Cell type expression phenotyping from deconvolution of bulk RNAseq data

Christopher K Tuggle

Department of Animal Science

Iowa State University

AG2PI Field Day

March 12, 2025

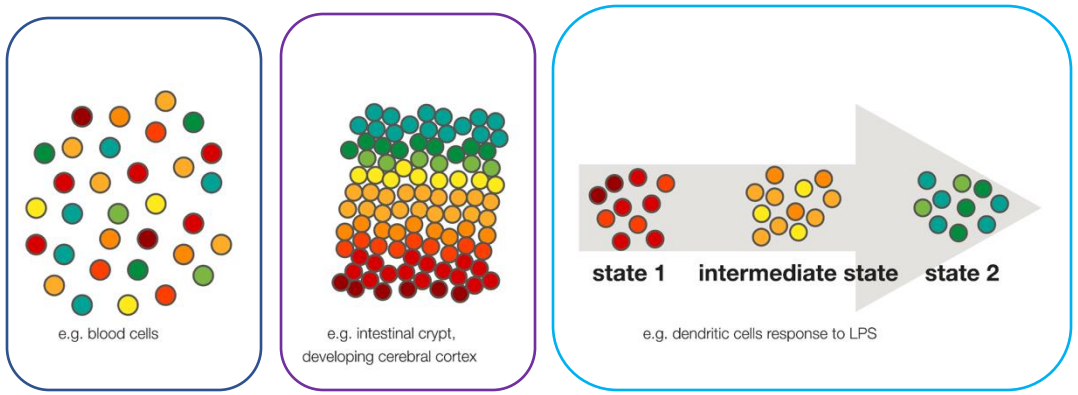
NIFA Project # 2022-67015-37055

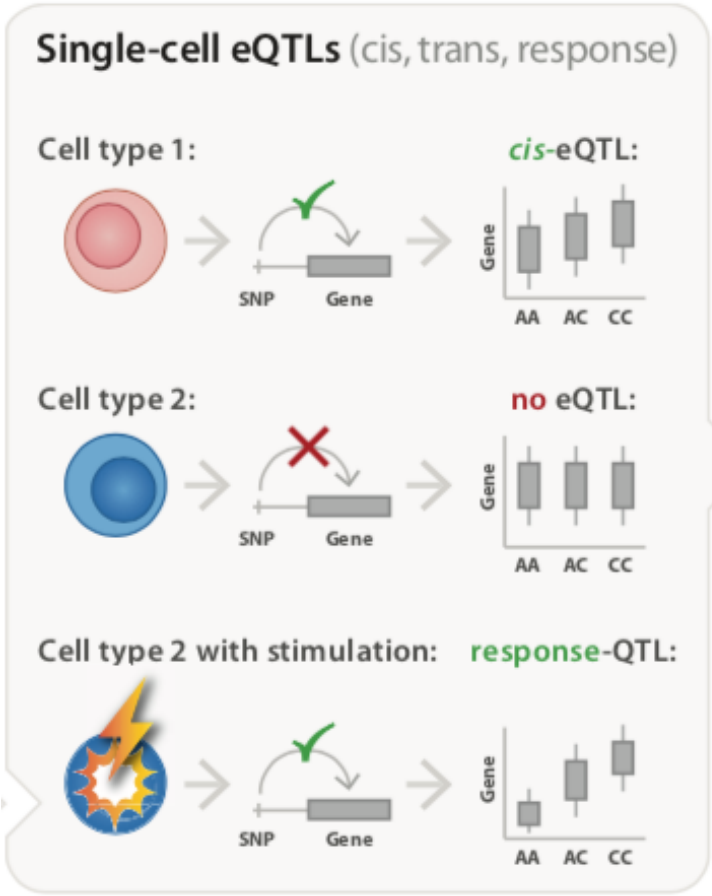
USDA United States Department of Agriculture National Institute of Food and Agriculture

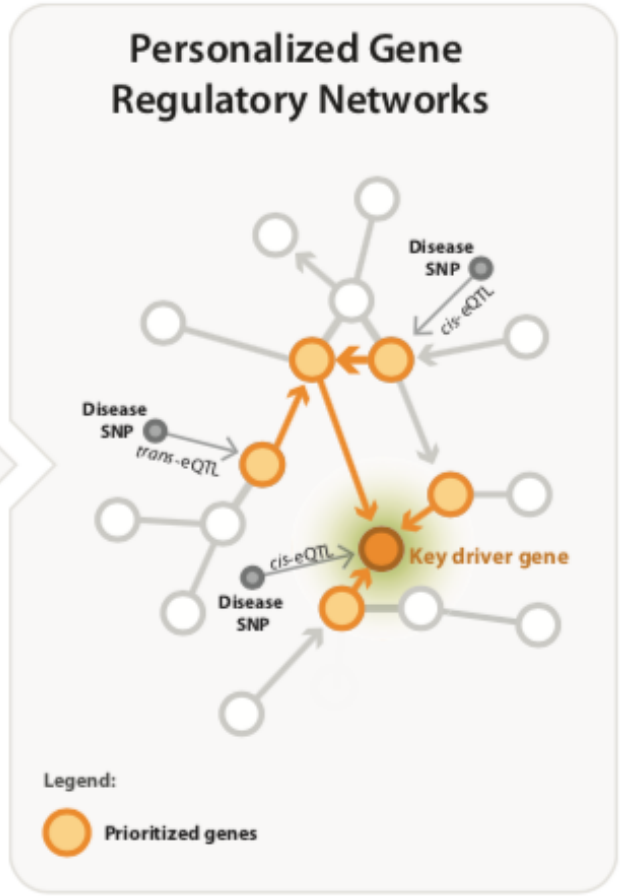
FAANG Functional Annotation of Animal Genomes

Value of single cell transcriptomics for biology and genetics

Heterogeneous cell populations in tissues- analyses of "bulk" tissue misses a lot….





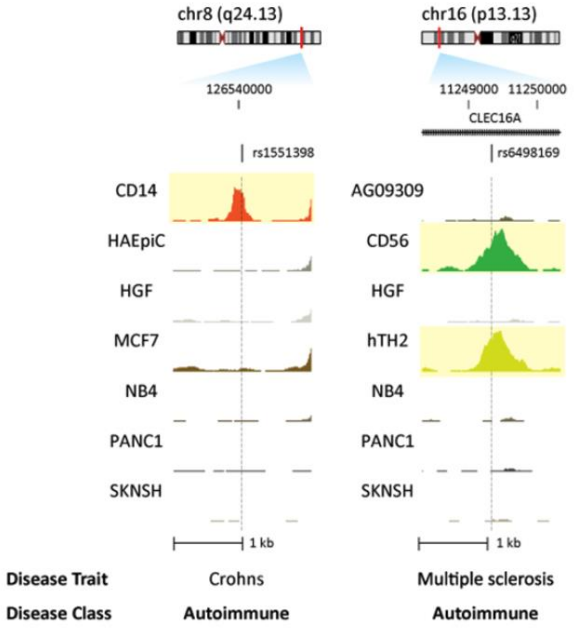


(Kolodziejczyk et al. 2015. Mol Cell. 2015 58(4):610-20)

van der Wijst et al. eLife 2020;9:e52155. DOI: https://doi.org/10.7554/eLife.52155

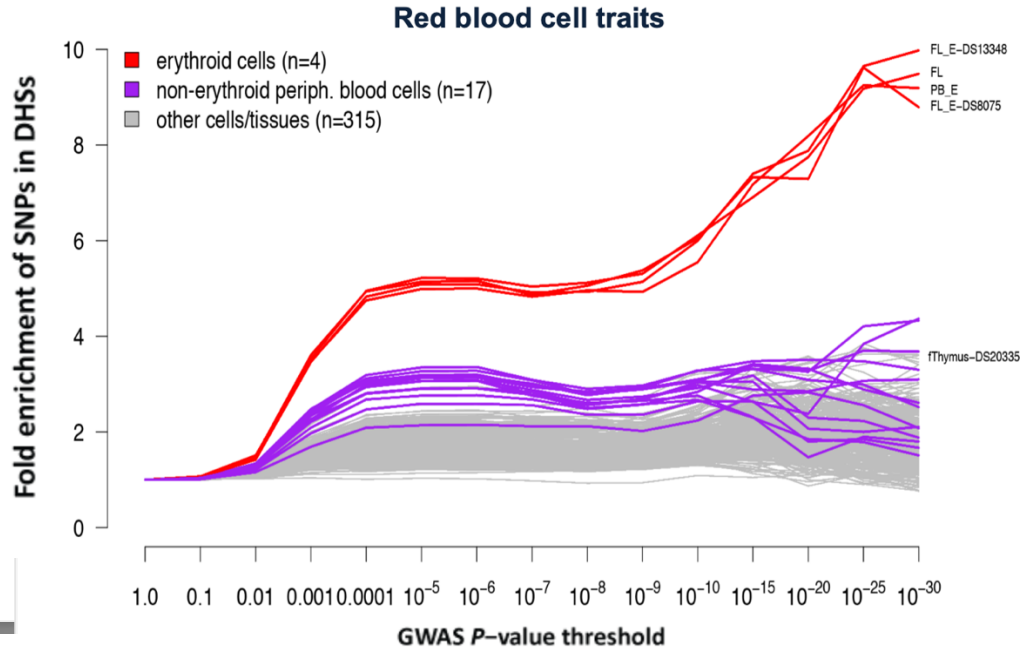
ENCODE project showed enrichment of trait-associated SNPs is *biological context specific.*

Disease associated variants often overlap with regulatory elements specific for target cell/tissue pathology



Maurano et al., Science 2012

Cell-selective enrichment of trait-associated variants



We need to test using context-specific information!

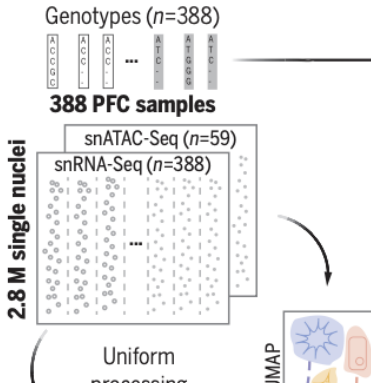
Linking to Breeding Goals:
Finding cell-type specific eQTL in tissues/cell mixes

RESEARCH ARTICLE SUMMARY

PSYCHENCODE2

Single-cell genomics and regulatory networks for 388 human brains

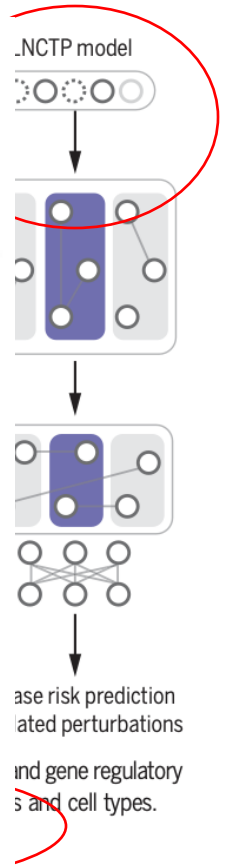
Emani *et al.*, *Science* **384**, 862 (2024) 24 May 2024



Prashant S. Emani[1,2]†, Jason J. Liu[1,2]†, Declan Clarke[1,2]†, Matthew Jensen[1,2]†, Jonathan Warrell[1,2]†, Chirag Gupta[3,4]†, Ran Meng[1,2]†, Che Yu Lee[5]†, Siwei Xu[5]†, Cagatay Dursun[1,2]†, Shaoke Lou[1,2]†, Yuhang Chen[1,2], Zhiyuan Chu[1], Timur Galeev[1,2], Ahyeon Hwang[5,6], Yunyang Li[2,7], Pengyu Ni[1,2], Xiao Zhou[1,2], PsychENCODE Consortium‡, Trygve E. Bakken[8], Jaroslav Bendl[9,10,11,12], Lucy Bicks[13], Tanima Chatterjee[1,2], Lijun Cheng[14], Yuyan Cheng[13,15], Yi Dai[5], Ziheng Duan[5], Mary Flaherty[14], John F. Fullard[9,10,11,12], Michael Gancz[1,2], Diego Garrido-Martín[16], Sophia Gaynor-Gillett[14,17], Jennifer Grundman[13], Natalie Hawken[13], Ella Henry[1,2], Gabriel E. Hoffman[9,10,11,12,18,19], Ao Huang[1], Yunzhe Jiang[1,2], Ting Jin[3,4], Nikolas L. Jorstad[8], Riki Kawaguchi[13,20], Saniya Khullar[3,4], Jianyin Liu[13], Junhao Liu[5], Shuang Liu[4], Shaojie Ma[21,22], Michael Margolis[13], Samantha Mazariegos[13], Jill Moore[23], Jennifer R. Moran[14], Eric Nguyen[1,2], Nishigandha Phalke[23], Milos Pjanic[9,10,11,12], Henry Pratt[23], Diana Quintero[13], Ananya S. Rajagopalan[7], Tiernon R. Riesenmy[24], Nicole Shedd[23], Manman Shi[14], Megan Spector[14], Rosemarie Terwilliger[25], Kyle J. Travaglini[8], Brie Wamsley[13], Gaoyuan Wang[1,2], Yan Xia[1,2], Shaohua Xiao[13], Andrew C. Yang[1,2], Suchen Zheng[1,2], Michael J. Gandal[26,27,28,29,30], Donghoon Lee[9,10,11,12], Ed S. Lein[8,31,32], Panos Roussos[9,10,11,12,18,19], Nenad Sestan[21], Zhiping Weng[23], Kevin P. White[33], Hyejung Won[34], Matthew J. Girgenti[25,35,36]*, Jing Zhang[5]*, Daifeng Wang[3,4,37]*, Daniel Geschwind[13,20,27,28,38]*, Mark Gerstein[1,2,7,24,39]*

ACKNOWLEDGMENTS

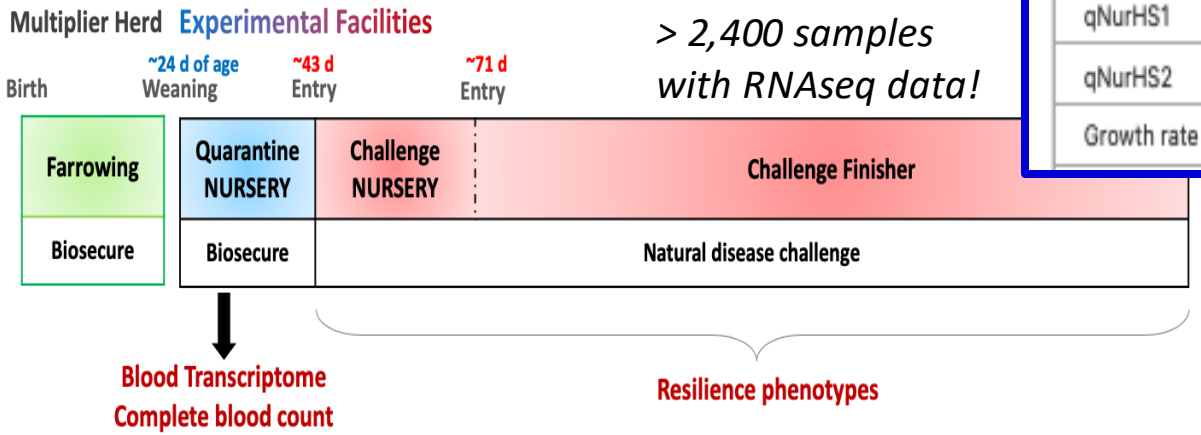
The authors thank the founder of the Allen Institute, P. G. Allen, for his vision, encouragement, and support. R.T. and M. J. Girgenti thank Keck Microarray Shared Resource (KMSR) and Yale Center for Genome Analysis (YCGA) at Yale University for their assistance with 10x Genomics single-cell RNA-seq services. **Funding:** Data were generated as part of the PsychENCODE Consortium, supported by U01DA048279, U01MH103339, U01MH103340, U01MH103346, U01MH103365, U01MH103392, U01MH116438, U01MH116441, U01MH116442, U01MH116488, U01MH116489, U01MH116492, U01MH122590, U01MH122591, U01MH122592, U01MH122849, U01MH122678, U01MH122681, U01MH116487, U01MH122509, R01MH094714, R01MH105472, R01MH105898, R01MH109677, R01MH109715, R01MH110905, R01MH110920, R01MH110921, R01MH110926, R01MH110927, R01MH110928, R01MH111721, R01MH117291, R01MH117292, R01MH117293, R21MH102791, R21MH103877, R21MH105853, R21MH105881, R21MH109956, R56MH114899, R56MH114901, R56MH114911, R01MH125516, R01MH126459, R01MH129301, R01MH126393, R01MH121521, R01MH116529, R01MH129817, R01MH117406, and P50MH106934 awarded to authors and collaborators A. Abyzov,



Comprehensive mining of the blood cell transcriptome for improved phenomics in swine

Rationale: Blood sampling is very practical and useful phenotype- but needs improvement to use in predictive genetics- a mixture of genetic effects on phenotypes

Whole blood RNA patterns have been associate disease phenotypes in pigs in Resilience project

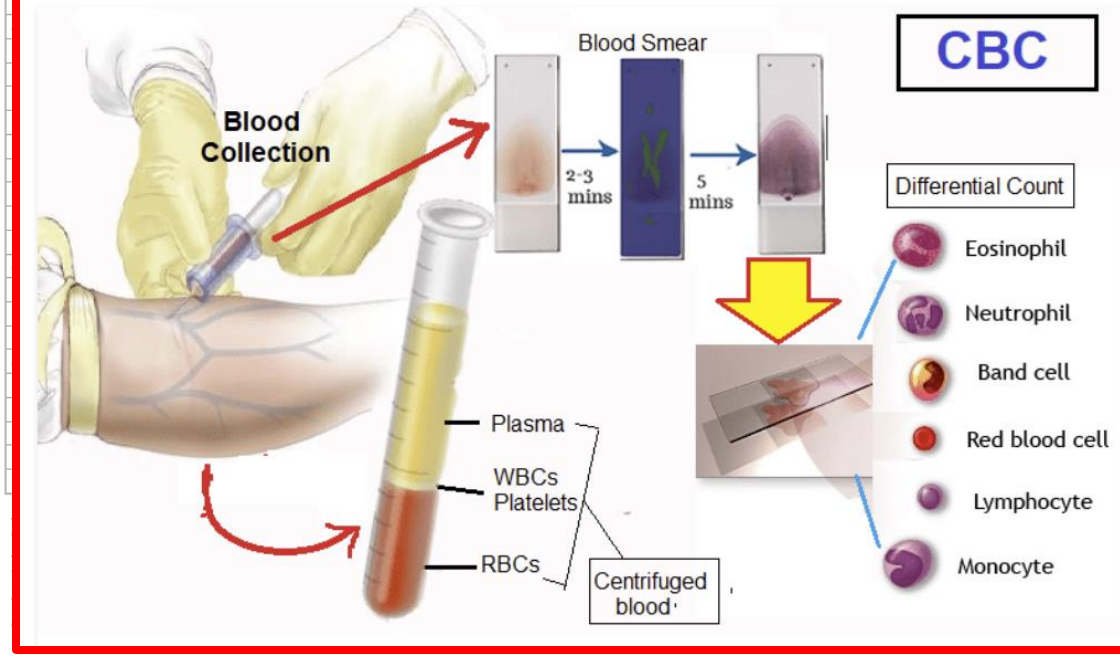


> 2,400 samples with RNAseq data!

Resilience Project natural disease challenge model (NDCM)

Table 3 The number of genes with expression levels in blood of young healthy pigs that were significantly ($q < 0.20$) associated with observed phenotypes, with or without accounting for blood cell composition, the

Trait measured during each phase [a]	Number of genes from expression residuals with or without adjustment for cell composition		
	Without	With	With+Without [b]
Quarantine Nursery			
qNurHS1	395 (1656) [c]	29 (1106)	395
qNurHS2	171 (982)	101 (0)	173
Growth rate	744 (3224)	830 (3357)	856



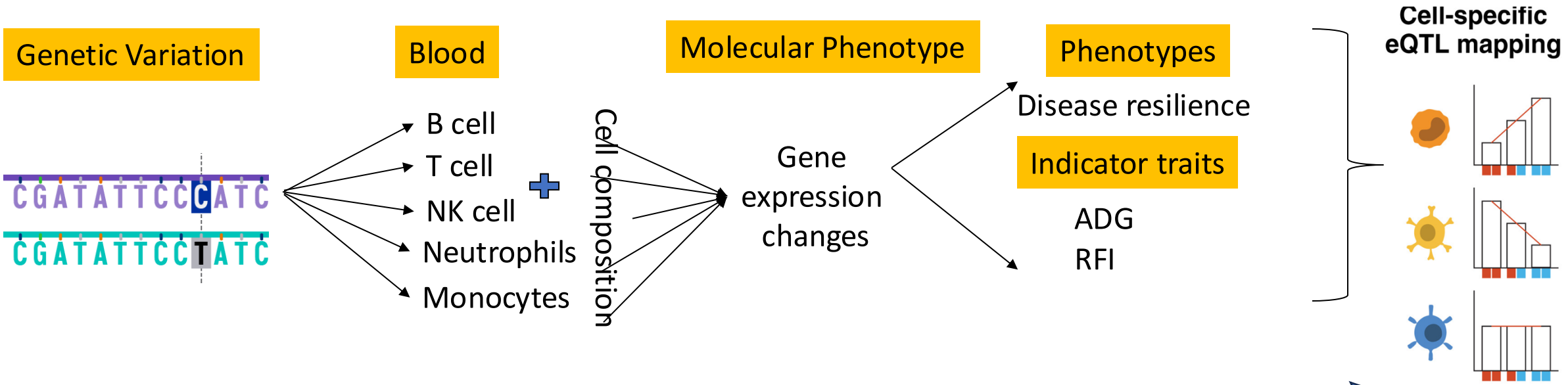
Quantitative analysis of the blood transcriptome of young healthy pigs and its relationship with subsequent disease resilience

Kyu-Sang Lim, Jian Cheng, Austin Putz, Qian Dong, Xuechun Bai, Hamid Beiki, Christopher K. Tuggle, Michael K. Dyck, Pig Gen Canada, Frederic Fortin, John C. S. Harding, Graham S. Plastow & Jack C. M. Dekkers

BMC Genomics **22**, Article number: 614 (2021) | Cite this article

Improving blood phenotypes: towards cell type specific phenotypes

Hypothesis: tools for accurate deconvolution of porcine whole blood transcriptome data into cell-type-specific transcriptome data will substantially improve molecular blood phenotypes as direct selection tools or as markers for animal traits



1000's of whole blood Rnaseq datasets!

Convert to:

- Cell composition phenotypes
- Cell-type specific RNA expression phenotypes

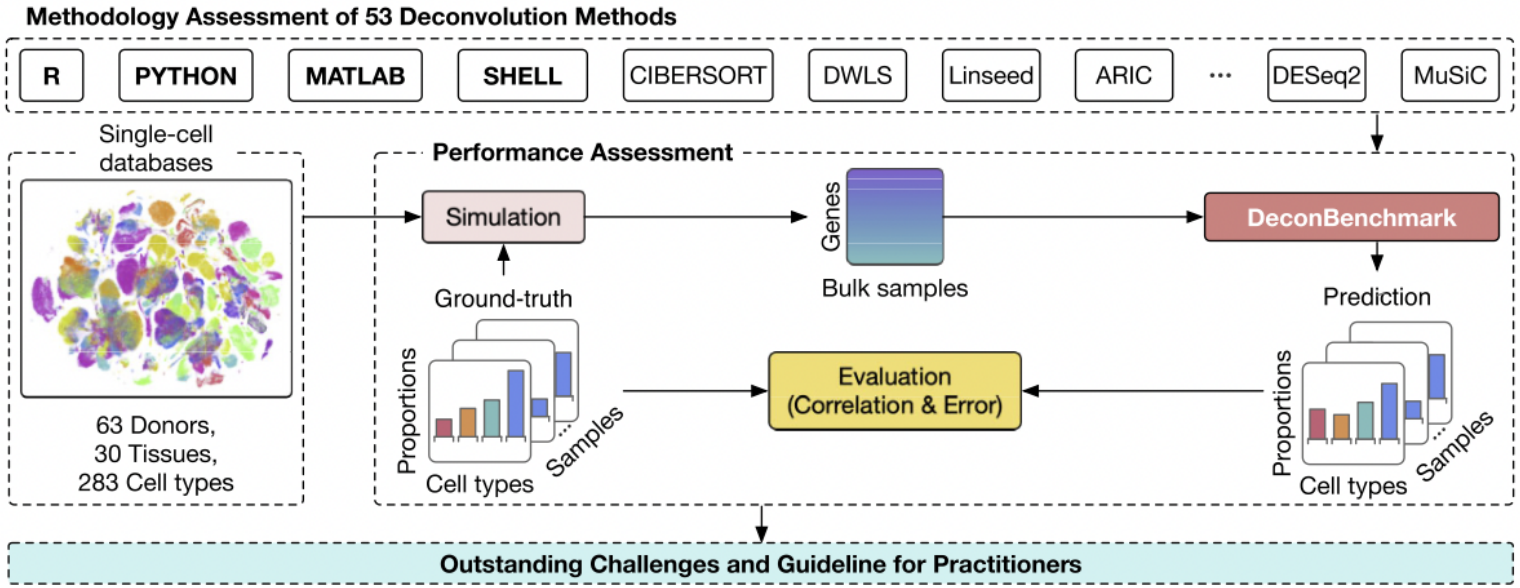
Cell Type Deconvolution of Cell Mixtures

Nucleic Acids Research, 2024, **52**, 4761–4783
https://doi.org/10.1093/nar/gkae267
Advance access publication date: 15 April 2024
Critical Reviews and Perspectives

Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges

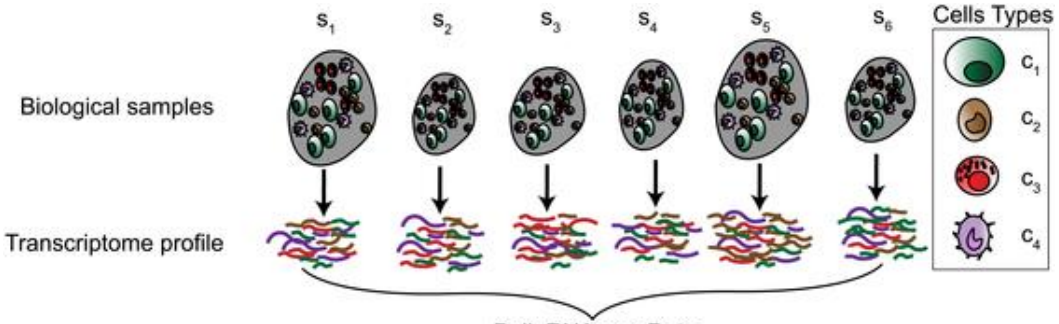
Hung Nguyen [1,†], Ha Nguyen [1,†], Duc Tran [2], Sorin Draghici [3,4,*] and Tin Nguyen [1,*]

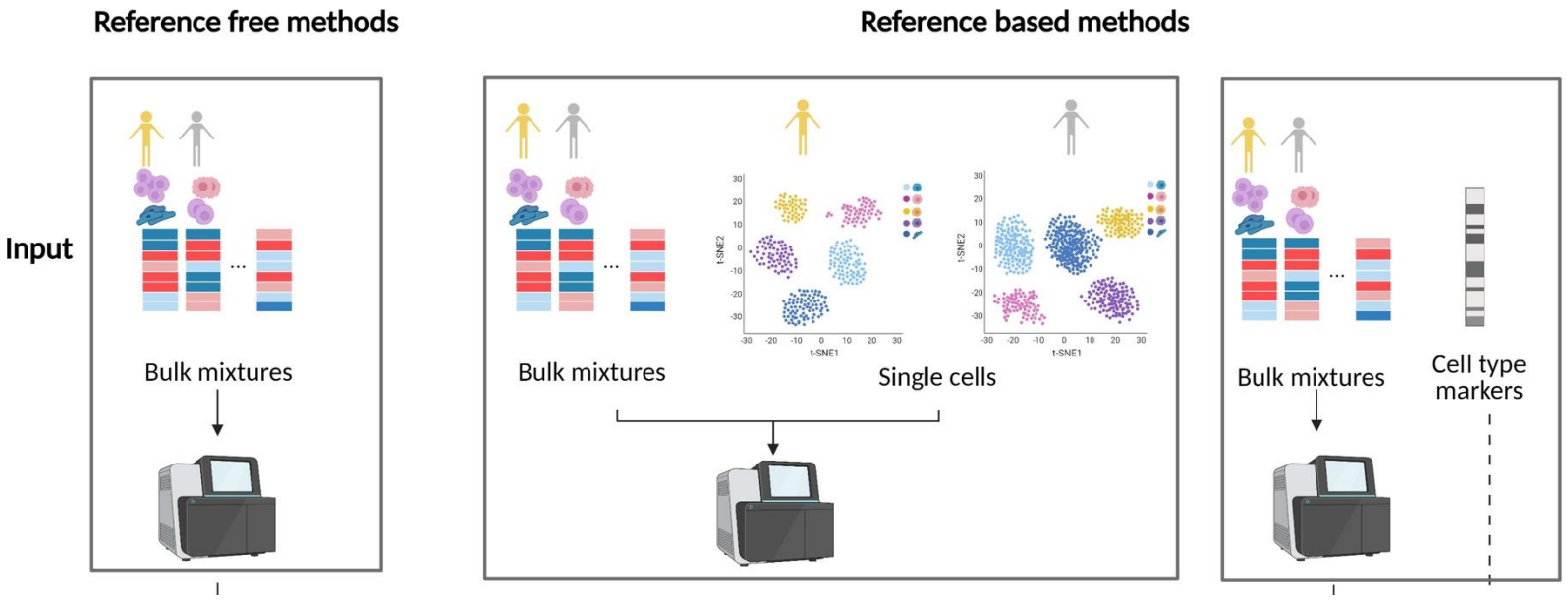
Graphical abstract



Cell Type Deconvolution

- Computational methods developed to infer cell type compositions and GEP within heterogenous samples
- Disease mechanism, immune response, transcriptional states, cell proportions in disease phenotypes





Cell Type Deconvolution of Cell Mixtures

Assumption: Total GE in bulk is *linear combination* of gene expression in individual cell types, weighted by cell type proportions.

Model: *M= S x F*

M= N genes x m samples

S= N genes x C cell types

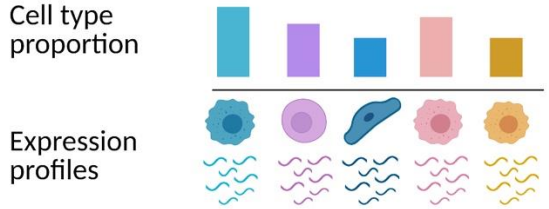
F= C cell types x m samples

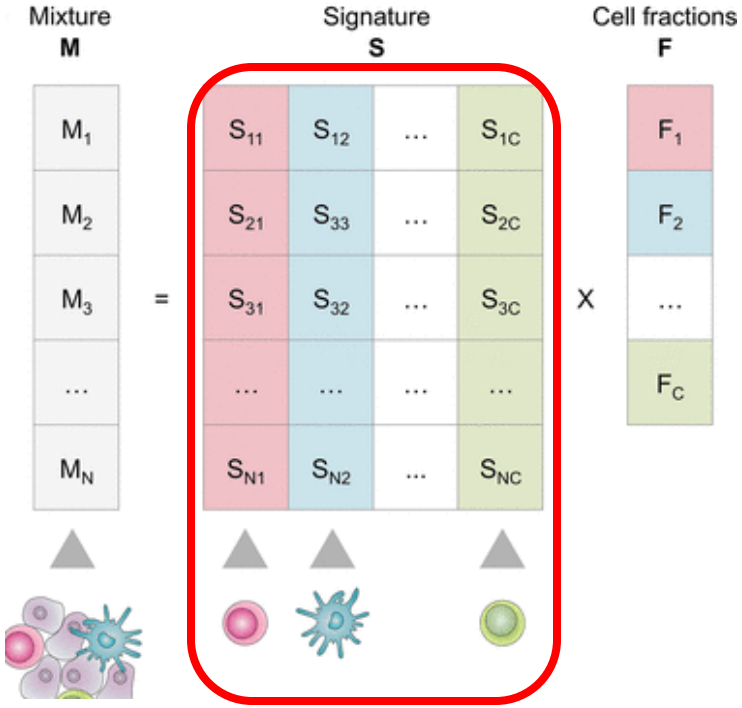
For every gene in bulk:

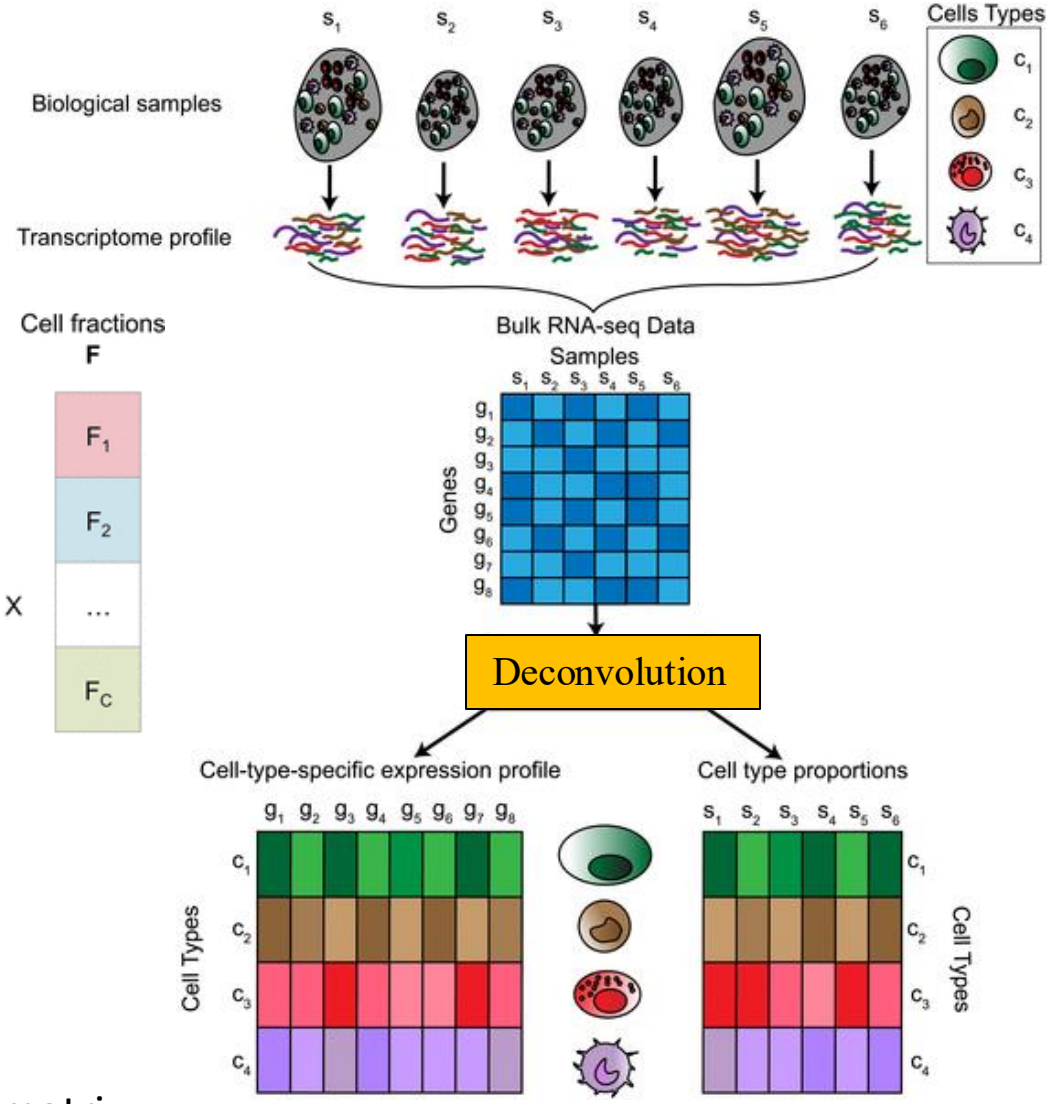
$M_{N1} = F_1 \times S_{N1, C1} + F_2 \times S_{N1, C2} \ldots.$

$M_{N2} = F_1 \times S_{N2, C1} + F_2 \times S_{N2, C2} \ldots.$

Output: cell type composition, CTS-GEP







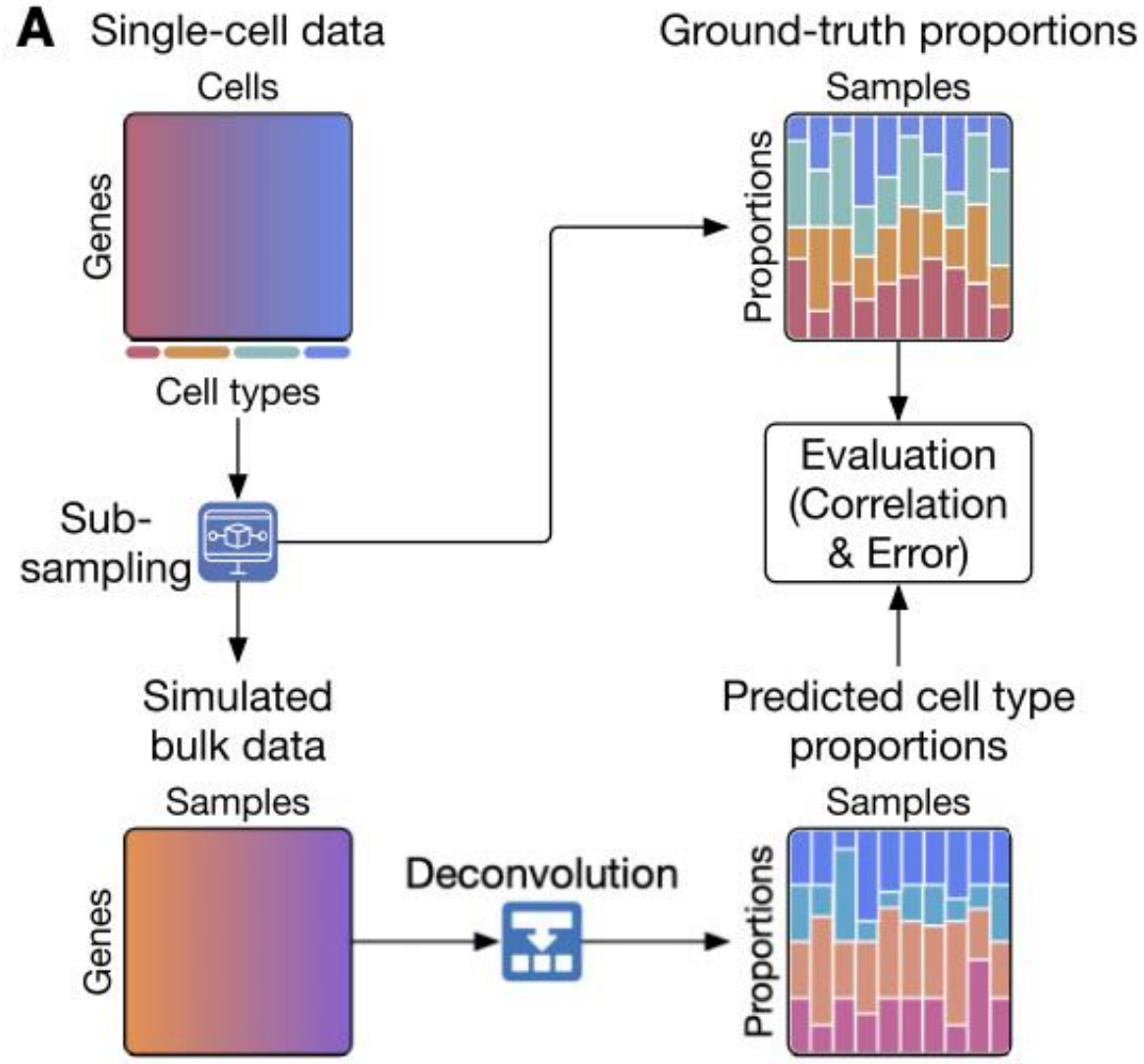
Gene expression signature matrix:

- Rows are genes
- Columns are cell types

Validation of results after Deconvolution

Starting Data:

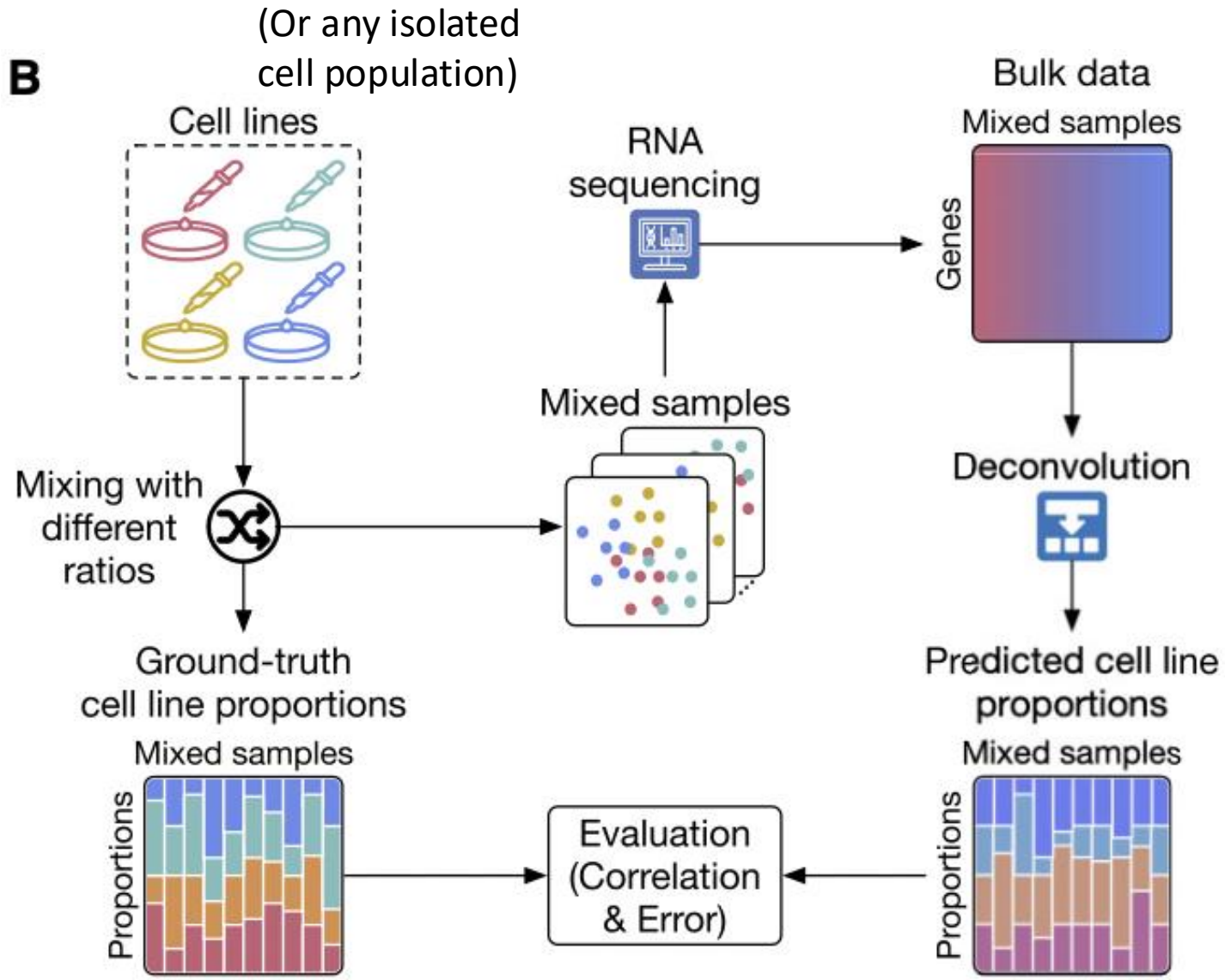
Using SC RNAseq for relevant samples, create "pseudobulk" samples to be deconvoluted



Validation of results after Deconvolution

Starting Data:

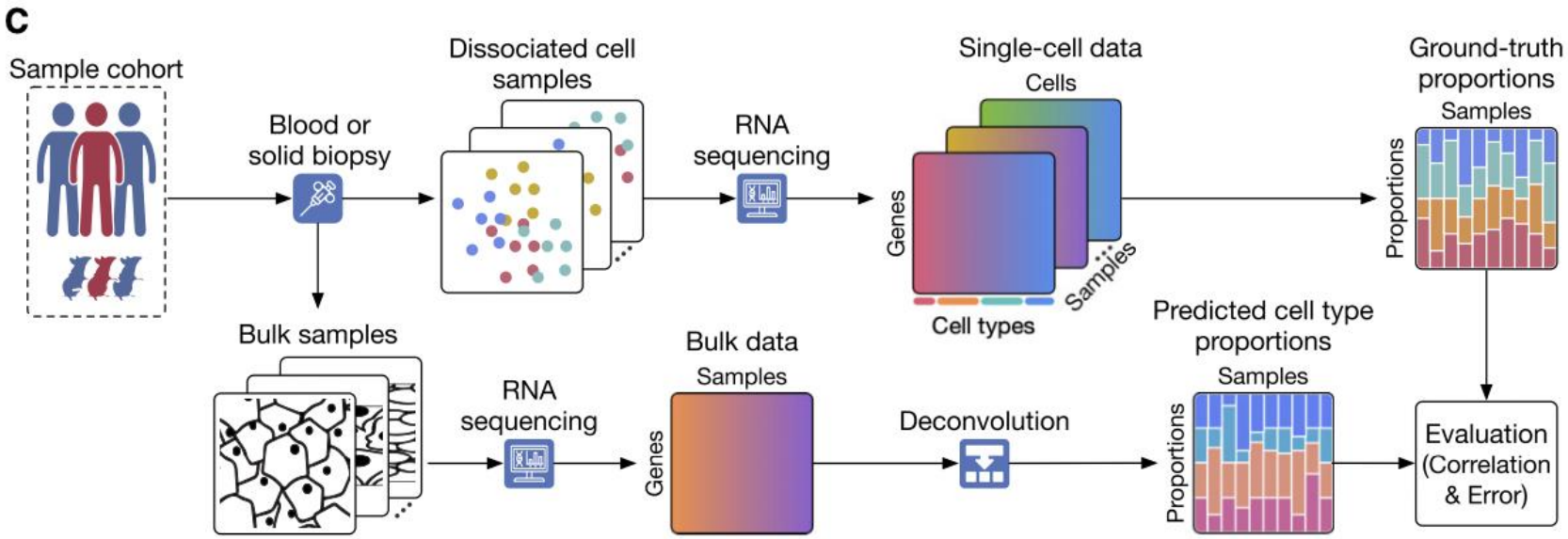
Collect data from separate cell populations and create artificial mixtures, run bulk RNAseq of individuals and mixtures with known proportions



Validation of results after Deconvolution

Starting Data:

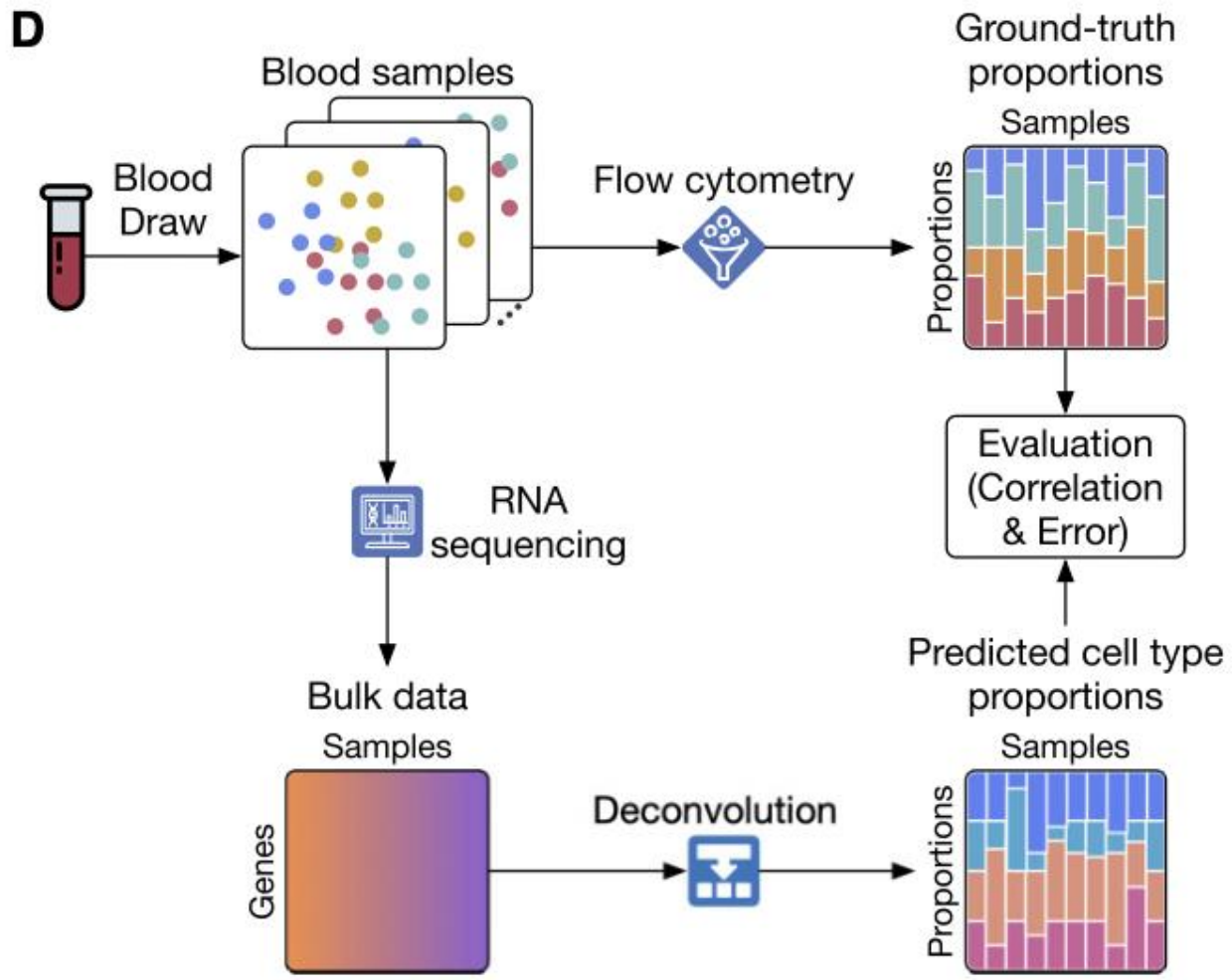
Use bulk RNAseq and dissociated cells from replicate samples in scRNAseq to determine cell proportions



Validation of results after Deconvolution

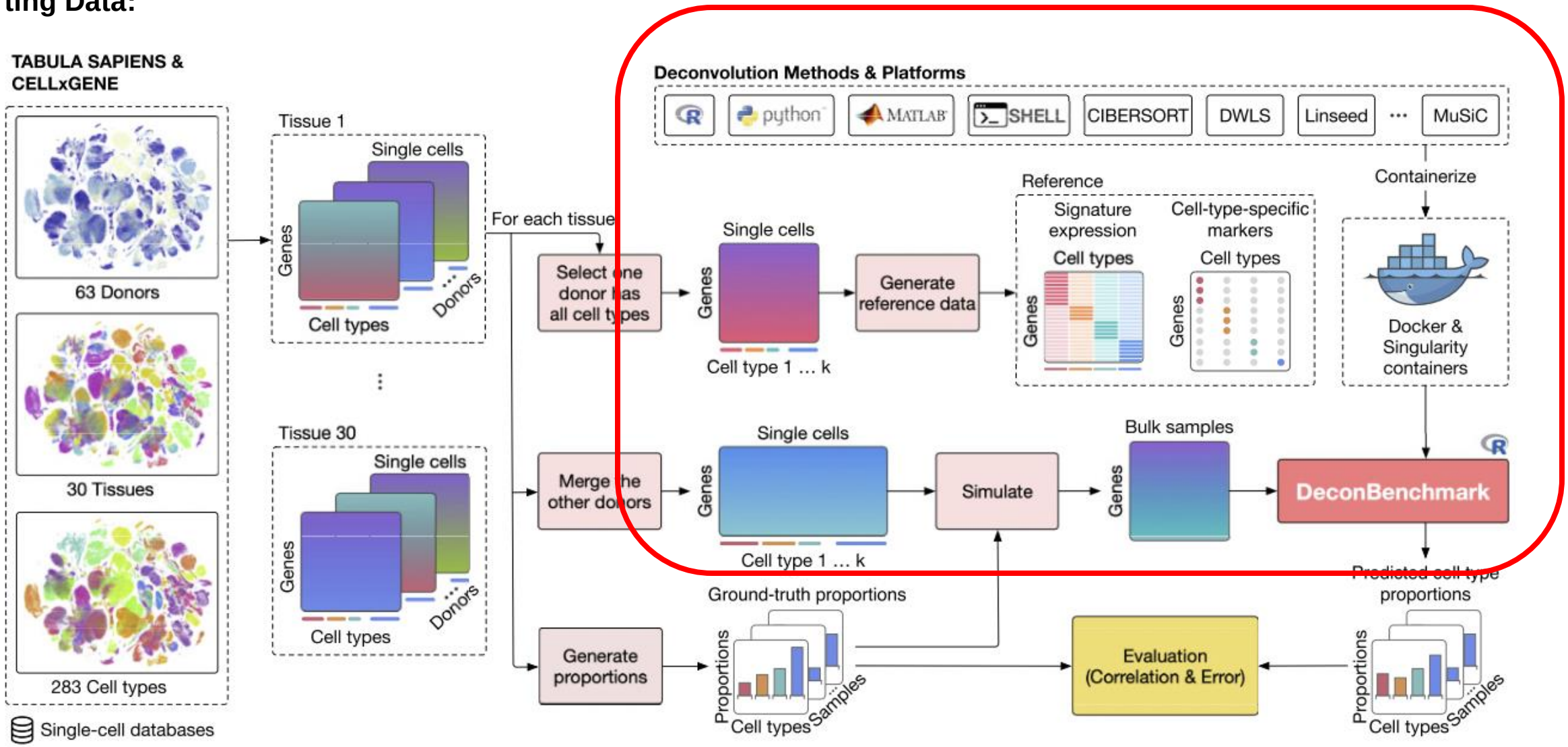
Starting Data:

Run flow cytometry for isolated cells from blood, or tissue for cell composition, run bulk RNAseq of original sample



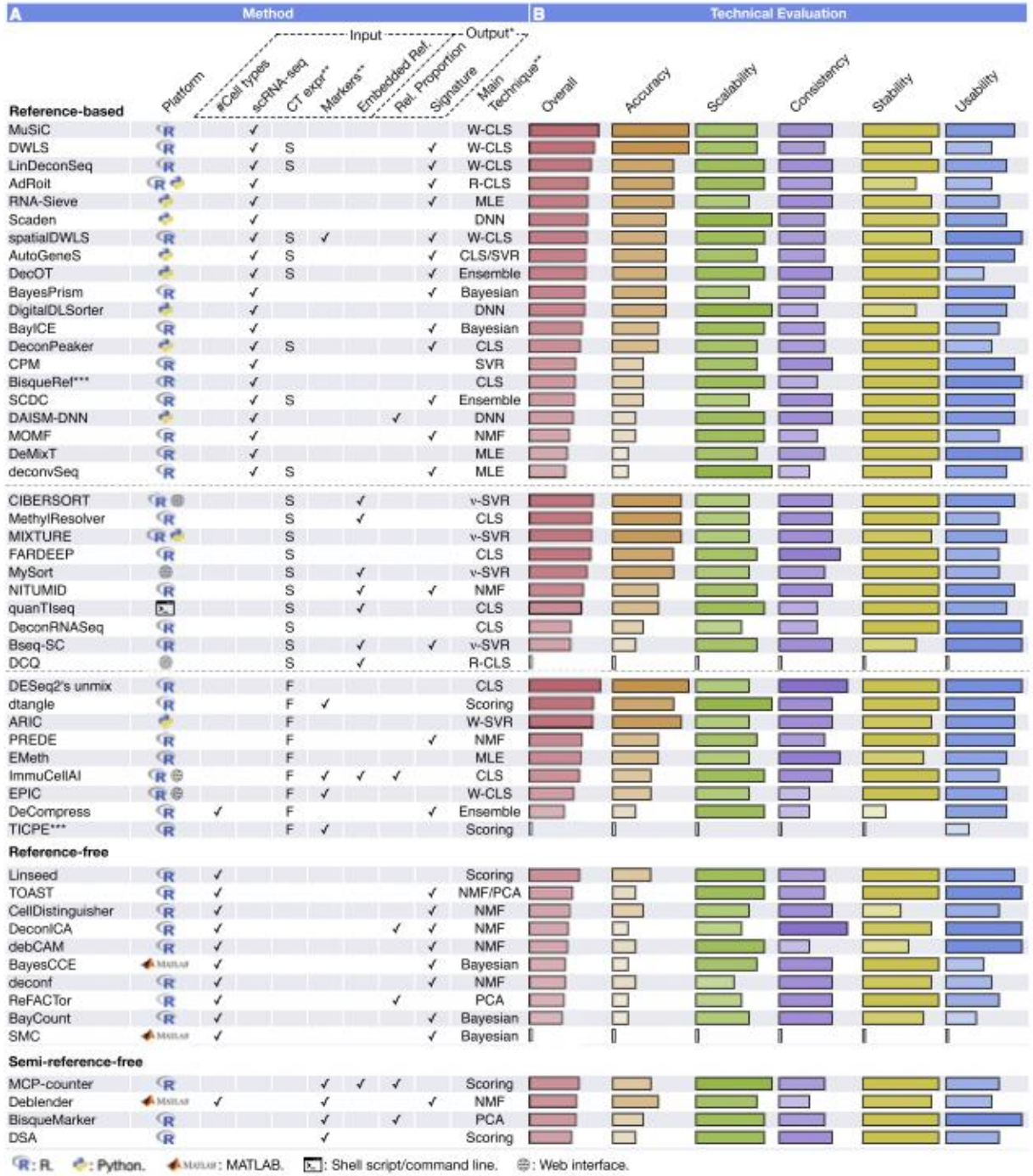
Comprehensive testing of methods using tissue atlases: scoring

Starting Data:

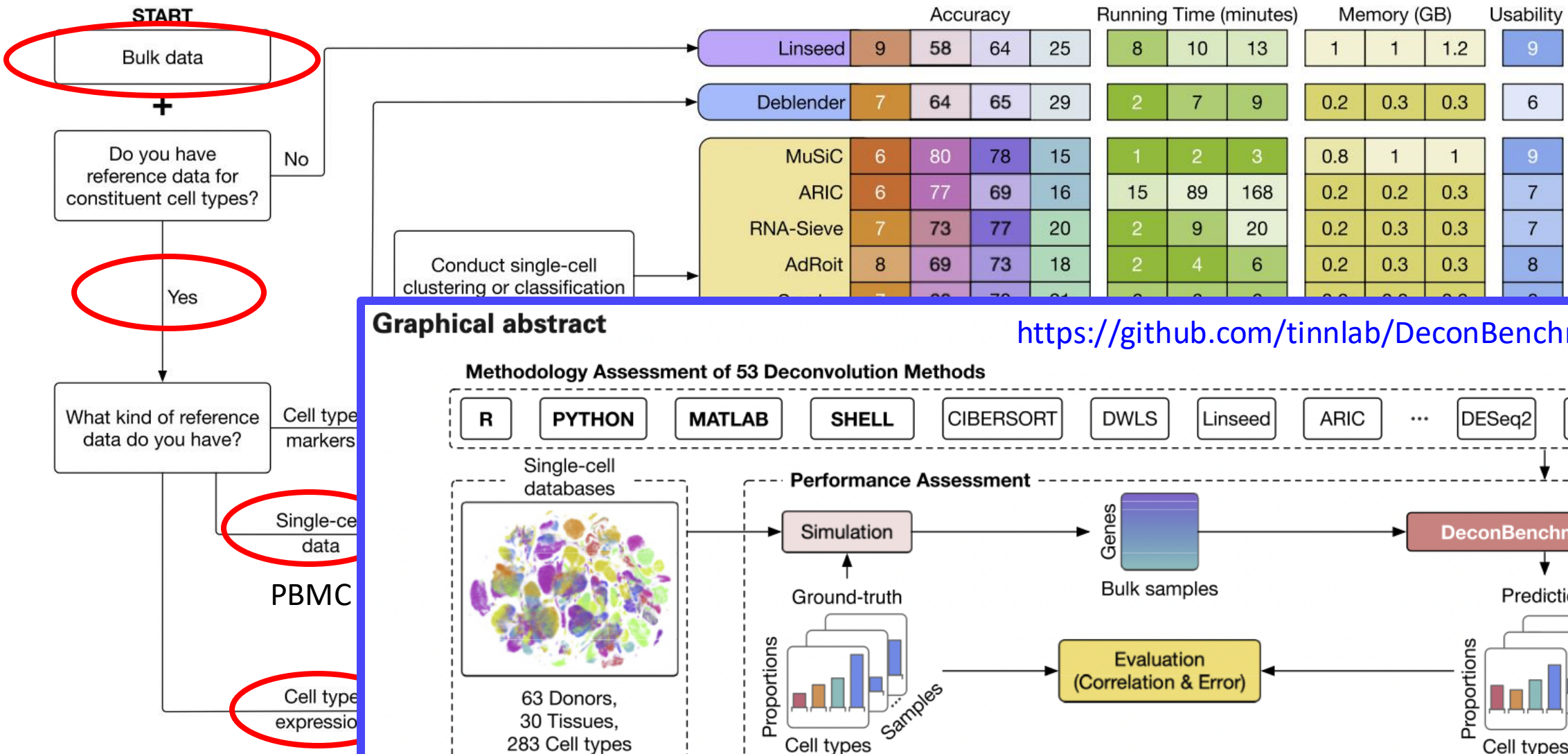


Many, many methods for Deconvolution...

Evaluated using the multi-tissue reference from Tabula Sapiens and CellxGene



Curated path to choose method for specific situations!



Application for deconvolution of whole blood RNAseq data: Create Cell-type-specific signatures


Carrie Meeks

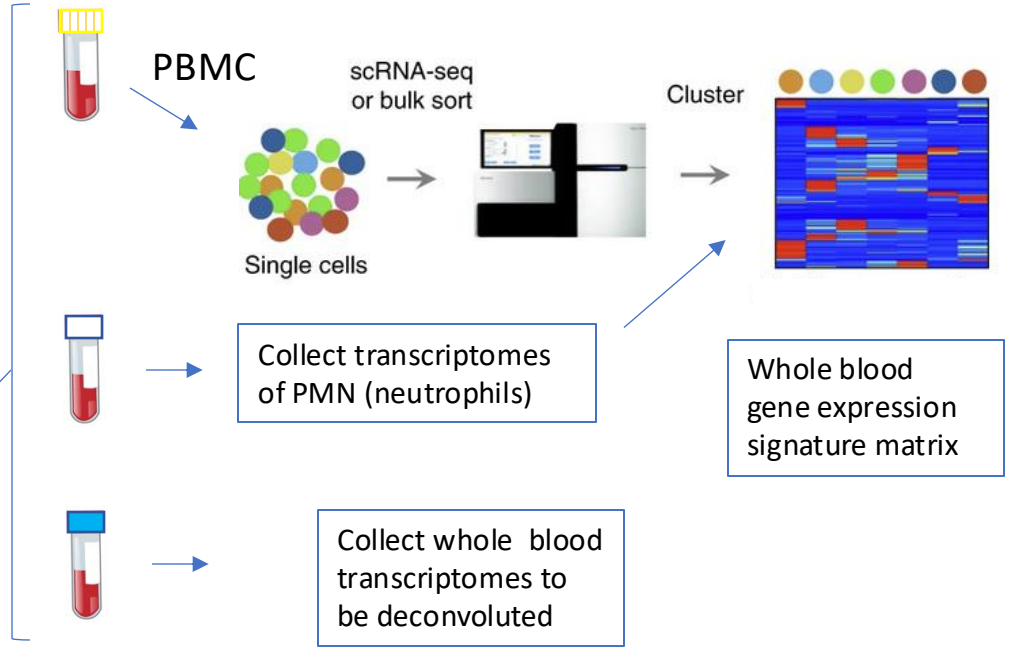

Kristen Byrne


Mehak Kapoor


Muskan Kapoor

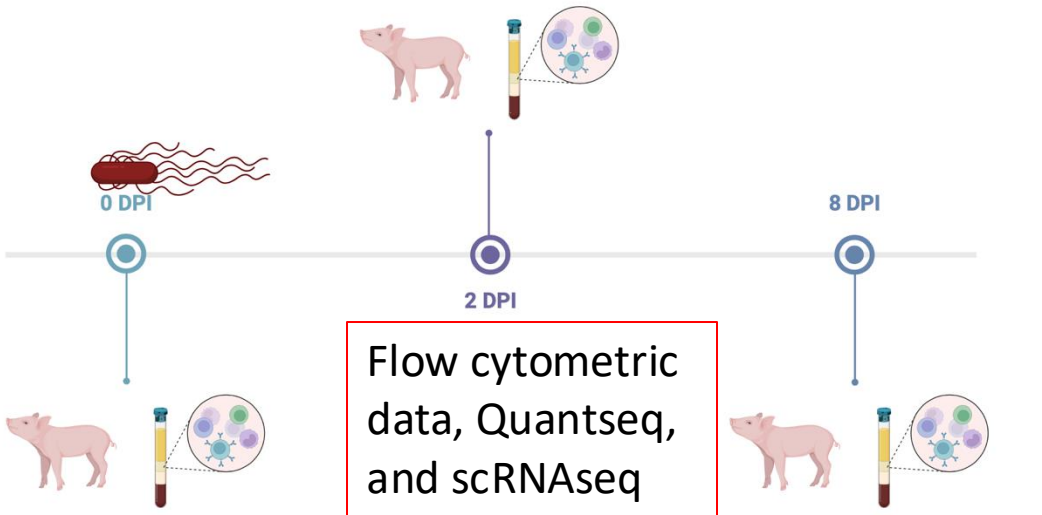
- Pigs challenged with bacteria or virus
- Blood collected during challenge 3 time points- total of 45 samples

Create new reference information on immune-stimulated pigs

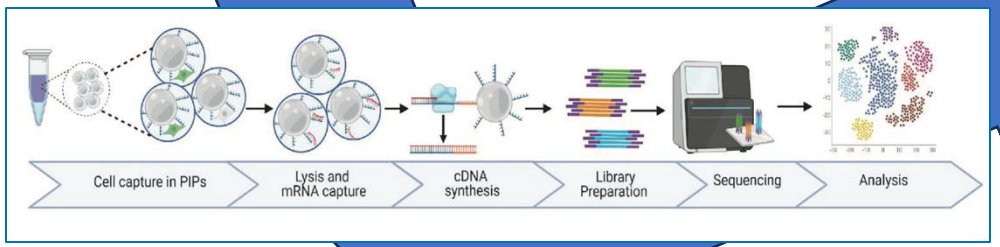


Nature Biotechnology **volume 37**, pages 773–782 (2019), adapted

Study Design: scRNAseq of PBMC from *Salmonella* challenged pigs



Flow cytometric data, Quantseq, and scRNAseq data collected



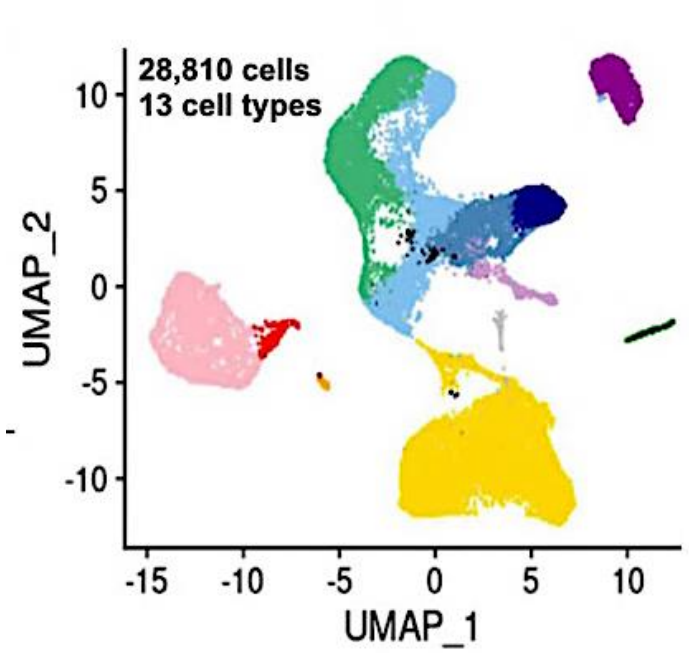
Cells passing QC for each sample

Pig ID	0D	2D	8D	Total
842	10,306	10,164	16,274	36,744
852	6,877	4,304	5,151	16,332
853	8,319	4,600	7,089	20,008
854	1,913	5,285	1,927	9,125
864	4,959	11,236	4,671	20,866
Total	32,374	35,589	35,112	103,075

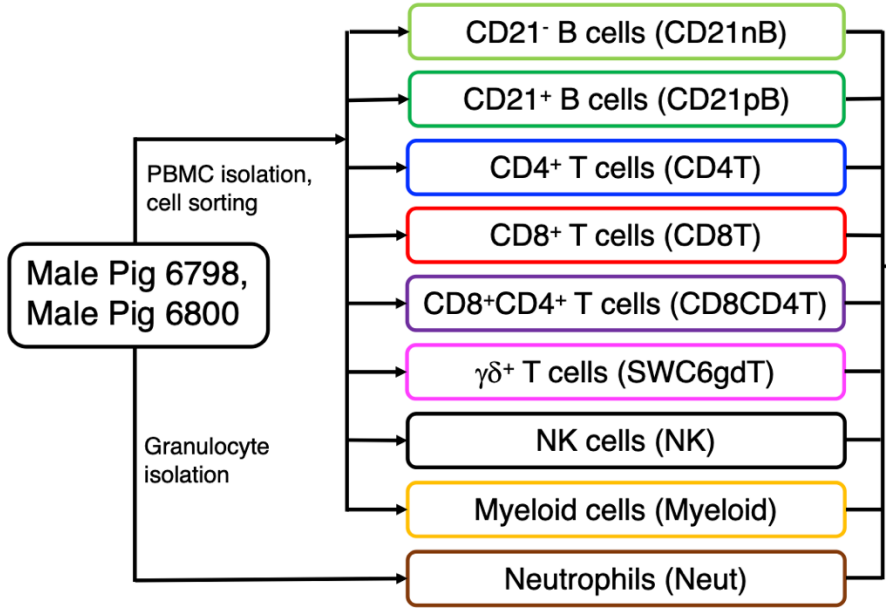
Cells with >300 genes, >400 UMIs and <15% mito sequences were retained, duplets removed.

Additional data available to project

- scRNAseq of PBMC samples of 7 healthy pigs (10X Genomics)
- Total of 28,810 cells.
- 36 clusters grouped to 13 major porcine cell types.



- RNAseq of nine sorted populations of blood cells from 2 healthy pigs
- Cell populations covering all nucleated cells in blood



Herrera-Uribe and Wiarda, et al. 2021;
Herrera-Uribe et al. 2023

Application for deconvolution of whole blood RNAseq data: Create Cell-type-specific signatures


Carrie Meeks


Kristen Byrne


Mehak Kapoor


Muskan Kapoor

- Pigs challenged with bacteria or virus
- Blood collected during challenge 3 time points- total of 45 samples

Create new reference information on immune-stimulated pigs

total of 15 PBMC samples plus 7 from prior publication

PBMC

Collect transcriptomes of PMN (neutrophils)

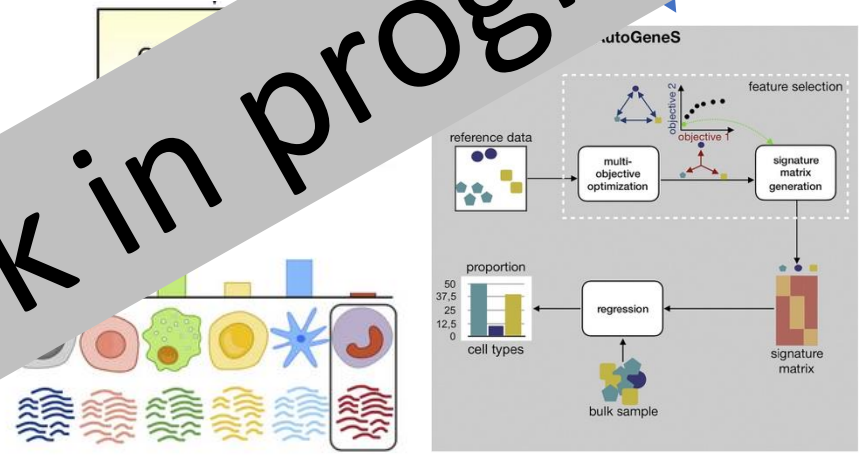
Whole blood gene expression signature matrix

Collect whole blood transcriptomes to be deconvoluted

Collect eight-cell type Flow cytometry - based Cell Counts (TRUTH) n=48

Is gene expression matrix valid?

Obtain pr… pro…

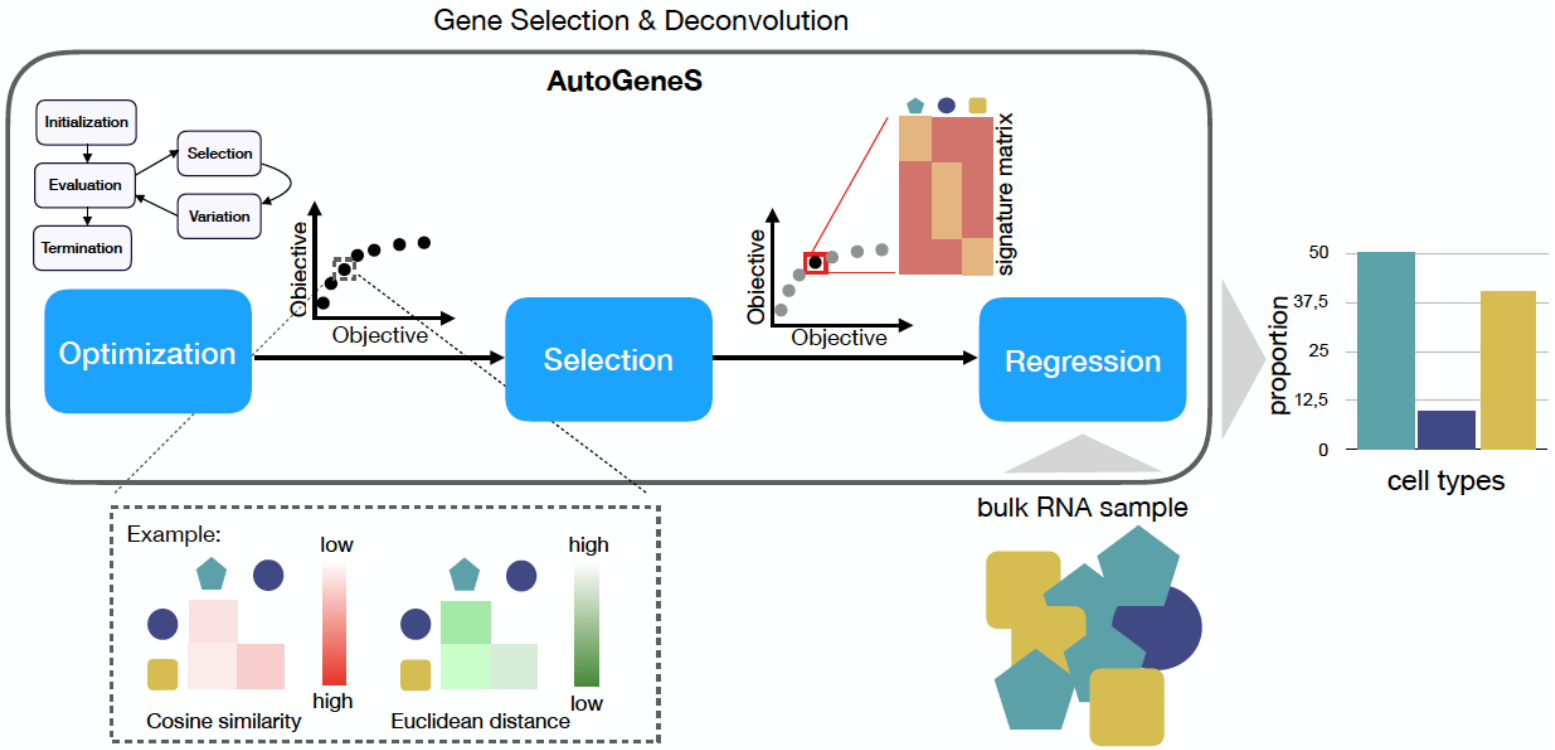


Work in progress!

Nature Biotechnology **volume 37**, pages 773–782 (2019), adapted

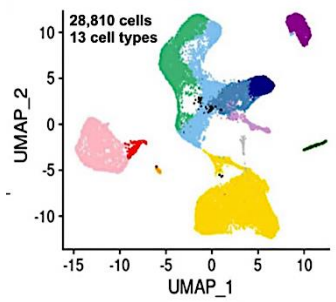
Feature Selection Tool: AutoGeneS

- Automatic feature selection, does not rely on pre-defined markers.
- Deals with multi-collinearity, a critical problem in deconvolution when dealing with closely related cell types
- Employs multi-objective optimization as a solution to select a set of non-collinear genes.
- Aims at minimizing correlation and maximizing Euclidean distance

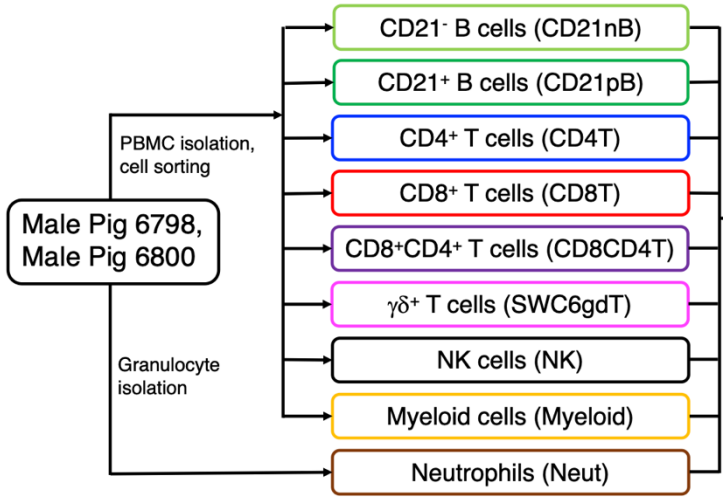


Validation methods we can use

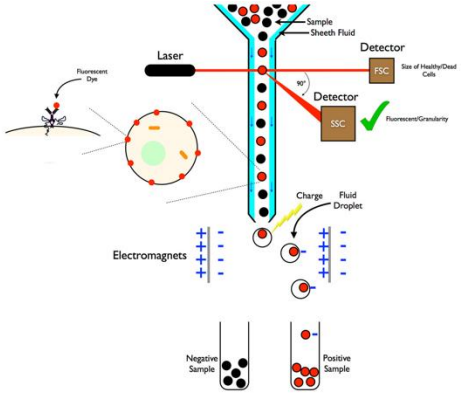
Thirteen pseudobulk populations of scRNAseq-based data

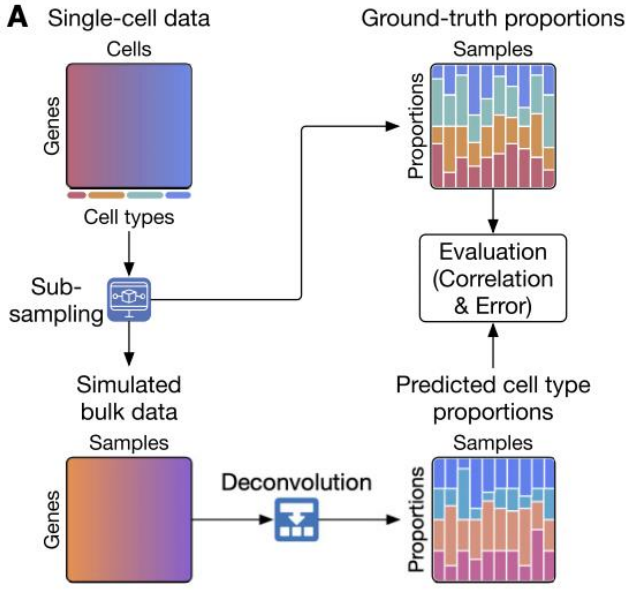


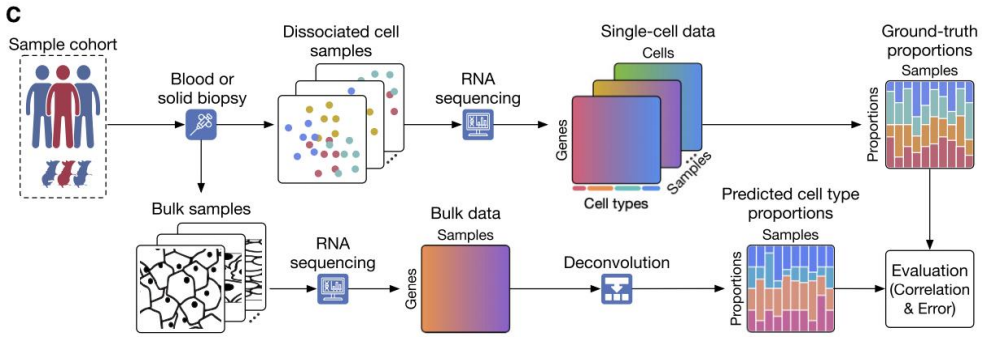
Nine sorted cell populations

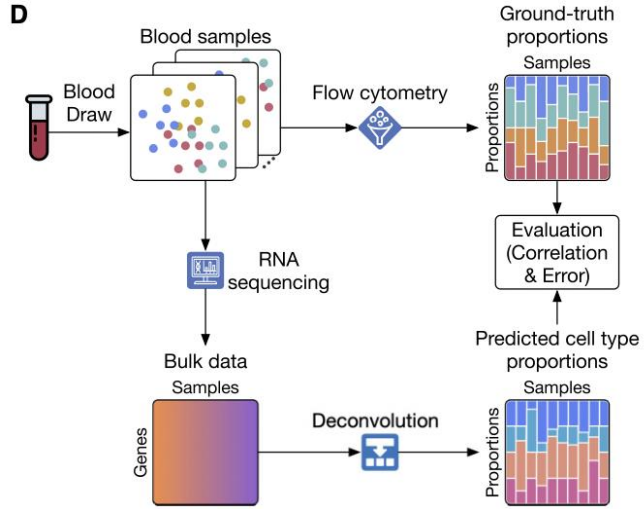


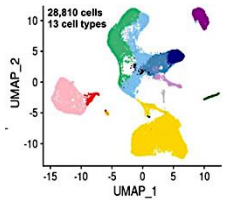
Eight cell type Flow cytometry - based Cell Counts of Whole blood RNAseq samples n=45





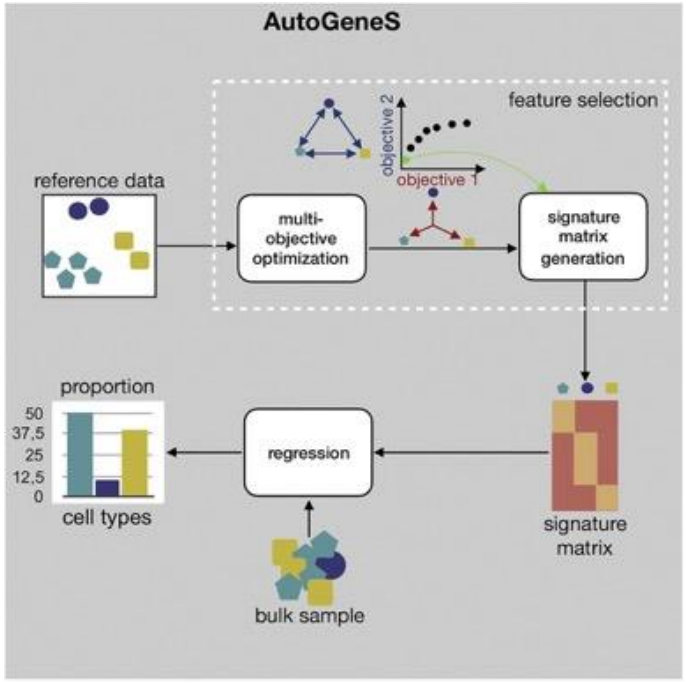


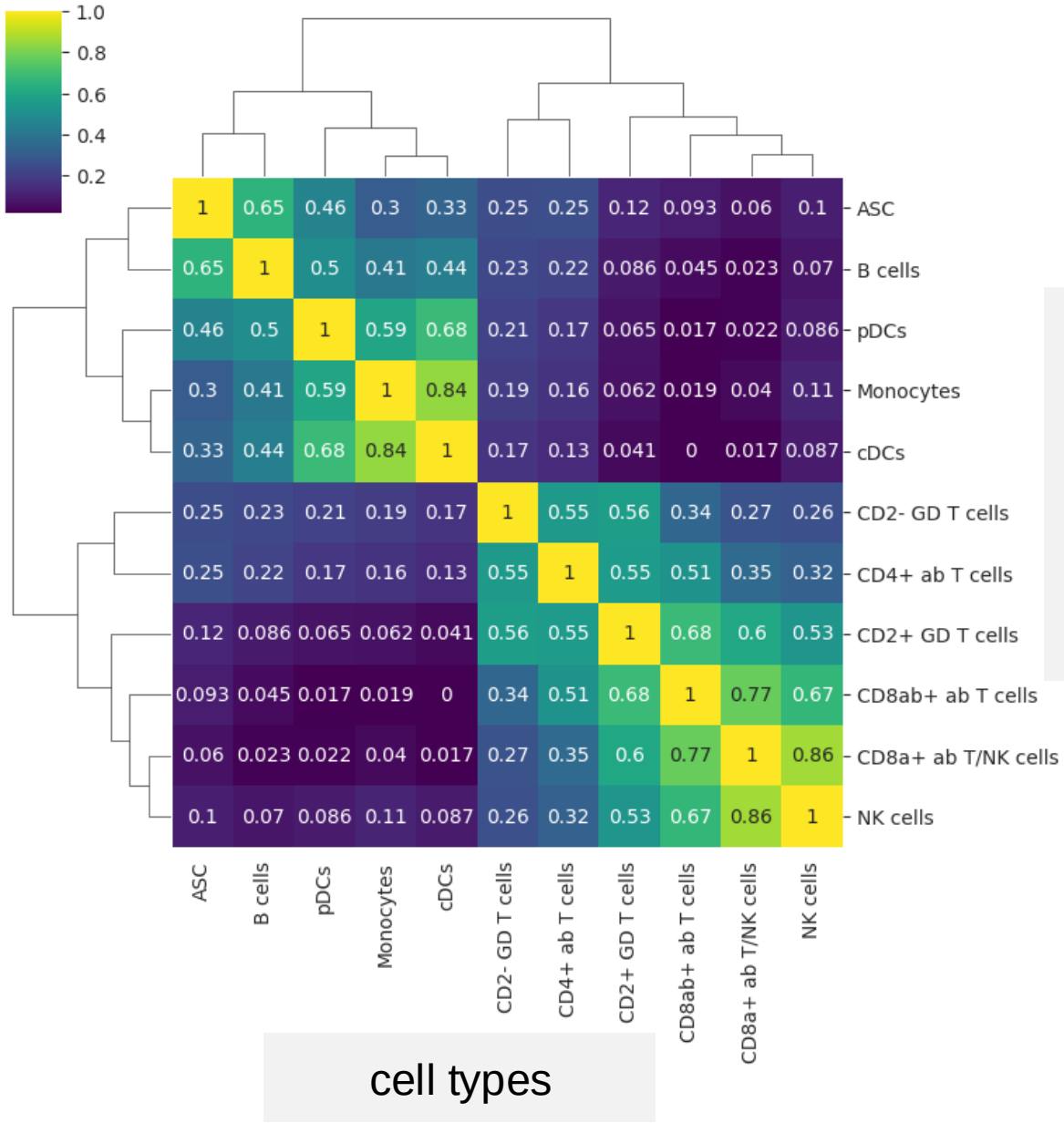




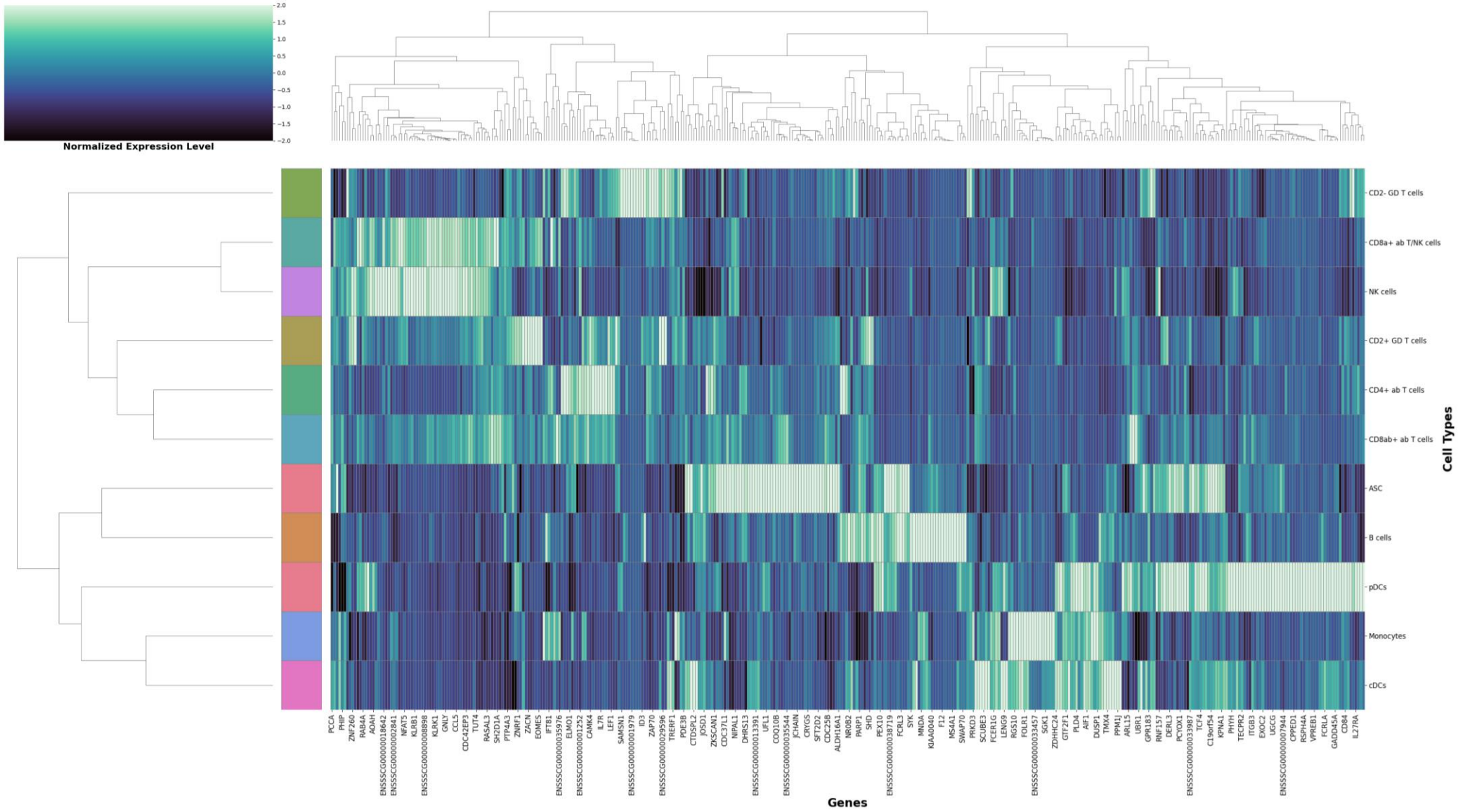
AutogeneS PBMC only

- Used 7 sample published scRNAseq data set as reference
- MOO to select features and create gene expression signature matrix
- Check which cell types are highly correlated (will be difficult to distinguish) and refine cell type markers if needed



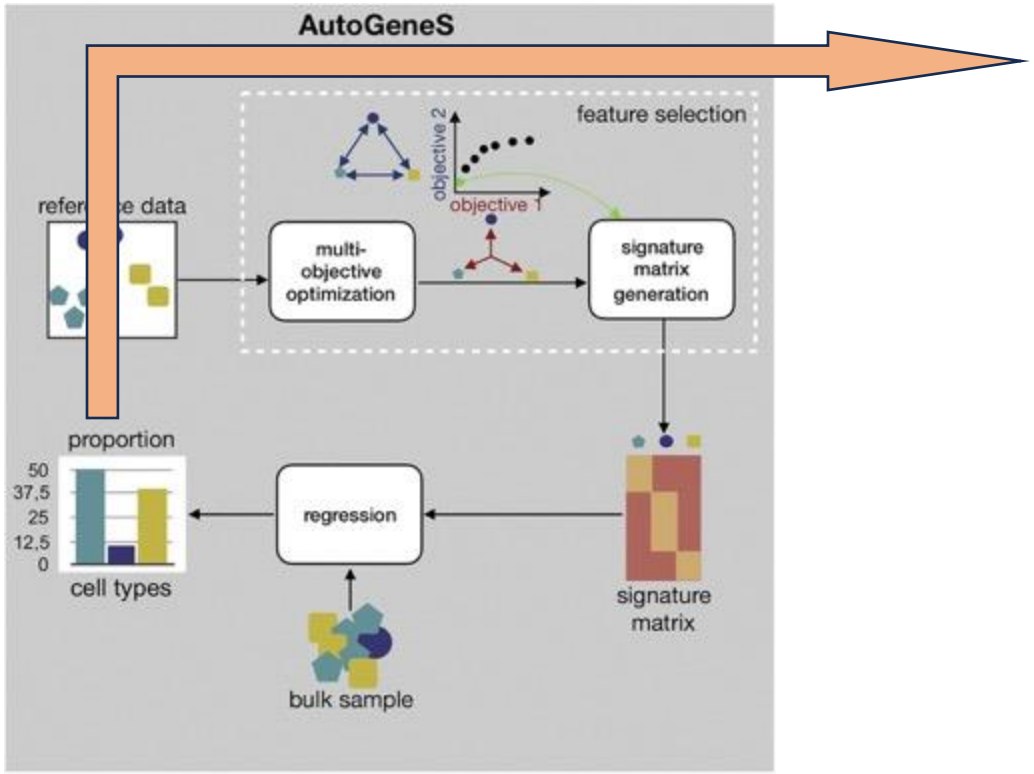


Heatmap of GE signature matrix (400 features)

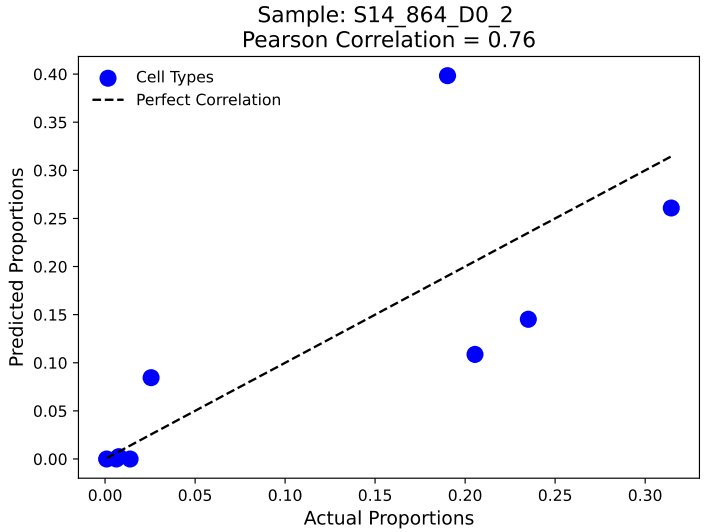


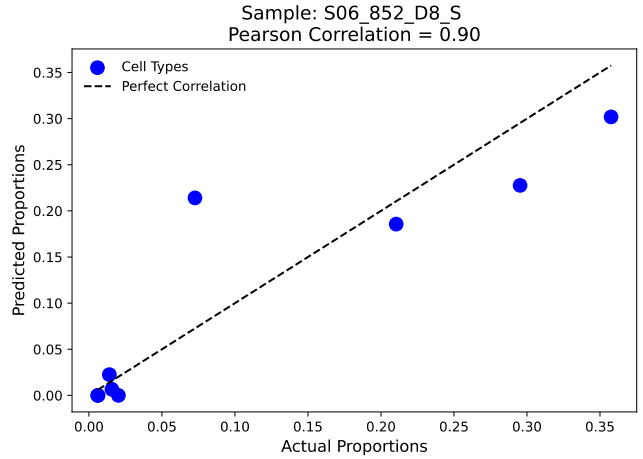
AutogeneS PBMC deconvolution results

- Use GE Sig matrix from 7 sample data to predict proportions of new 15 sample scRNAseq data set (*Salmonella* infection)
- Used pseudobulking of 15 datasets
- Two models (NNLS, NuSVR)



Within sample across cell types

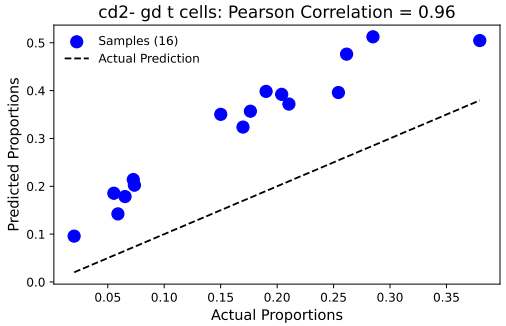


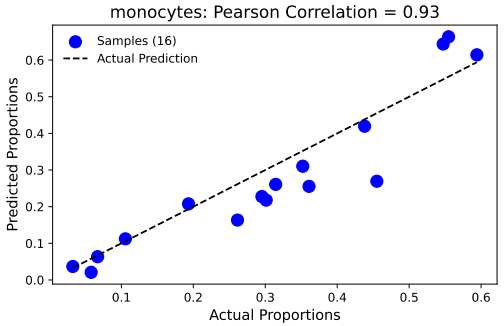


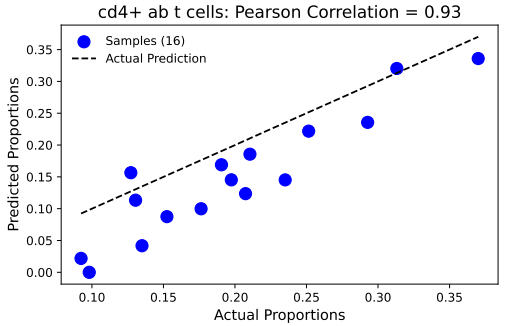
Pearson correlation range 0.67-0.96 across 15 samples

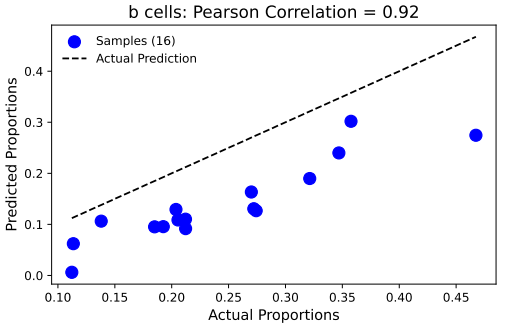
AutogeneS PBMC deconvolution results

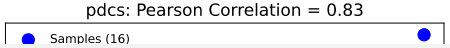
Across samples for each cell type











Summary

- Good predictions for most cell types
- This matrix could be used to deconvolute PBMC RNAseq datasets for major cell types
- More work needed for low abundance cells
- For whole blood, we need to add a major cell type not present in PBMC: PMNs (primarily neutrophils)

Future Plans- Deconvolution

1. **Create new GE signature matrix (GESM) from sc PBMC+PMN to test methods against datasets (with TRUTH):**
   - **48 Sal and Flu samples with flow cytometry cell type data and Quantseq RNA data**
   - **42 RFI WB samples with flow cytometry cell type data and RNAseq data**

   Also test other tools such as CIBERSORTx, etc.

   Will also use purified cell populations as additional verification/GESM creation

2. **Accurate deconvolution of NDCM dataset: 2400+ Quantseq samples**

3. **Develop "short-hand" version of deconvolution GE signature with Nanostring genes**

   → practical tool for deconvolution for phenotyping at cell type level to eliminate RNA prep, Quantseq and bioinformatics costs

Acknowledgments


Christopher Tuggle


Juber Herrera Uribe


Ryan Corbett


Pengxin Yang


Muskan Kapoor


Carrie Meeks


Mehak Kapoor


Crystal Loving


Jayne Wiarda


Kristen Byrne


Sathesh Sivasankaran


Jack Dekkers


Luke Kramer

IOWA STATE UNIVERSITY

NIFA Project #2021-67015-34562


Joan Lunney

USDA ARS Agricultural Research Service


Graham Plastow

UNIVERSITY OF ALBERTA

Application of Genomics to Improve Disease Resilience and Sustainability in Pork Production

GenomeAlberta GenomeCanada GenomePrairie

USDA United States Department of Agriculture National Institute of Food and Agriculture

NIFA 2017-67007-26144 + 2018-67015-2701

FAANG Functional Annotation of Animal Genomes