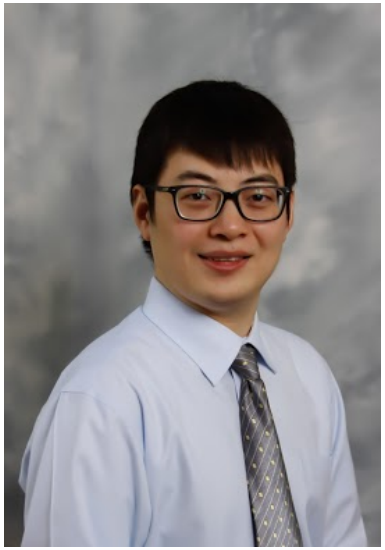
Tutorial T2 : Cell-type-aware Differential Analysis for Bulk Transcriptome Data

March 20, 2023 @ ENAR

Instructors



- Hao Feng
  Assistant Professor
  Department of Population and Quantitative Health Sciences
  Case Western Reserve University

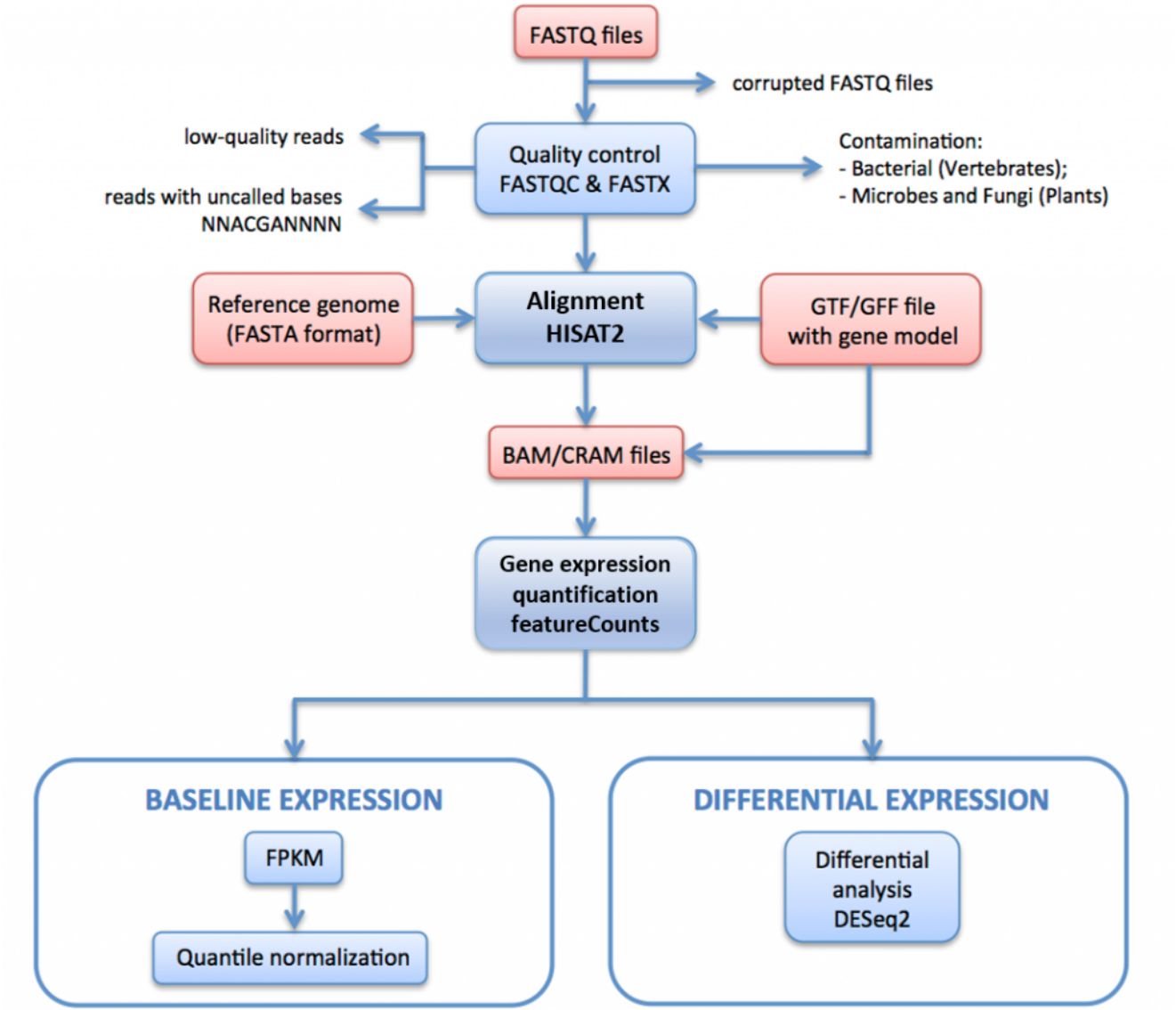
- Guanqun Meng
  PhD Student
  Case Western Reserve University

Tutorial session outline

1. Backgound in differential expression and deconvolution
2. Hands-on tutorial
   - TOAST, CellDMC, TCA, CARseq, DESeq2, CeDAR, LRCDE, csSAM
3. Methods comparison and conclusion

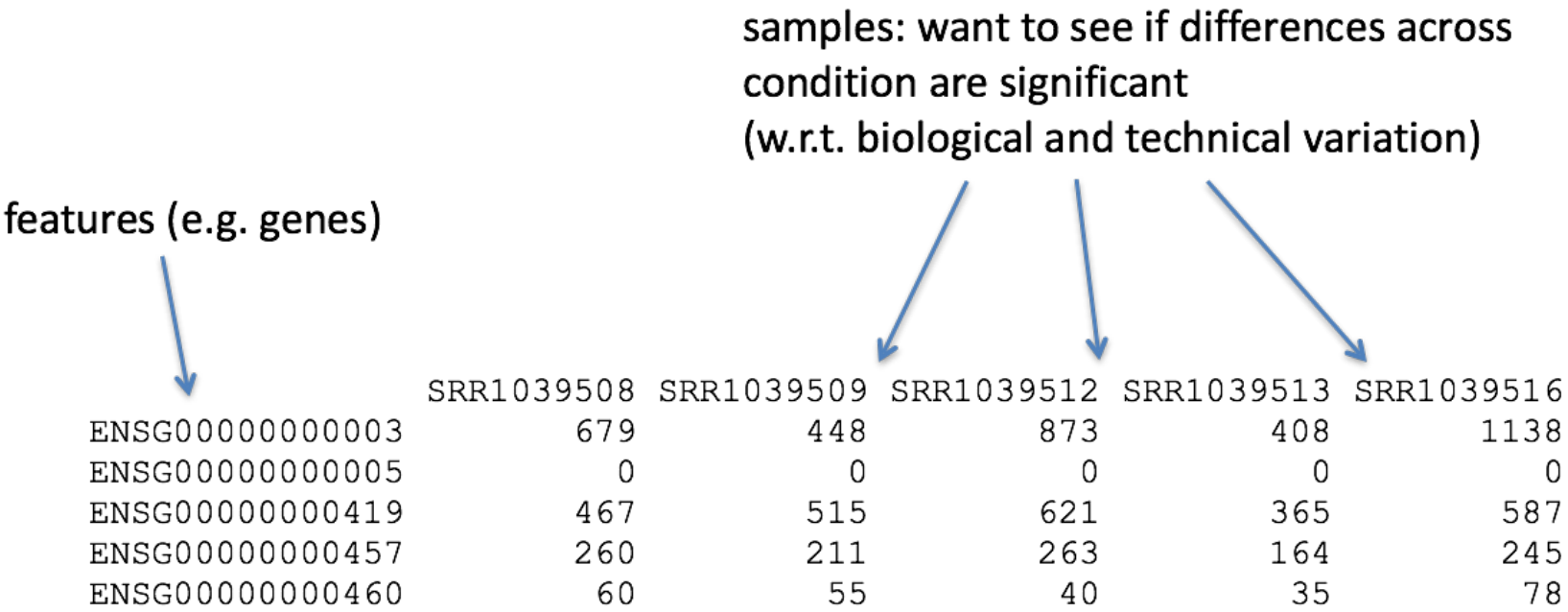
Backgound in differential expression and deconvolution

Transcriptome data processing



EMBL-EBI Training, Functional genomics II.

Differential gene expression analysis



	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

Harvard Chan Bioinformatics Core training modules. https://github.com/hbctraining

Differential gene expression analysis

Goal: find genes that are expressed differently between conditions.

1. Assign a score for each gene to represent its statistical significance of being different.
2. Rank the genes according to the score.
3. Find a proper threshold for the score for calling DE.

Easy solutions:

- Hypothesis testing (t-test, ANOVA, linear model, etc.) to get p-values and use as scores
- Use canonical cutoff (0.05) to call DE.

Potential problems

- Small sample size in hypothesis testing.
- Gene expression values are not necessarily Normally distributed.
- Multiple testing problem (p=0.05 cutoff).

Empirical Bayes method from *limma*

Smyth et al. (2004) Statistical Applications in Genetics and Molecular Biology

- Highly cited (>13,000 citations)
- Use a Bayesian hierarchical model in multiple regression setting.
- Borrow information from all genes to estimate gene specific variances.
    - As a result, variance estimates will be "shrunk" toward the mean of all variances. So very small variance scenarios will be alleviated.
- Implemented in Bioconductor package "limma".

Empirical Bayes method from *limma*

Let β_{gj} be the coefficient (difference in means in two-groups setting) for gene g, factor j, assume:

$$\hat{\beta}_{gj} \,|\, \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2) \qquad s_g^2 \,|\, \sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g} \quad \textbf{with priors:}$$

$$P(\beta_{gj} \neq 0) = p_j. \quad \beta_{gj} \,|\, \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2). \quad \frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}.$$

Posterior variance :

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}.$$

Moderated t-statistics for testing $\beta_{gj} = 0$:

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}.$$

RNA-seq differential expression using *DESeq2*

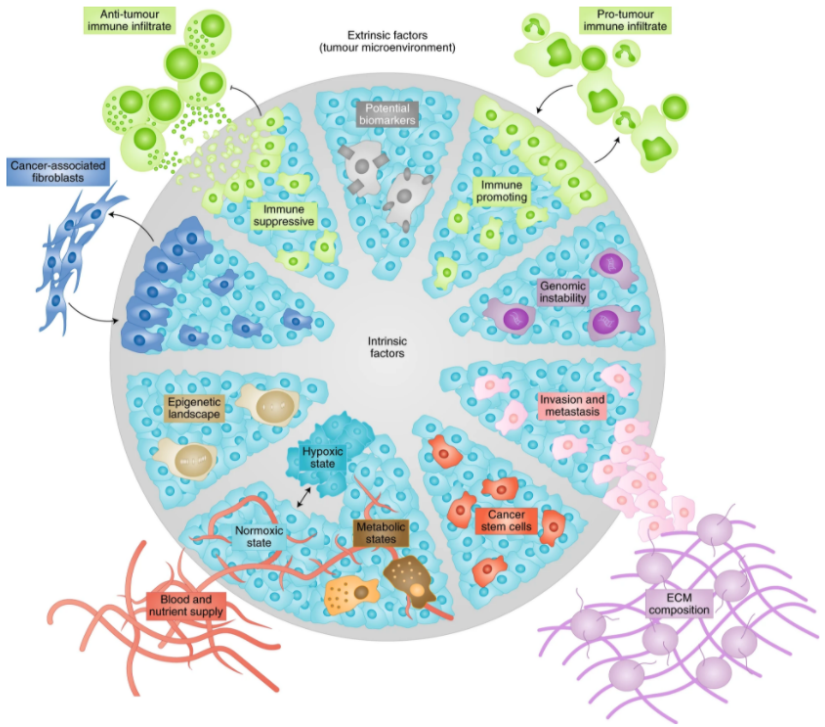
Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014).
Cited by $>$47,000
The read count K_{ij} for gene i in sample j, using GLM of NB family with a log link:

$$K_{ij} \sim \text{NB}(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_i)$$
$$\mu_{ij} = s_{ij} q_{ij}$$

$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}.$$

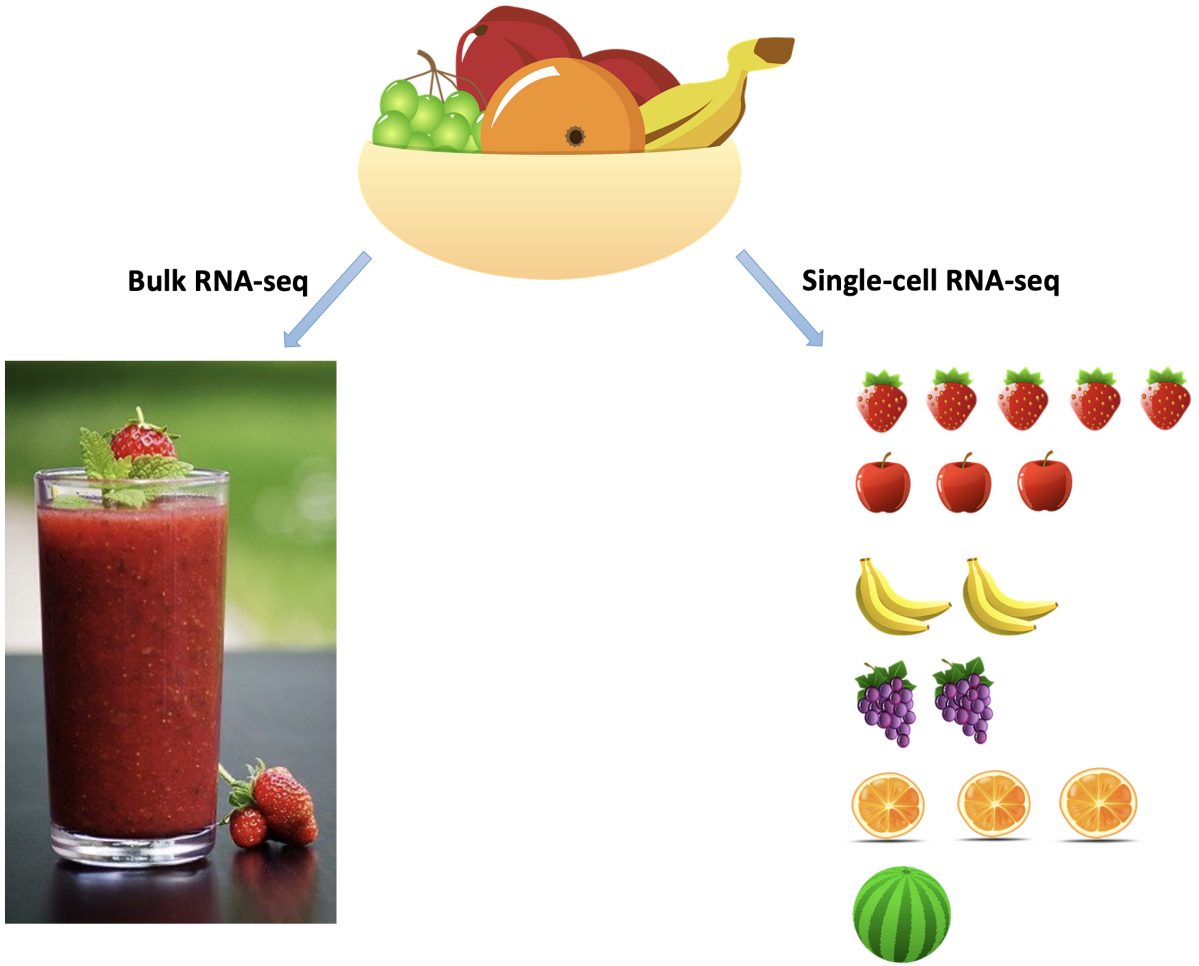
What was missing? heterogeneous mixture

- Human tissues are **heterogeneous**, as they have diverse cell types/states.
- Traditional RNA-seq ("bulk" RNA-seq) can measure **averaged signal** across millions of cells.

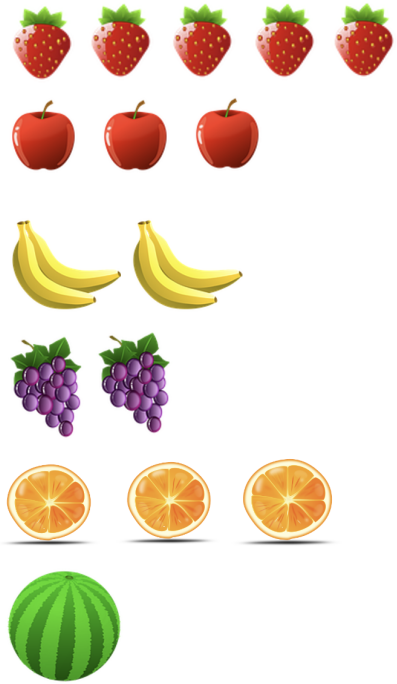


Lawson et al. Nature. https://www.nature.com/articles/s41556-018-0236-7

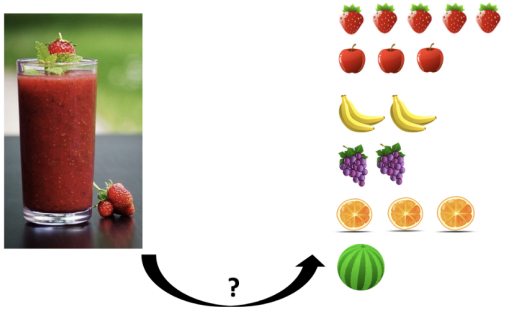
Bulk vs single-cell



?



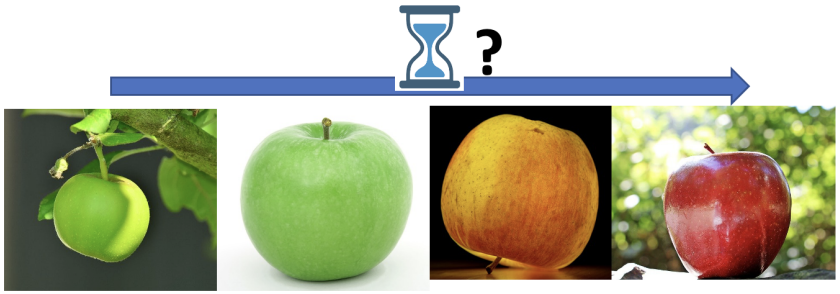
Deconvolution and beyond



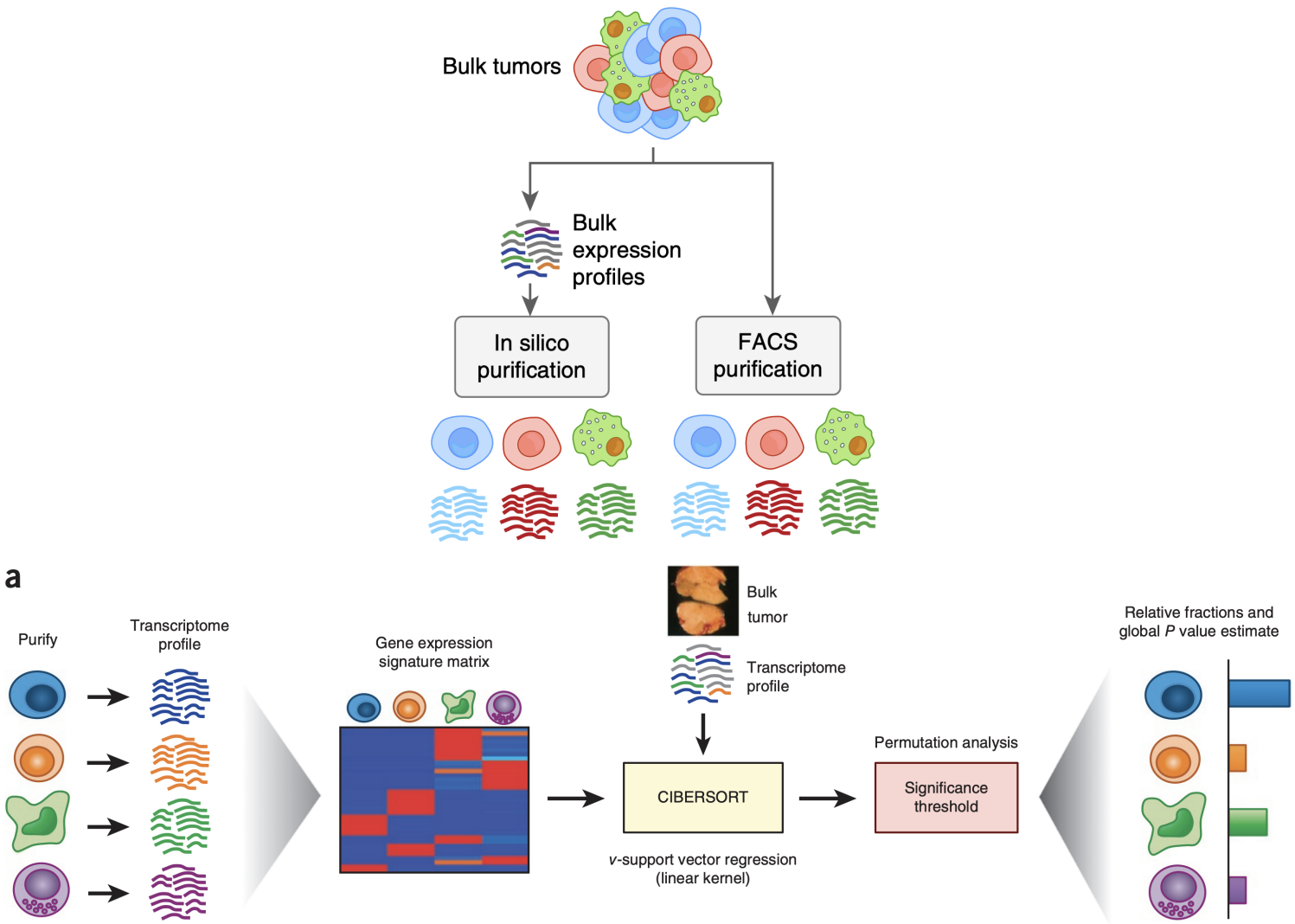
Smoothie A use:

?

Smoothie B use:

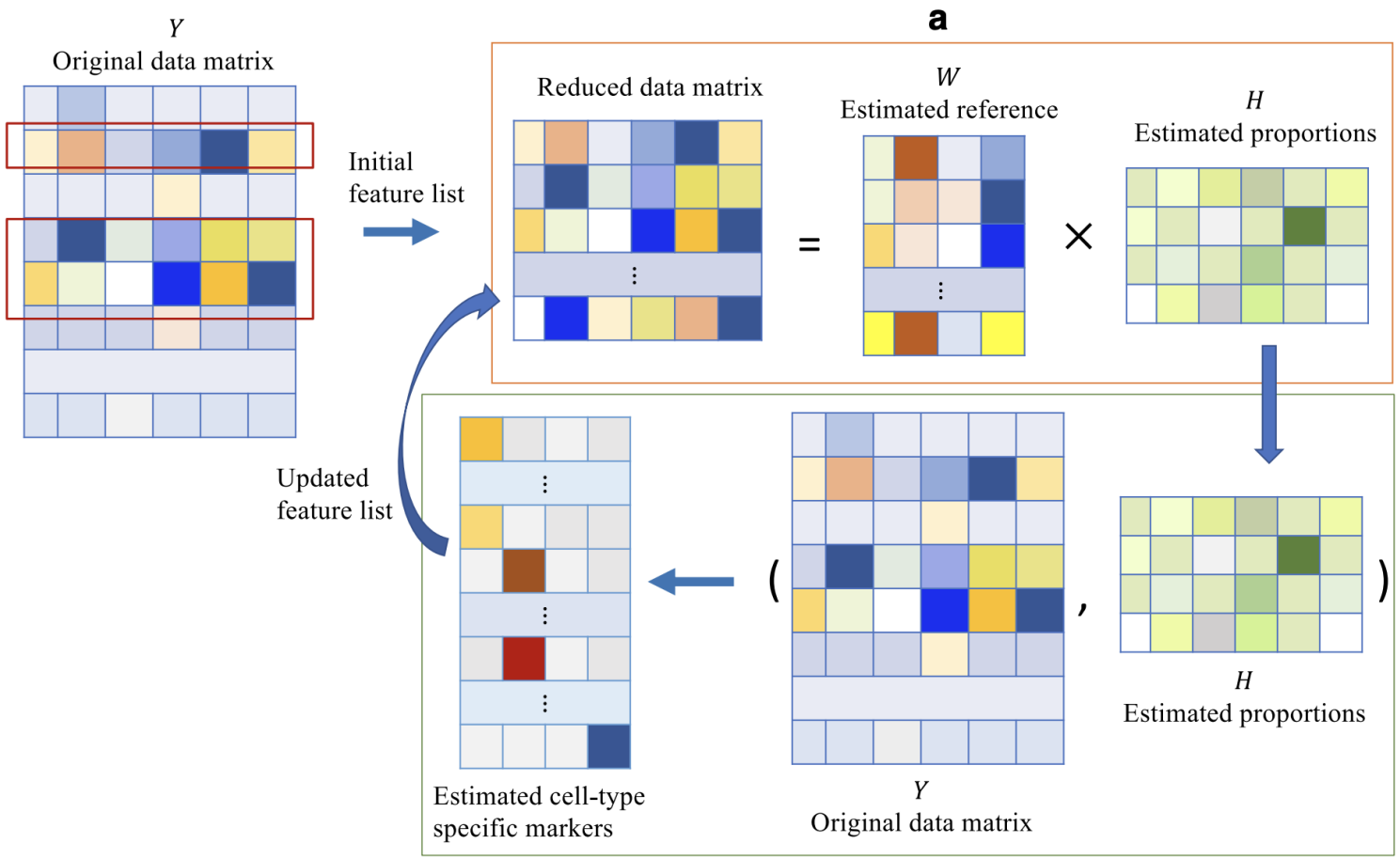


Cell composition of complex tissues



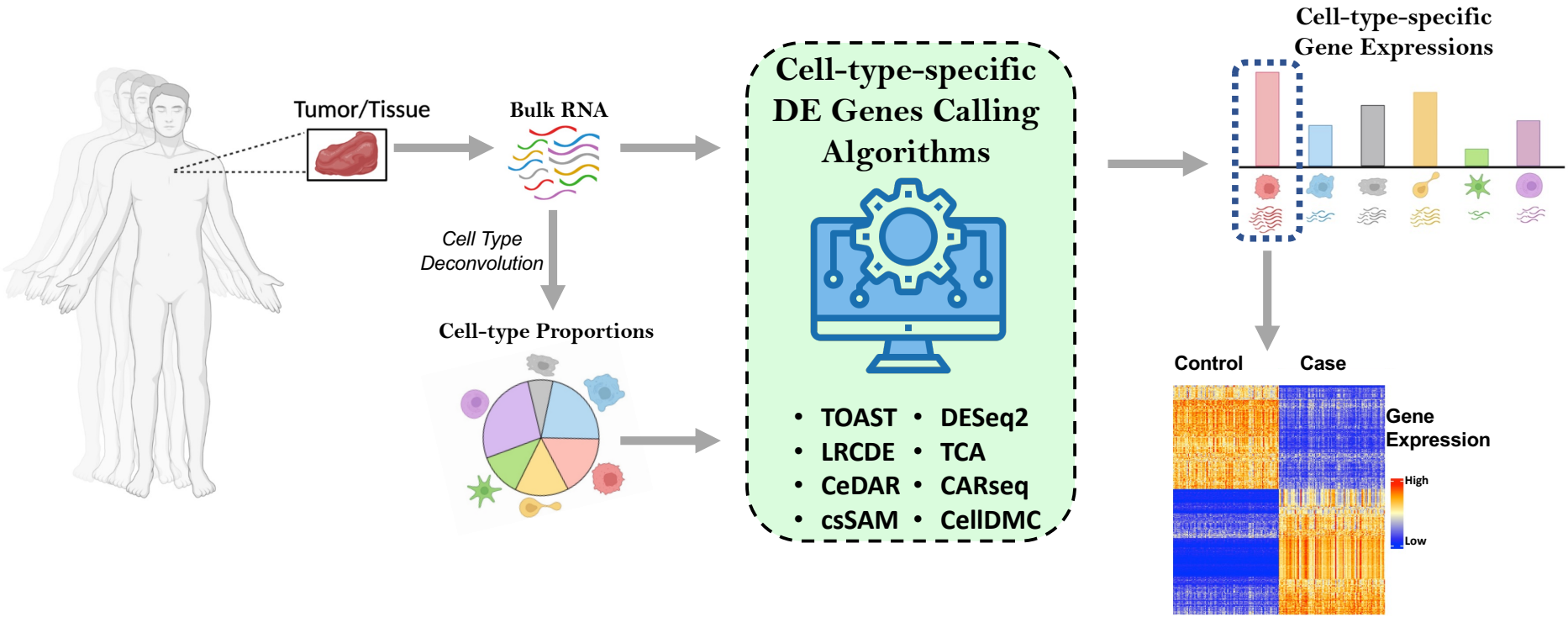
Newman et al. Nat Biotechnol. 2019; Newman et al. Nat Methods. 2015

Cell composition of complex tissues



Li et al. Genome Biology 2019

Cell-type-specific DE analysis



Meng et al. Briefings in Bioinformatics 2023

Cell-type-specific DE analysis

Method	Package/Year	Input	Algorithm
Cell type-specific Significance Analysis of Microarray (csSAM)	csSAM/ 2010	Gene expression microarray data	Linear regression; deconvolute cases and controls separately. Inferences of csDEG are based on *t*-statistics of permutation.
Differential gene expression based on NB[1] distribution (DESeq2)	*DESeq2/ 2014*	Gene expression RNA-seq data	Apply generalized NB[1] linear model and empirical Bayesian method to estimate the shrunk posterior of dispersion and LFC[2]. Adopt Wald tests under Normal distribution.
Linear Regression-based Cell type- specific Differential Expression (LRCDE)	lrcde/ 2016	General gene expression	Multivariate linear regressions: compare csDEG coefficients of different phenotypes. Inferences are based on two-sample *t*-test.
Identification of Differentially Methylated Cell types (CellDMC)	EpiDISH/ 2018	DNA methylation	Multivariate linear regression solved by LSE.
Tools for the Analysis of heterogeneouS Tissues (TOAST)	TOAST/ 2019	Gene expression and methylation data	Linear model framework: incorporate cell type proportions, phenotype information, and subject-specific covariates.
Tensor Composition Analysis (TCA)	TCA/ 2019	DNA methylation	Apply tensor to deconvolute 2D matrices into 3D tensors, which further allows statistical inference on variables of interest.
Cell type-aware Analysis of RNA-seq (CARseq)	CARseq/ 2021	Gene expression RNA-seq data	NB regression with parameters estimated iteratively by IWLS. Inferences based on likelihood ratio test.
CeDAR	TOAST/ 2022	Gene expression or methylation data	Stemmed from TOAST, further incorporating cell type DE/DM state correlations through hierarchical clustering.

Hands-on tutorial

- TOAST, CellDMC, TCA, CARseq, DESeq2, CeDAR, LRCDE, csSAM

See R markdown tutorial

Methods comparison and conclusion

Benchmark

OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–13

https://doi.org/10.1093/bib/bbac516

Review

A comprehensive assessment of cell type-specific differential expression methods in bulk data

Guanqun Meng, Wen Tang, Emina Huang, Ziyi Li and Hao Feng

Corresponding author. Hao Feng, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, E-mail: hxf155@case.edu

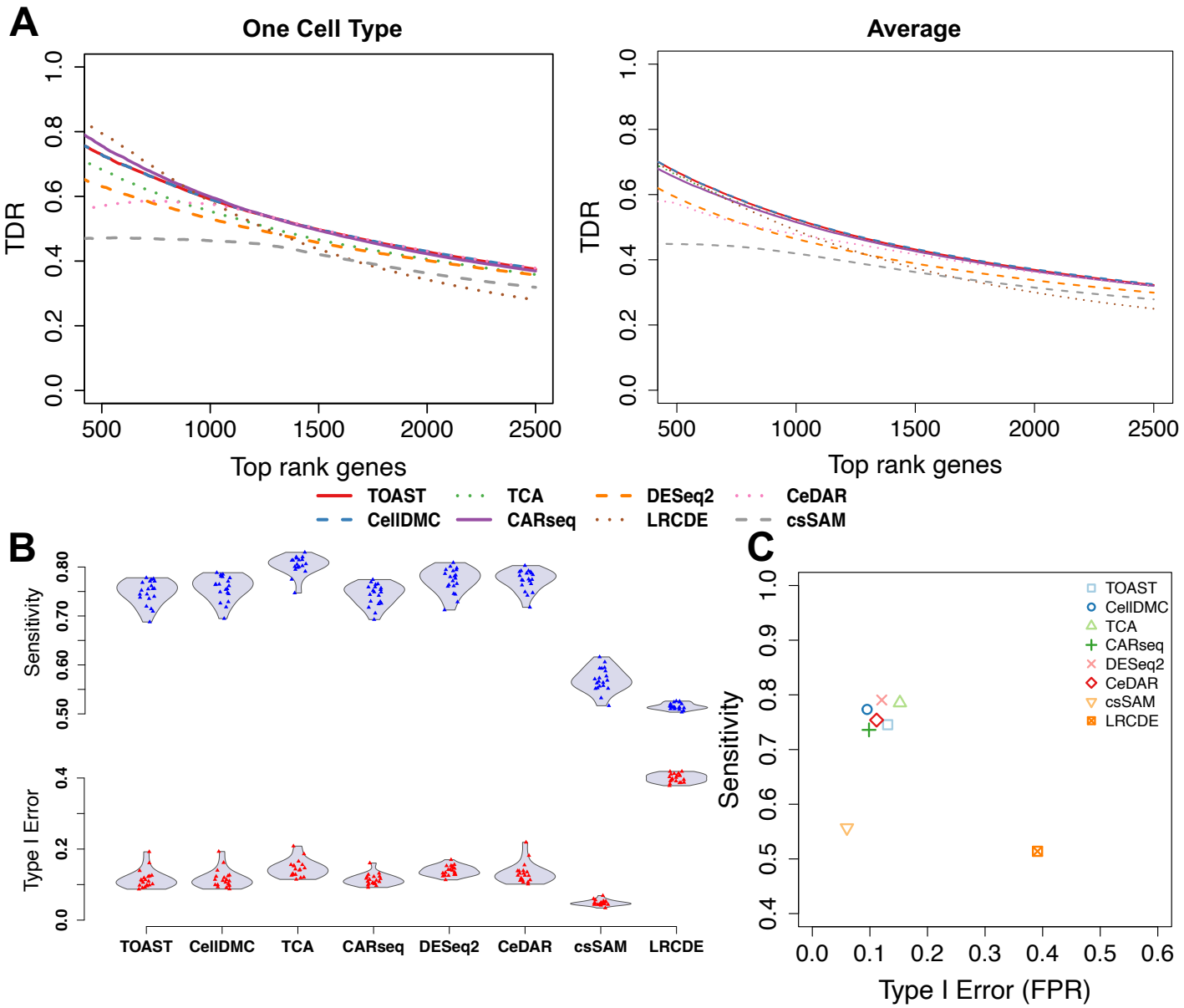
Simulation setup

1. $$\boldsymbol{\mu}_{g,K\times 1} \sim MVN(\hat{\boldsymbol{m}}, \hat{\boldsymbol{\Sigma}}_m)$$
$$\boldsymbol{\phi}_{g,K\times 1} \sim MVN(\hat{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_d)$$
2. $$\boldsymbol{M}_{G\times K} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_G]^T; \boldsymbol{\Phi}_{G\times K} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_G]^T$$
3. $$\boldsymbol{X}_{G\times K} \sim Gamma\{shape = \frac{1}{\exp(\boldsymbol{\Phi})}, scale = \exp(\boldsymbol{M}) \cdot \exp(\boldsymbol{\Phi})\}$$
4. $$\boldsymbol{\theta}_i \sim Dir(\boldsymbol{\alpha})$$
5. $$\boldsymbol{r}_i = \boldsymbol{X}\boldsymbol{\theta}_i$$
$$\boldsymbol{y_i}|\boldsymbol{r}_i \sim Poisson(\boldsymbol{r}_i)$$

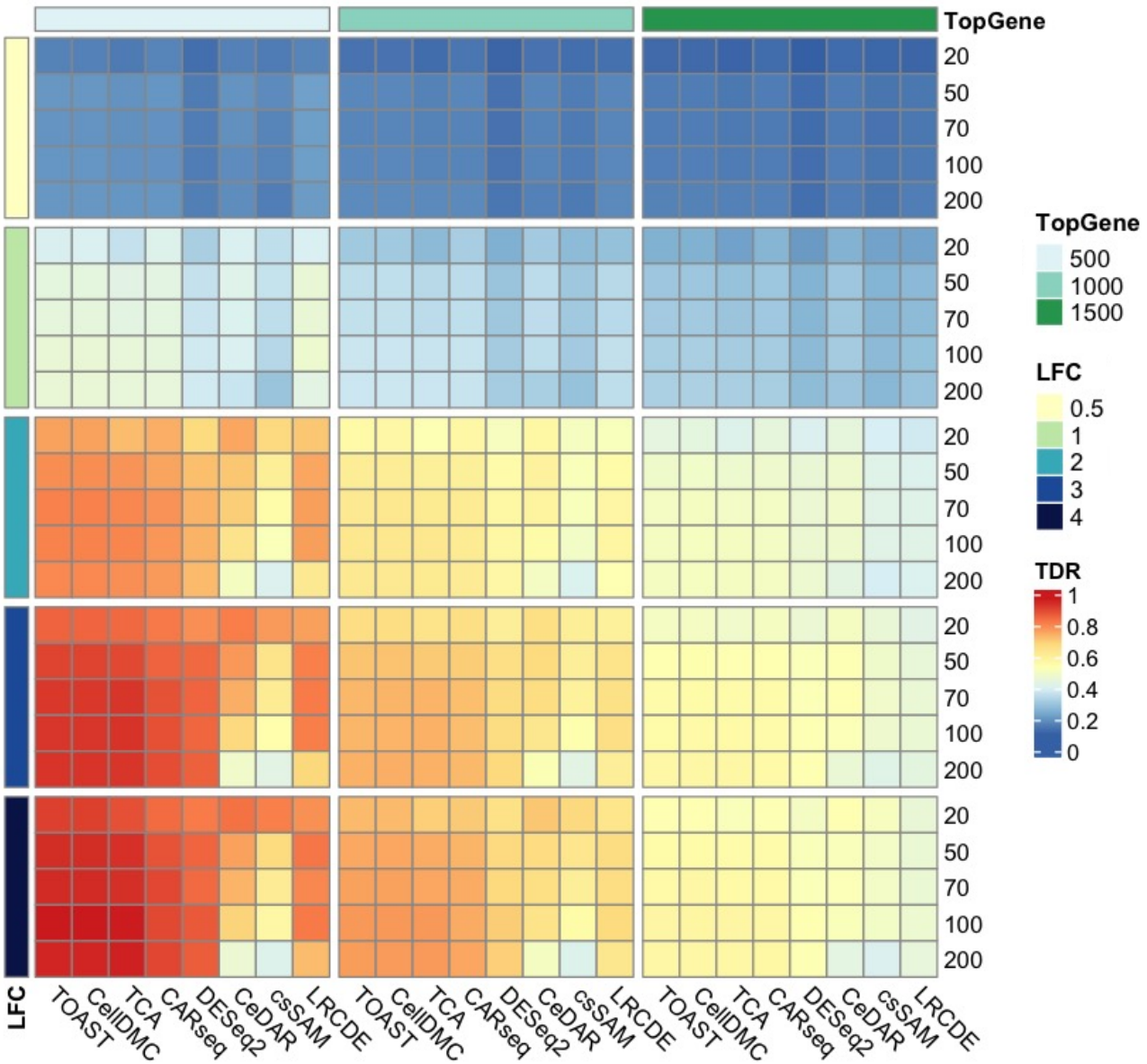
Simulation setup

- N = 50, 100, 150, 200
- LFC = 0(null), 0.5, 0.75, 1.0, 1.25, 1.5.
- 10% or 0%(null) csDEG.
- 6 cell types
- Reference panel generated from real bulk cell line.
- Proportions from Dirichlet with parameters from scRNA-seq data.
- Gamma-Poisson for observed counts.

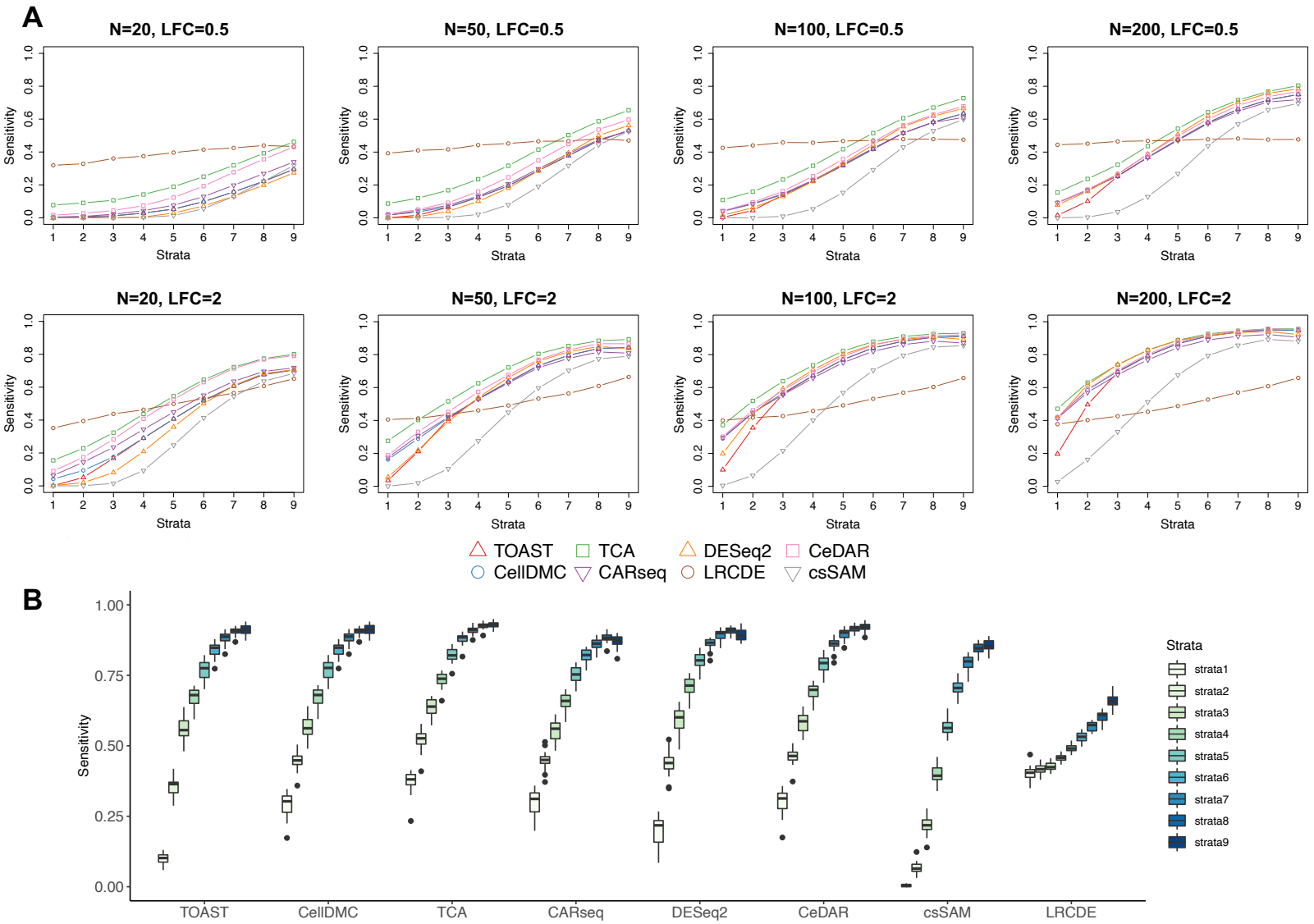
Comparisons of csDEG detection accuracy



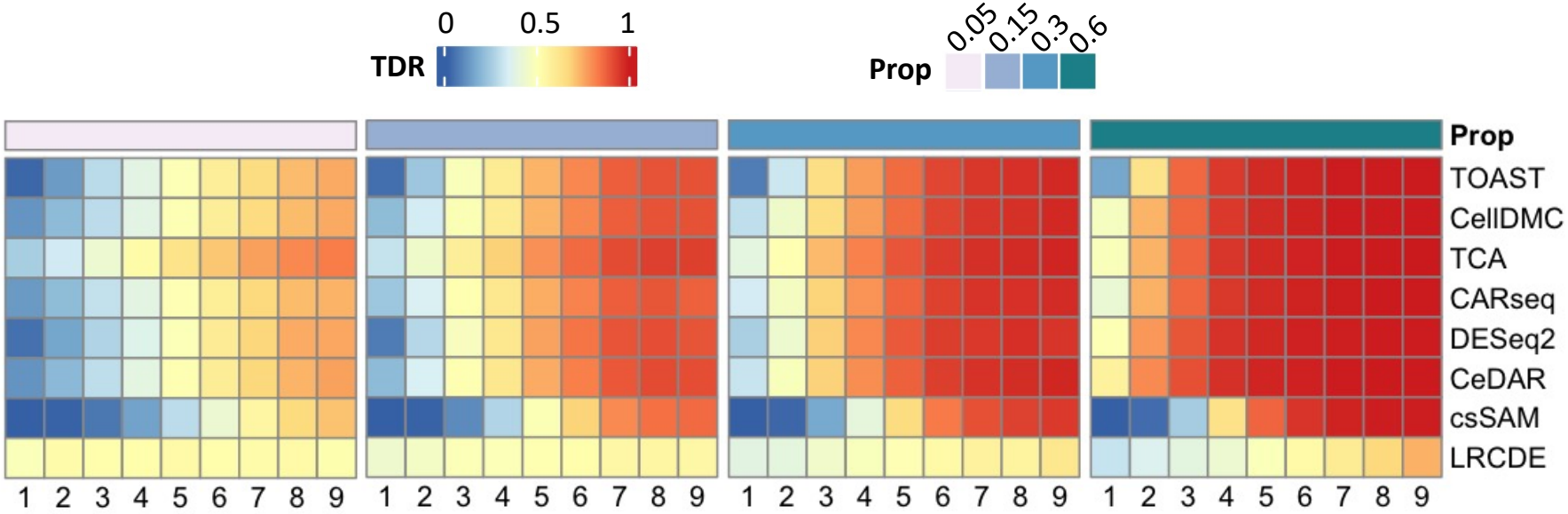
Precision at various N and LFC



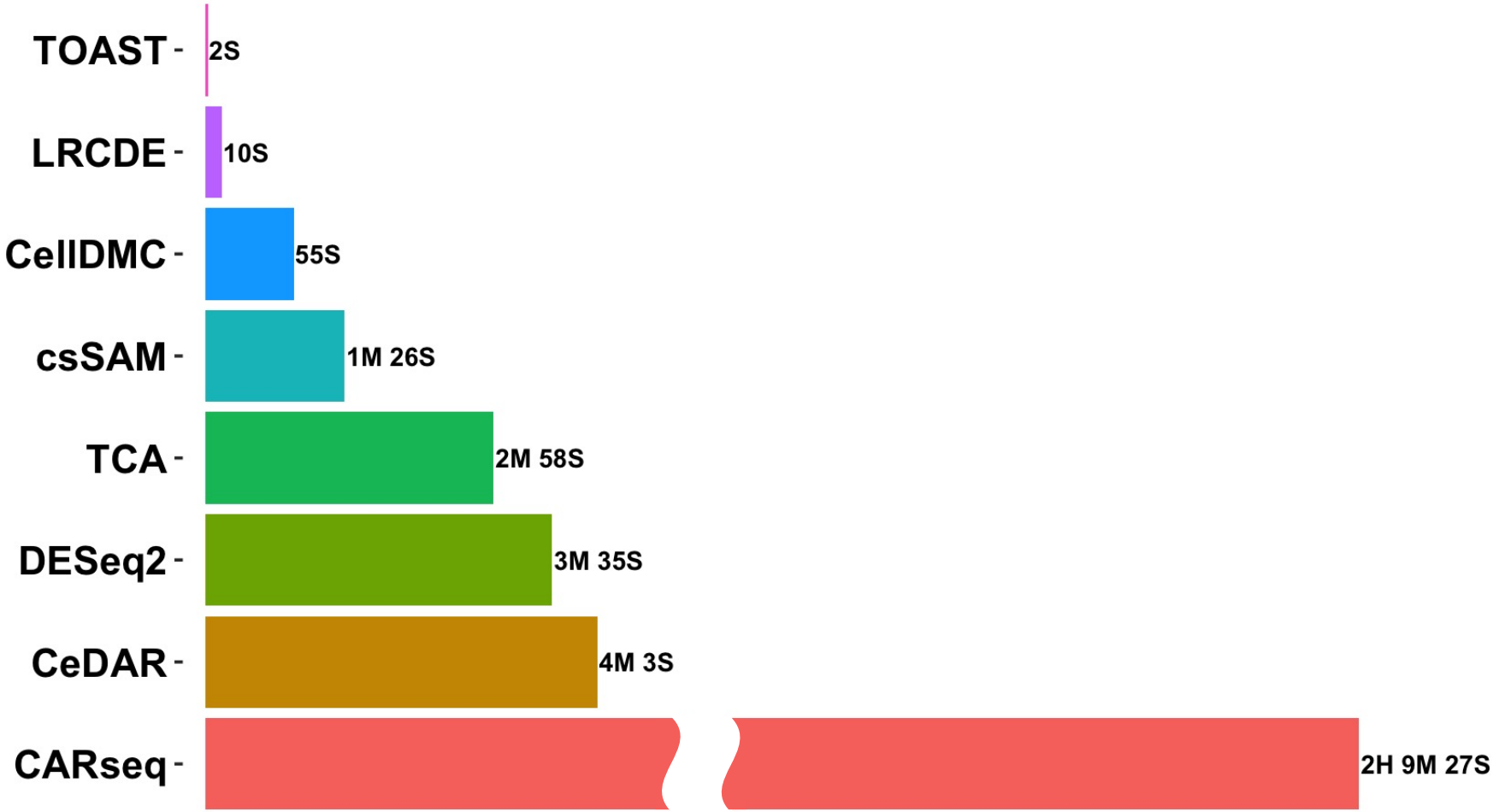
Expression stratification



Impact of cell type proportions



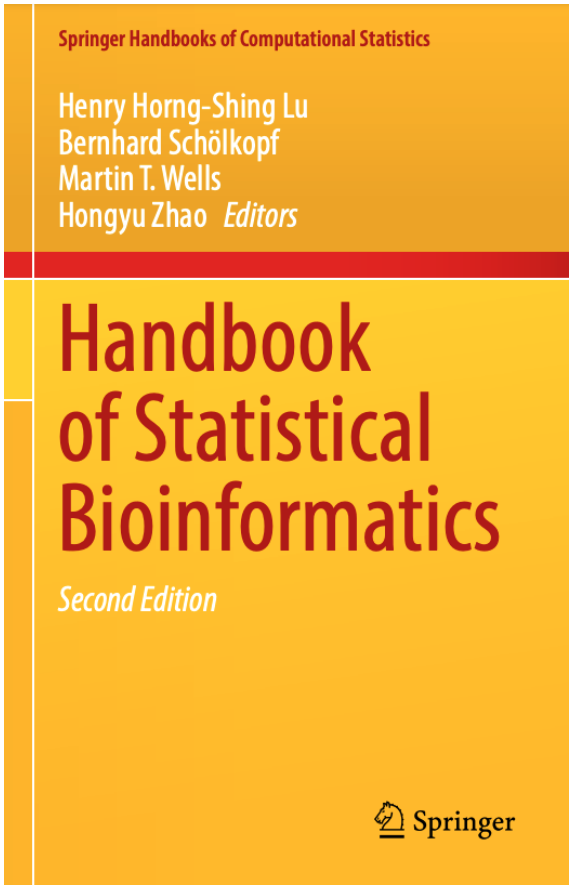
Software runtime



Summary

- Cell type-specific differentially expressed genes (csDEG) analysis is successful at dissecting bulk RNA-seq data and identifying biomarkers in a finer resolution.
- Effect size, baseline expression level and cell type composition are the leading factors affecting csDEG calling accuracy.
- CARseq, TOAST, CellDMC and TCA are the most reliable methods in terms of precision and sensitivity.
- Insufficient power can be expected for low expression genes. Larger sample size is needed compared with traditional DE analysis.
- csDEG is a challenging task itself, with room to improve to properly handle low signal-to-noise ratio and low expression genes.

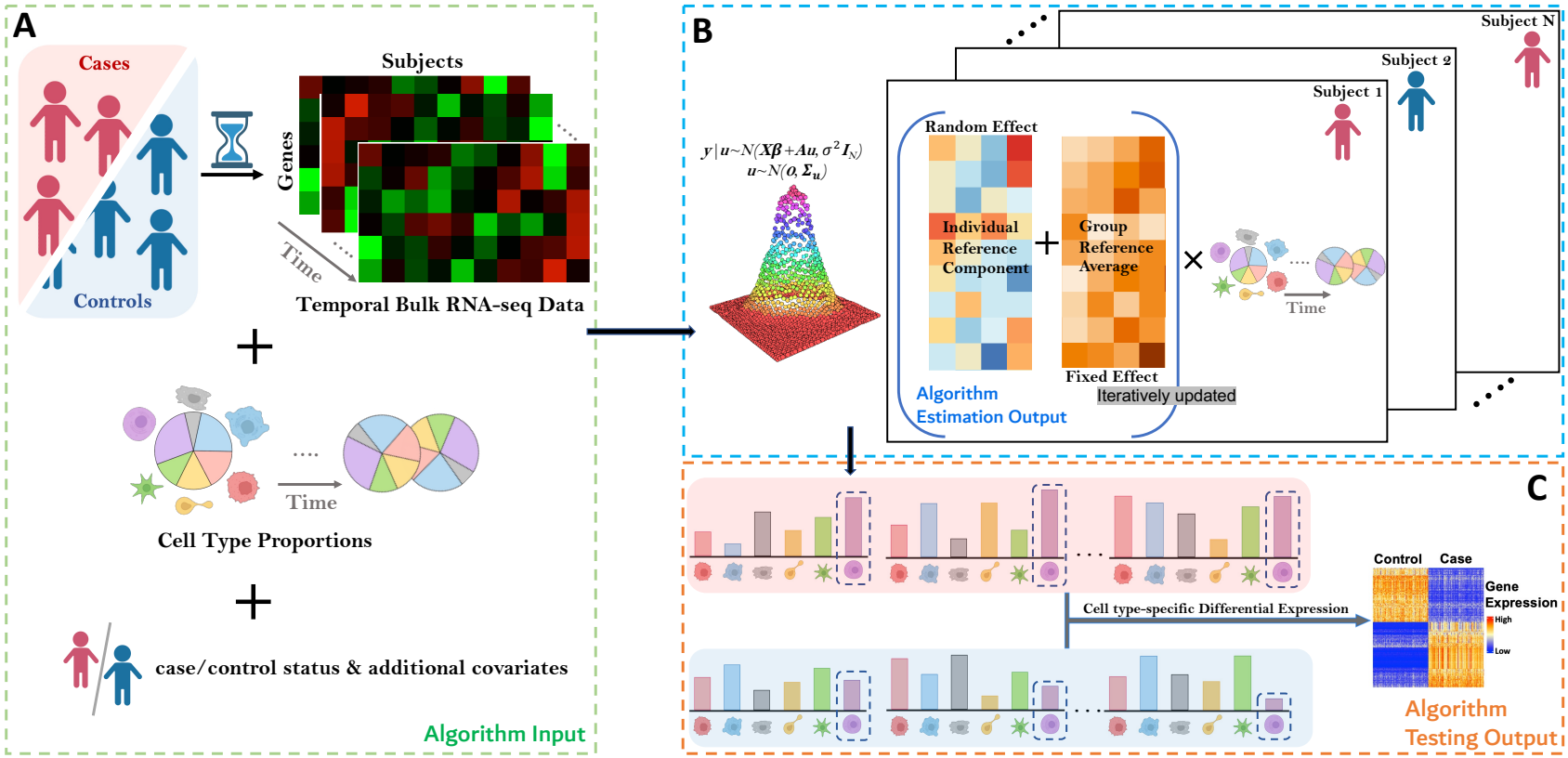
Additional Resources



Part III: *Cell Type-Specific Analysis for High-throughput Data*. Covers tools TOAST, CellMix, EpiDISH, RefFreeEWAS, and MuSiC.

ISLET: **I**ndividual-**S**pecific Ce**L**l Typ**E** Referencing **T**ool

Wednesday, March 22. session 95. 9:15 am – 9:30 am



Decomposing Admixed Genomics Data: Cell-type-aware Analysis Methodology Advances

Chair & Organizer: Hao Feng, Case Western Reserve University

Speakers:

Aaron Newman, Stanford University

Stephanie Hicks, Johns Hopkins Bloomberg School of Public Health

Wenyi Wang, The University of Texas MD Anderson Cancer Center

Rafael Irizarry, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health.

Tuesday, March 21. 8:30 am – 10:15 am