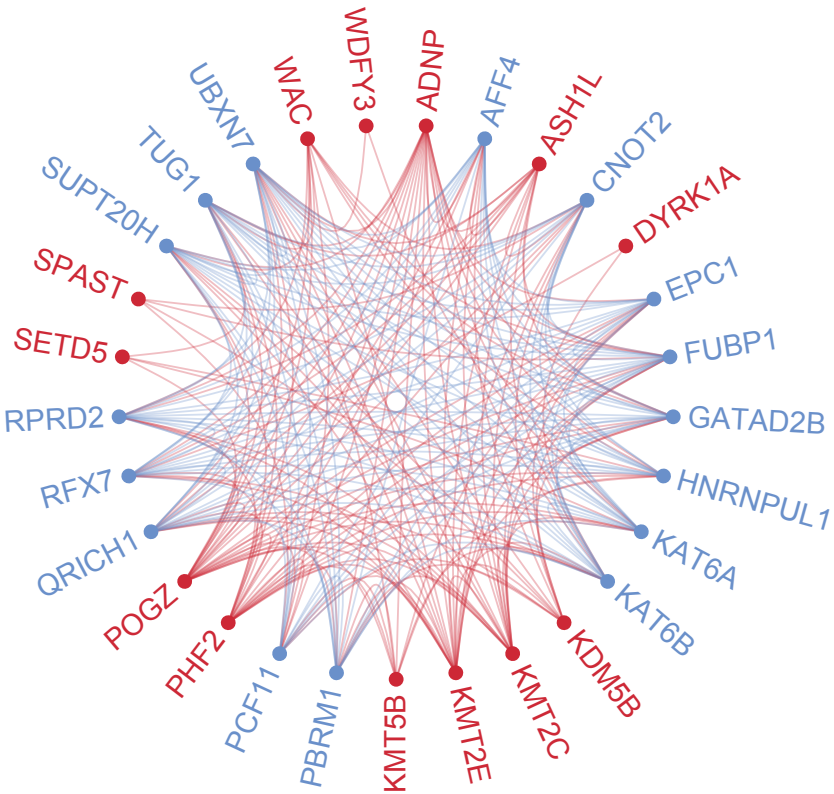


(Number of connections per gene)



(Network for ASD genes in immature neurons)

Case study: using MIND identifies new ASD genes



red: known ASD genes

blue: ASD-correlated genes identified based on MIND

- play regulatory roles
- are evolutionarily conserved (essential)
- are related to neurodevelopmental disorders

Larger Question for Progress

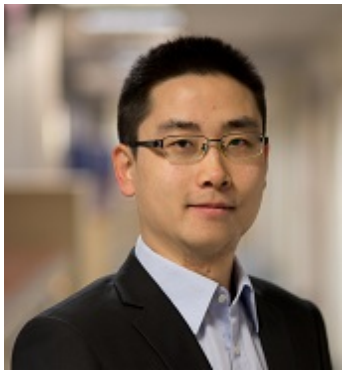
Seek gene-gene correlations computed by cell type

- Single cell data provides this, but the cells are from a very small number of tissue samples
- Deconvolved tissue samples can be obtained from hundreds of samples, but require at least 3 reps per sample
- Which variation is important for co-expression?
- Hard to determine which genes are co-expressed when the expressions are at the maximum of the range of the genes

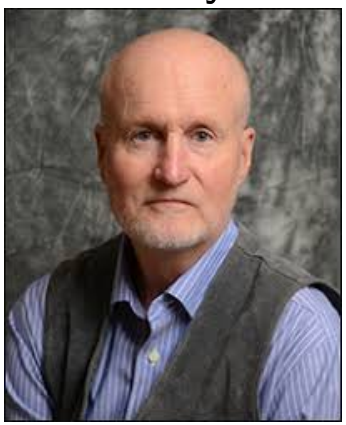
Can we combine information from both types of data to construct better gene networks?

Acknowledgements

Jiebiao Wang
Carnegie Mellon University



Bernie Devlin
University of Pittsburgh



NIH National Institute of Mental Health

SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE