scNMT-seq: multi-modal integration and feature selection using projection to latent structures

Al JalalAbadi

LêCao Lab

Melbourne Integrative Genomics, School of Mathematics & Statistics

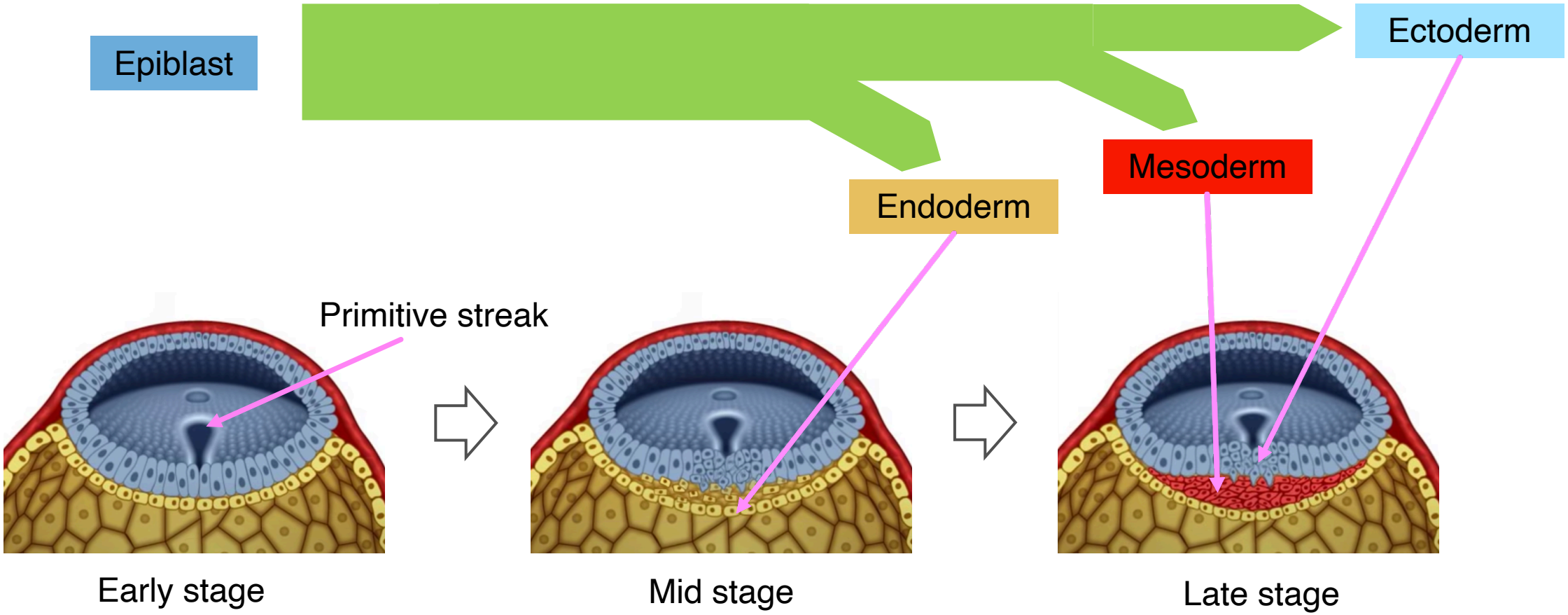
Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Types (Online)

17 June 2020

THE UNIVERSITY OF MELBOURNE

@aljabadi

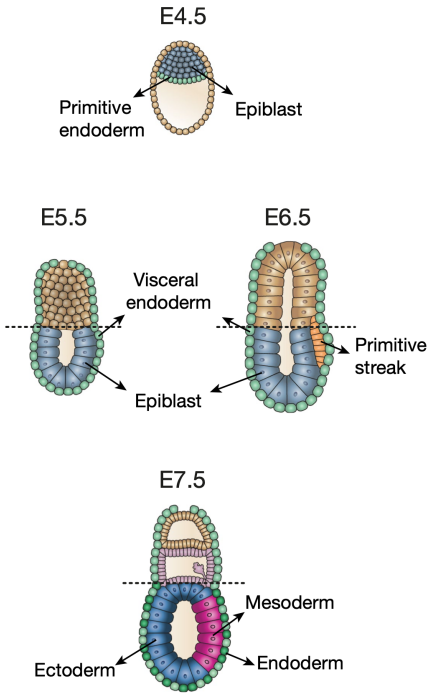
Mouse gastrulation

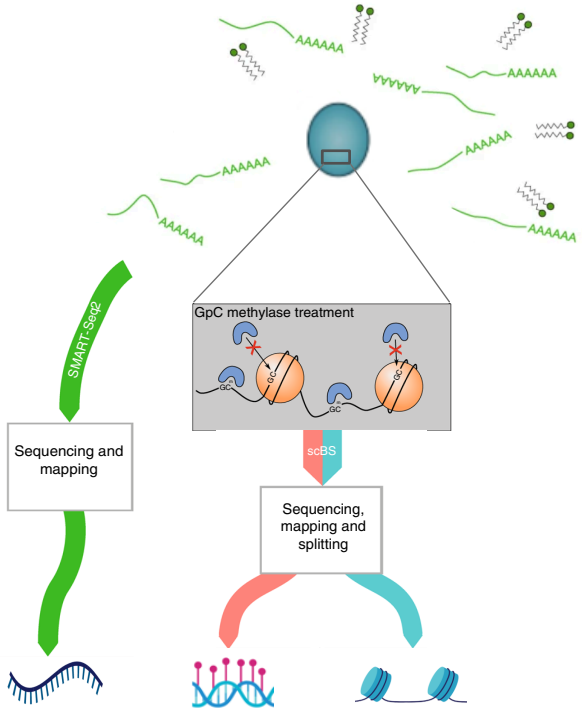


What are the coordinated changes across genomic modalities leading to each lineage commitment?

https://youtu.be/ADlYn0ImTNg

Single cells from different stages of mouse gastrulation





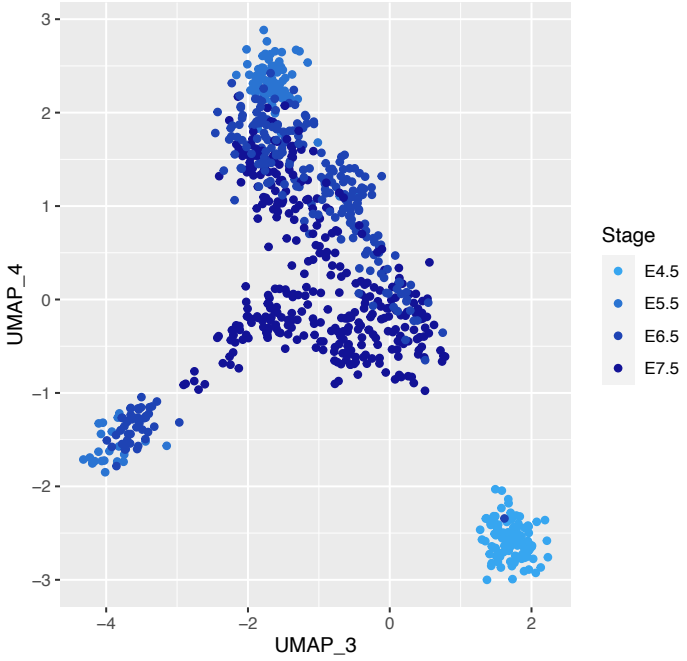
	E4.5	E5.5	E6.5	E7.5
Ectoderm	0	0	0	43
Endoderm	0	0	0	81
Epiblast	60	84	146	44
Mesoderm	0	0	28	141
Primitive_endoderm	43	0	0	0
Primitive_Streak	0	0	43	33
Visceral_endoderm	0	24	45	0

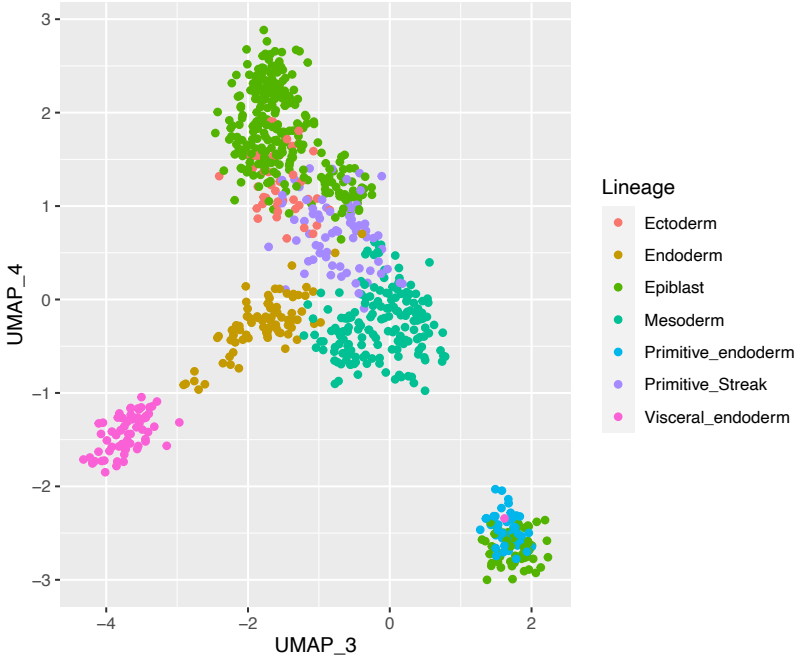
- Multiple donors (mice)

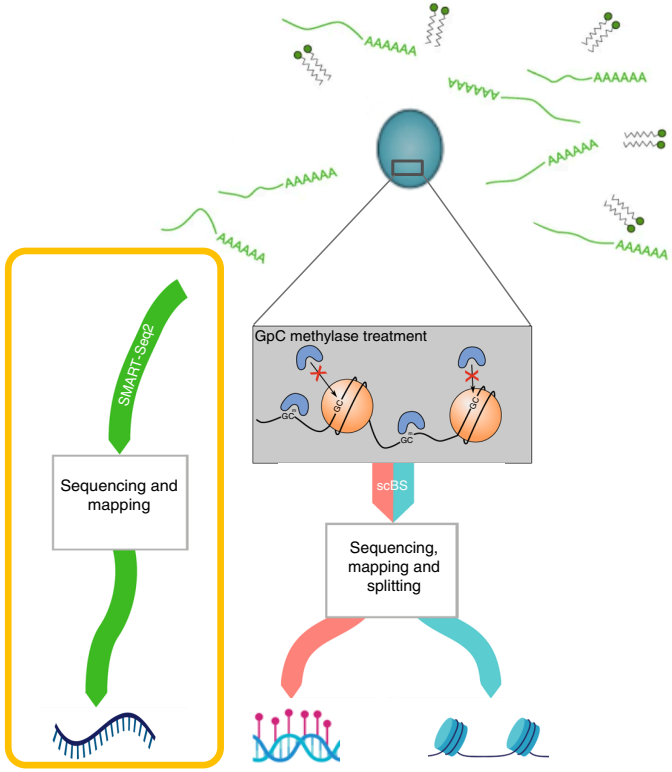
Argelaguet *et al*. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 576, 487–491 (2019)

Transcriptome

Non-linear dimension reduction using UMAP







- Early stage cells are transcriptionally distinct from others
- Putative lineages assigned using transcriptome data

McInnes, Leland, and John Healy. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction."

DNA-level measurements are binary calls

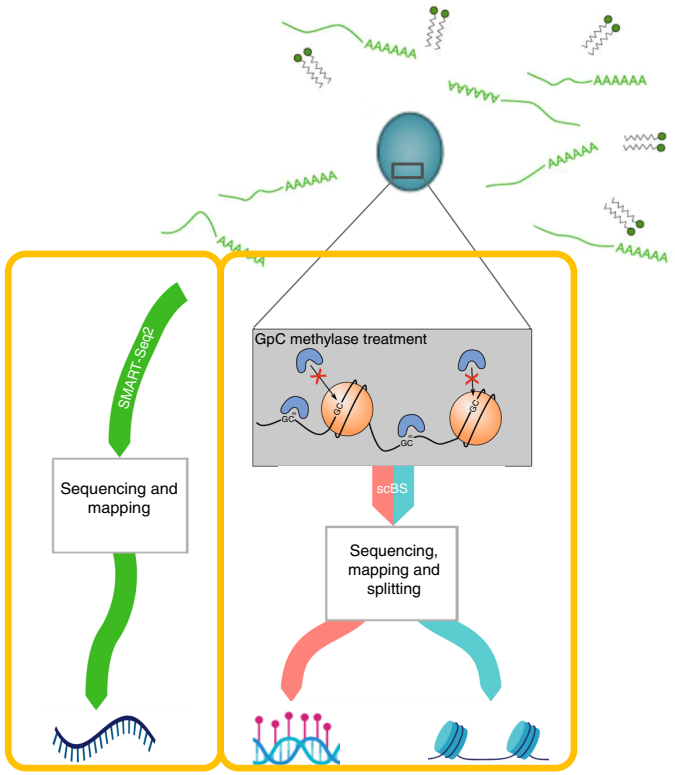
Region	# of methylated/accessible (c^+)	# of unmethylated/inaccessible (c^-)
Chr. IV 12,222,872-12,227,112	4	1
Chr. III 81,782,112- 81,837,335	453	103

$$C^+ \sim Binomial\ (c^+ + c^-, \mathrm{r})$$

MAP estimate with $\beta(1,1)$ (pseudo-count) prior: $\hat{r} = \frac{c^+ + 1}{c^+ + c^- + 2}$

$$SE[\hat{r}]^2 = \frac{\hat{r}\ (1\ - \hat{r})}{c^+ + c^-}$$

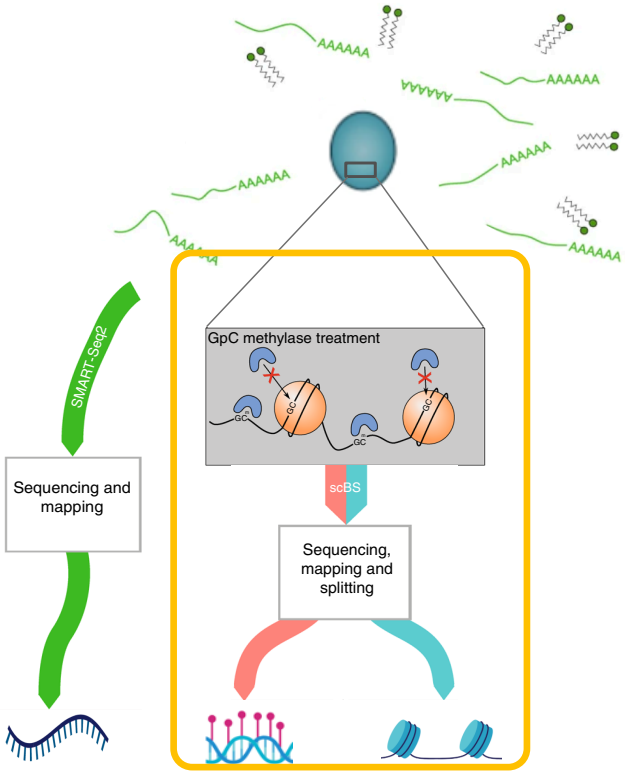
$$W_{feature,cell} = \frac{1}{SE[\hat{r}]^2}$$



Smallwood *et al*. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014

DNA-level measurements summarised over various genomic contexts

Genomic context	Dataset name	Dataset name
genebody	met_genebody	acc_genebody
promoter	met_promoter	acc_promoter
p300 binding sites	met_p300	acc_p300
DHS	met_DHS	acc_DHS
CG island	met_cgi	acc_cgi



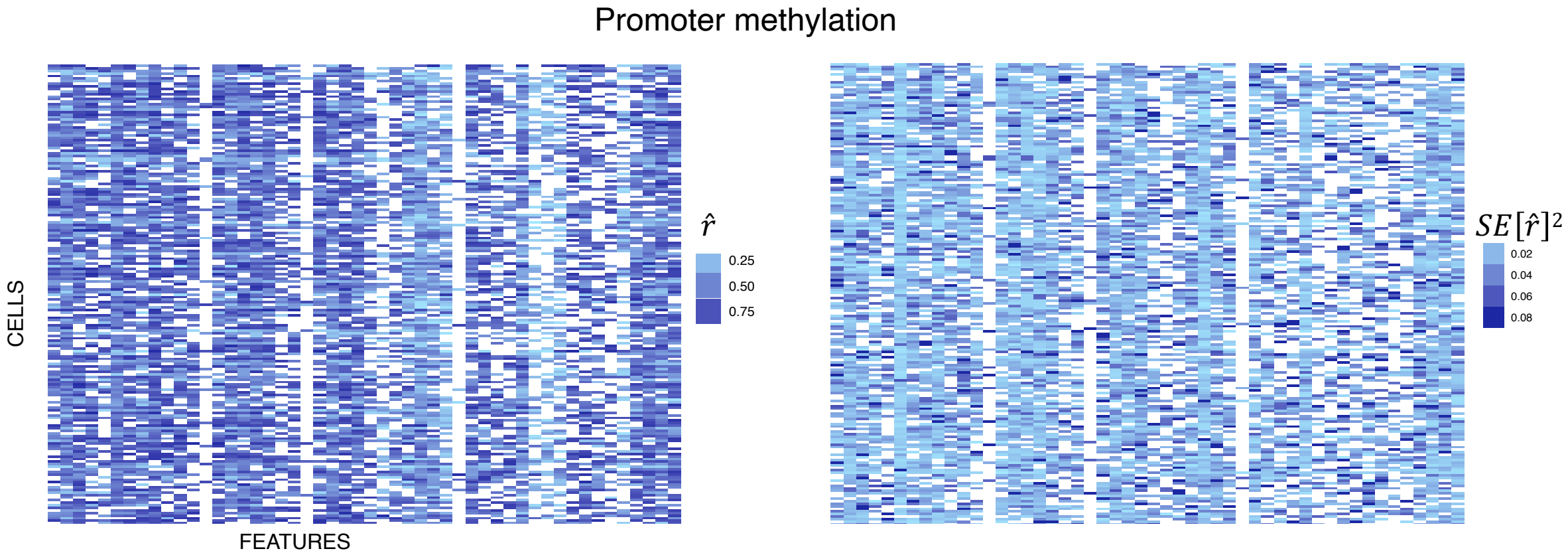
Argelaguet *et al*. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 576, 487–491 (2019)

Heterogeneity in size of datasets

	P	N
rna	14501	815
met_genebody	15837	815
met_promoter	12092	815
met_cgi	5536	815
met_p300	101	815
acc_DHS	290	815
acc_genebody	17139	815
acc_promoter	16518	815
acc_cgi	4459	815
acc_p300	138	815
acc_DHS	290	815

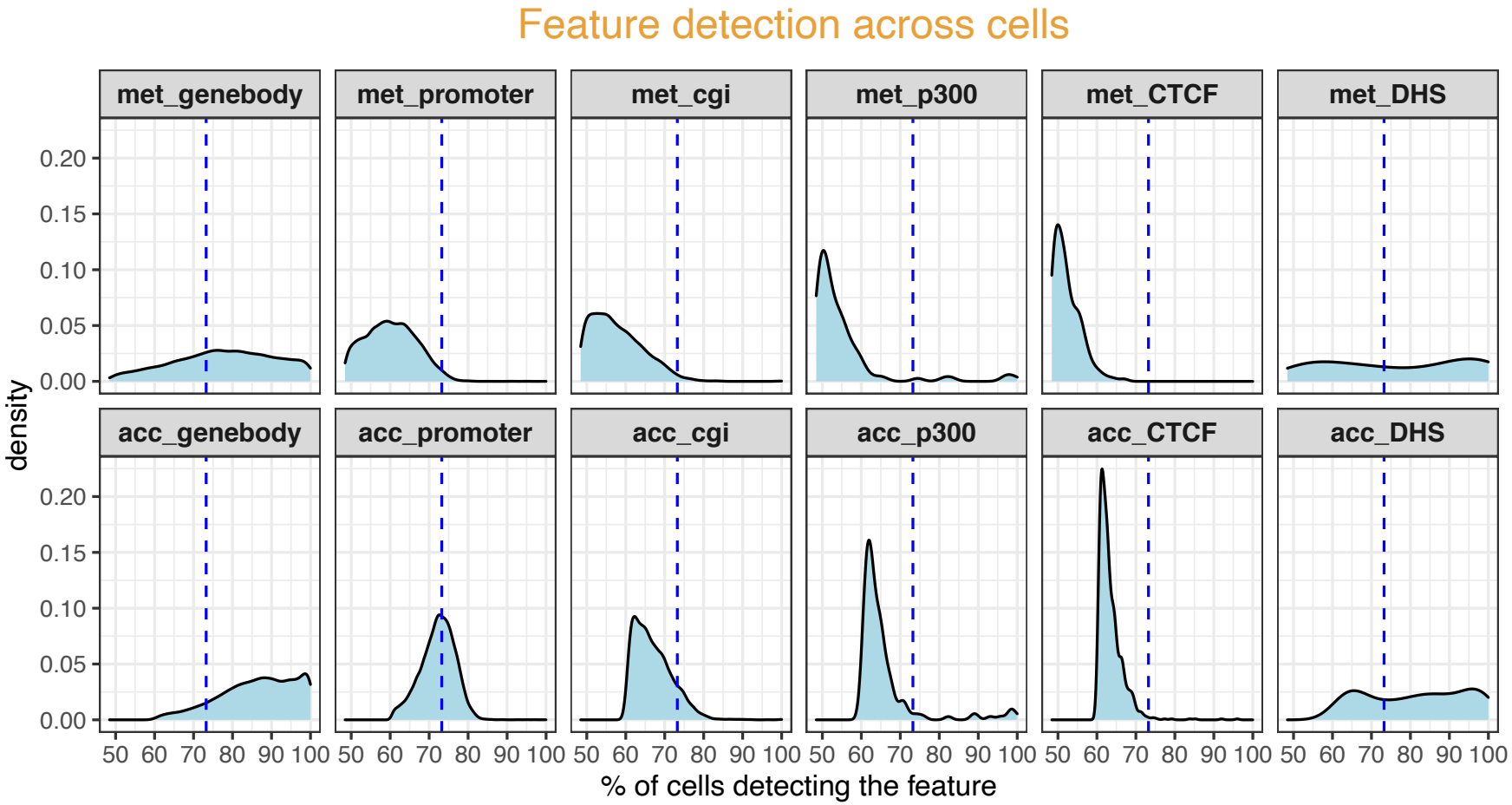
- Integrative analyses could be sensitive to size of te datasets

DNA-level estimates are sparse and noisy



- Missing values (dropouts) due to limited coverage
- Estimates vary in levels of uncertainty

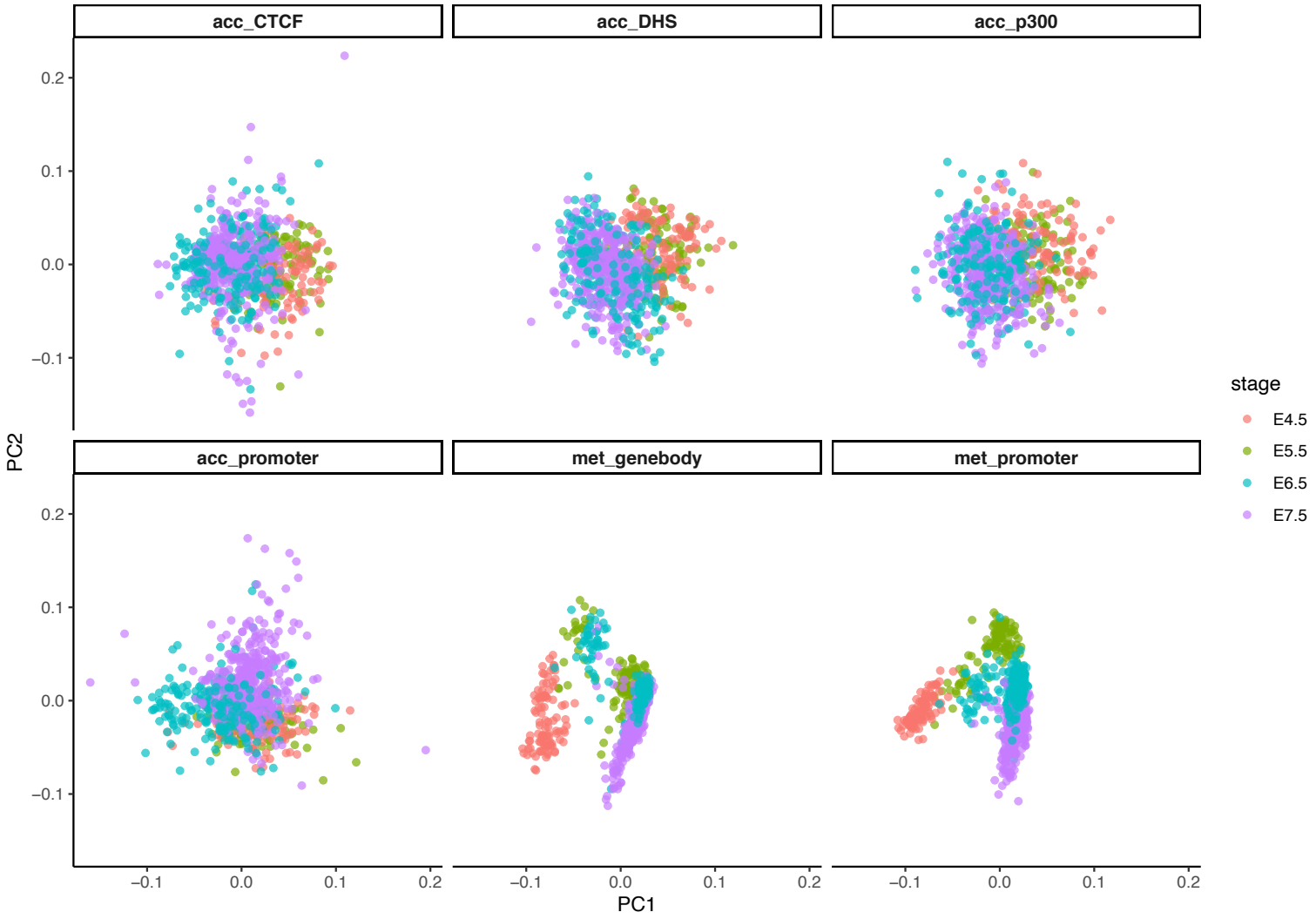
DNA-level estimates are sparse and noisy



https://en.wikipedia.org/wiki/CpG_site

- Methylation modalities have consistently lower feature detection

DNA-level estimates are sparse and noisy



Challenge: effective noise-weighting required to enhance the biological signal

Current methods for single-cell multi-modal analysis

- Most integrative methods developed for integration of **multiple scRNA-seq** datasets
- **Seurat** uses shared latent space (CCA) to identify corresponding cells in query datasets. Takes one dataset as ‘reference’.
- **MOFA** and **LIGER** use NMF to get **shared and data-specific** axes of variation across modalities.

MOFA:

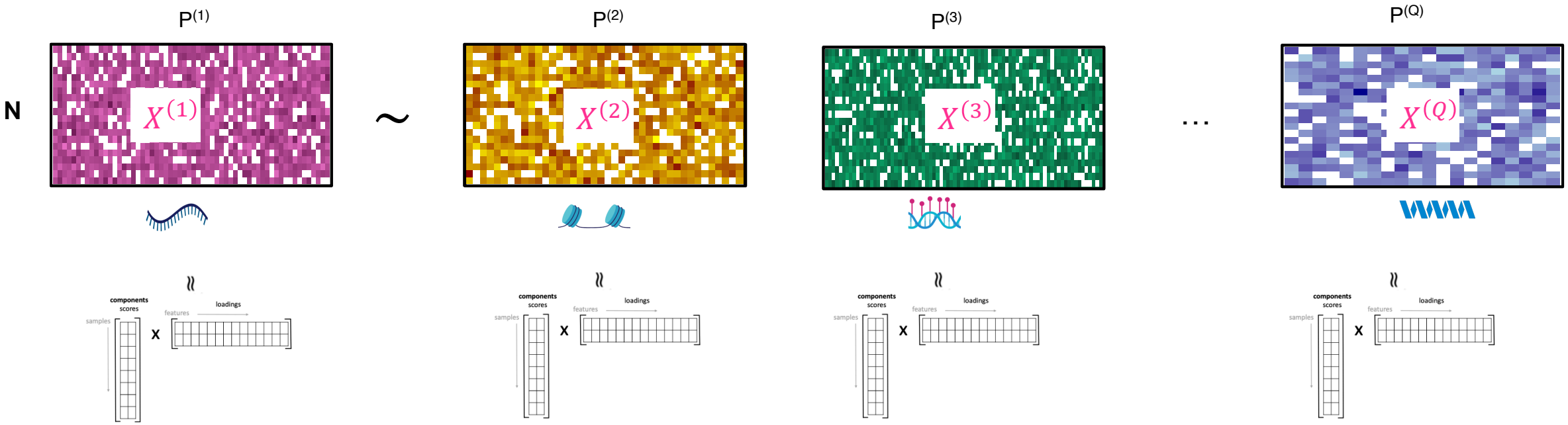
- ❑ Works best when data are assumed to follow a Gaussian distribution (although other models are supported)
- ❑ Learned factors could be sensitive to initialization and could be biased towards larger datasets (size homogenisation required)

LIGER:

- ❑ Needs shared features across modalities

Argelaguet et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 21, 111 (2020)
Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411–420 (2018)
Liu et al. Jointly Defining Cell Types from Multiple Single-Cell Datasets Using LIGER. bioRxiv 2020.04.07.029546

Multi-modal sparse PLS integrative approach



$$\max \sum_{i=2}^{Q} cov(X^{(1)}a^{(1)}, X^{(i)}a^{(i)})$$

$$s.t.\ \ \|a^{(q)}\|_2 = 1 \ \ and \ \ \|a^{(q)}\|_1 \leq \lambda^q \ for \ \ all \ \ 1 \leq q \leq Q$$

- Variable-selection using the LASSO
- Able to handle (ignore) missing values without the need to impute

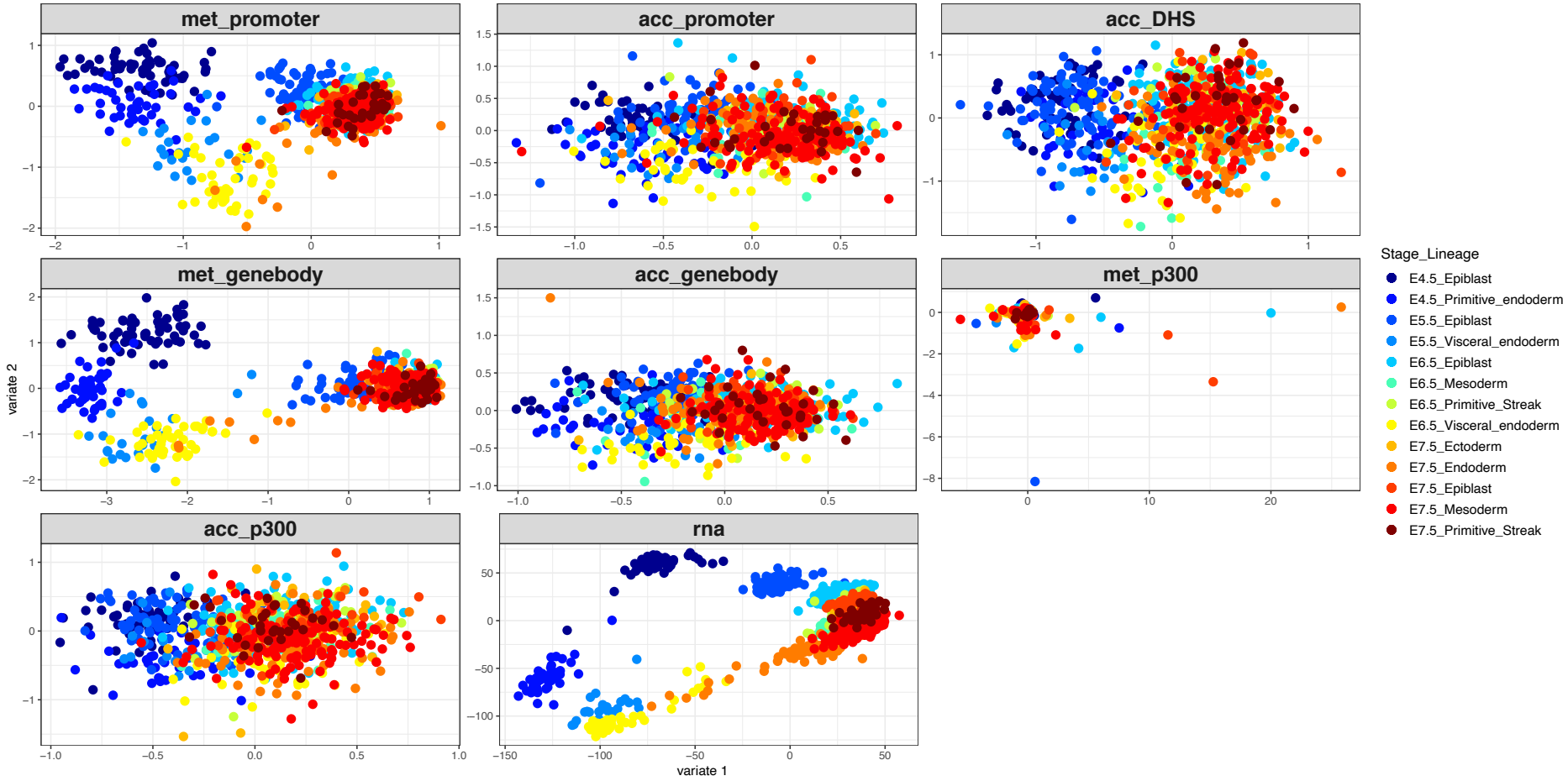
Rohart et al. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS computational biology (2017)

Data integration & feature selection

Transcriptome ~ DNA Methylation & Chromatin Accessibility

# of components: **2**

# of features selected per component: **variable (min 25, max 50)**



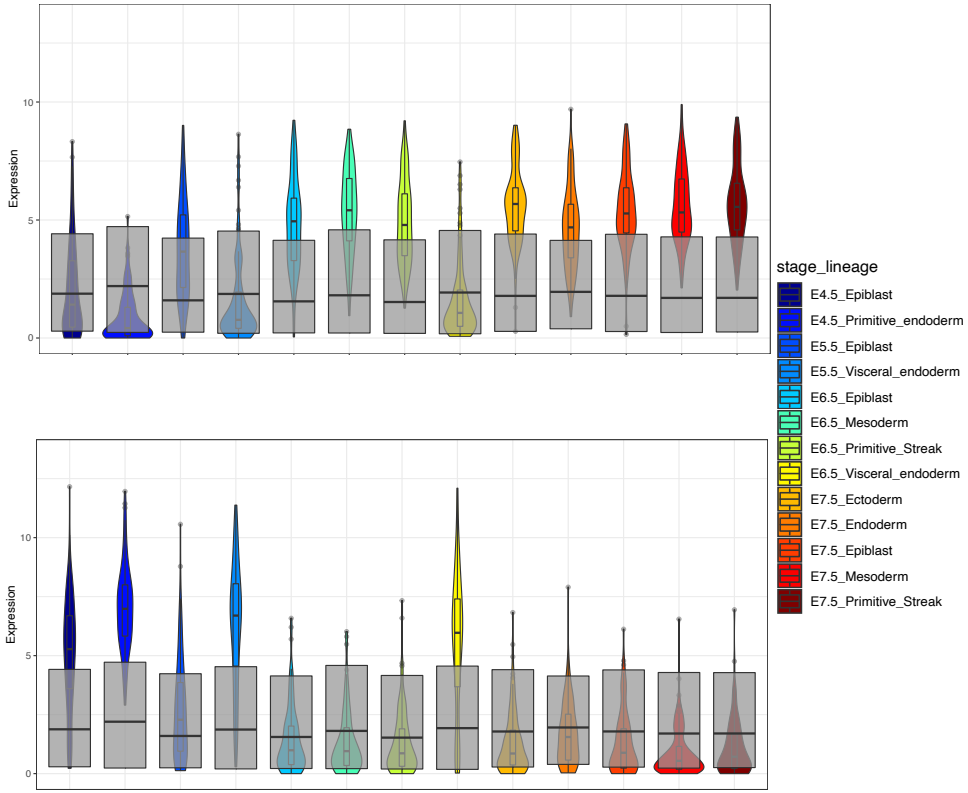
- Coordinated variation mainly driven by stages and lineages
- Different modalities contain different levels of covariance with transcriptome (or none!)

Data integration & feature selection

RNA markers (component 1)

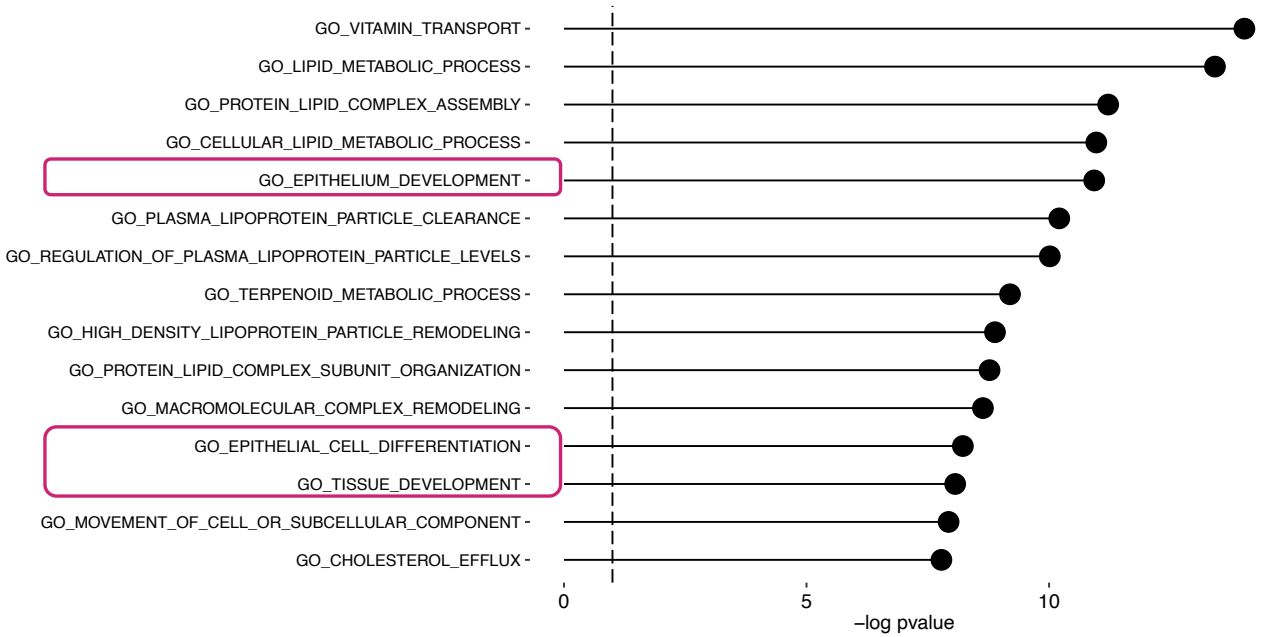
Positive

coefficient on loading vector



Negative

coefficient on loading vector

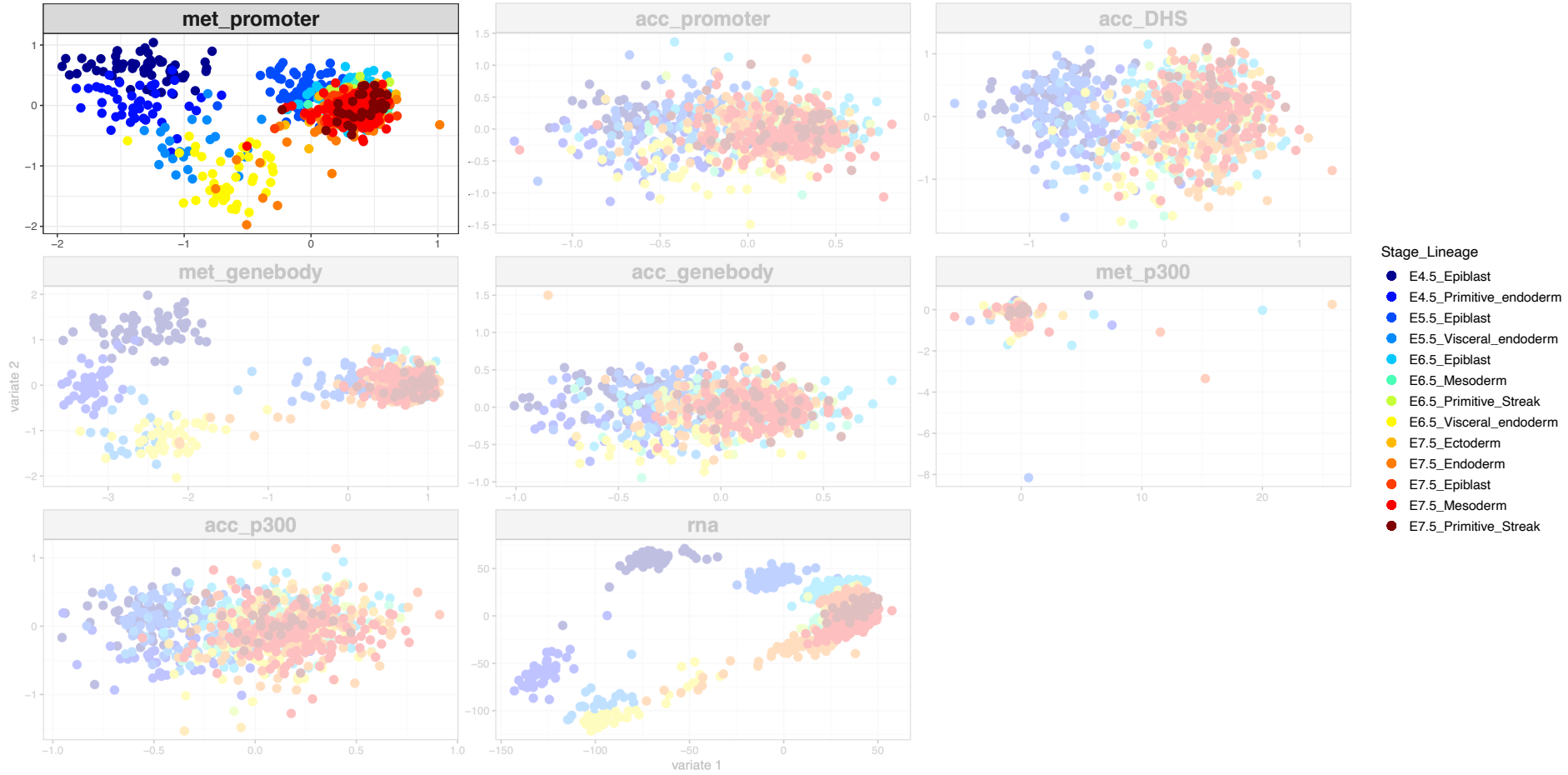


https://bioconductor.org/packages/release/data/experiment/html/MOFAdata.htm
https://github.com/bioFAM/MOFA2

- First RNA component selects for genes that either switch on or switch off in later stages

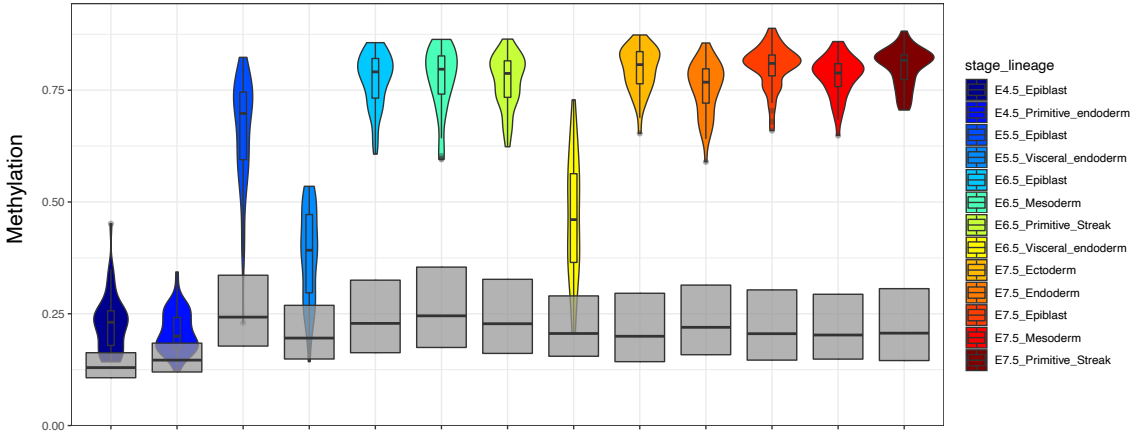
Data integration & feature selection

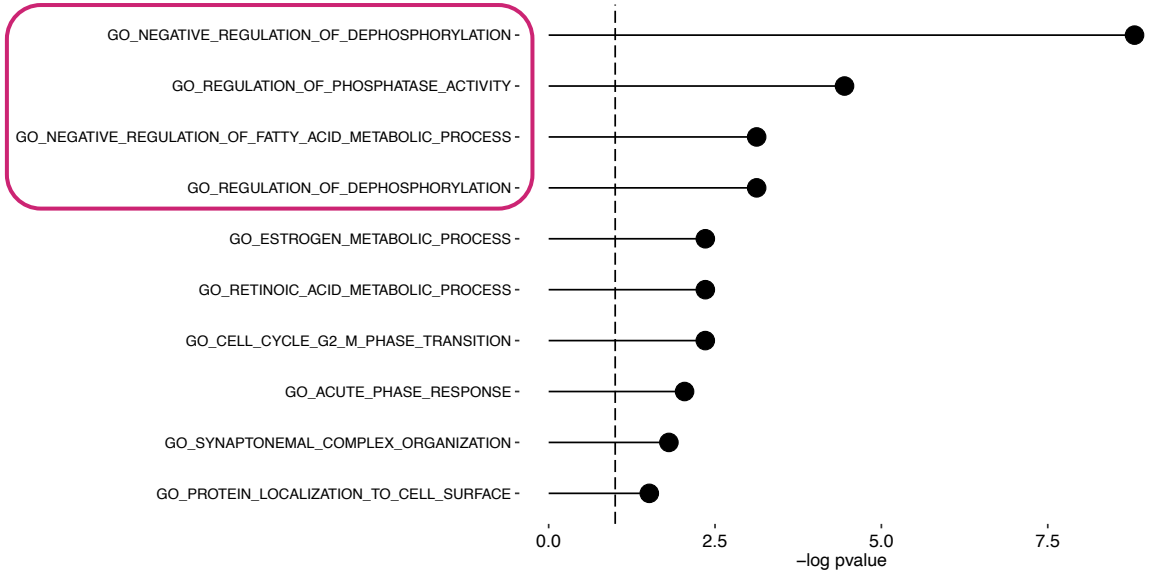
Promoter methylation markers



Data integration & feature selection

Promoter methylation markers (component 1)





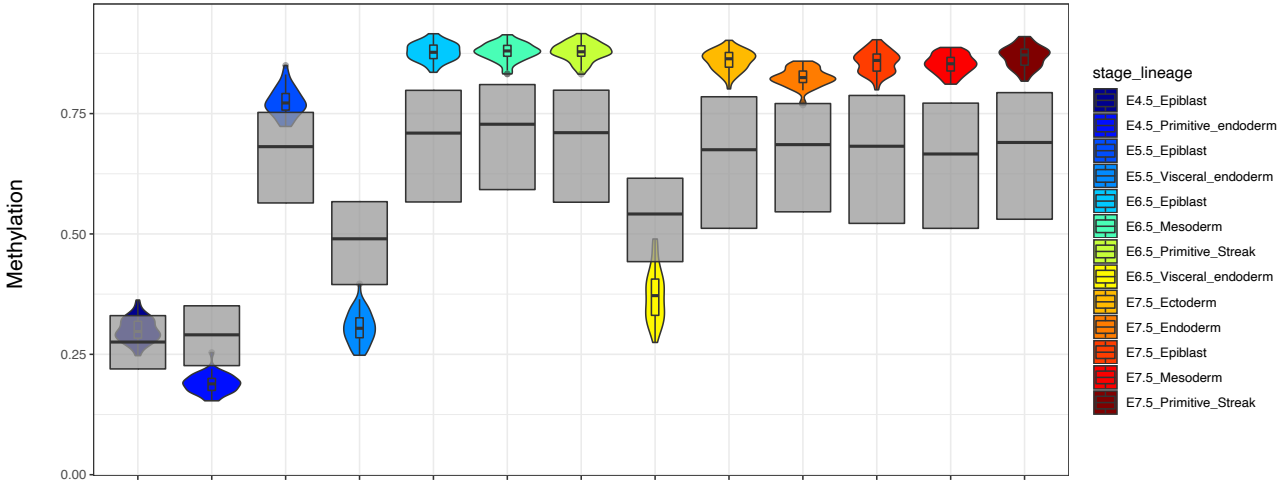
- Selected promoter regions are hypermethylated in late-stage embryonic cells
- Enrichment of regulatory pathways in selected promoter regions

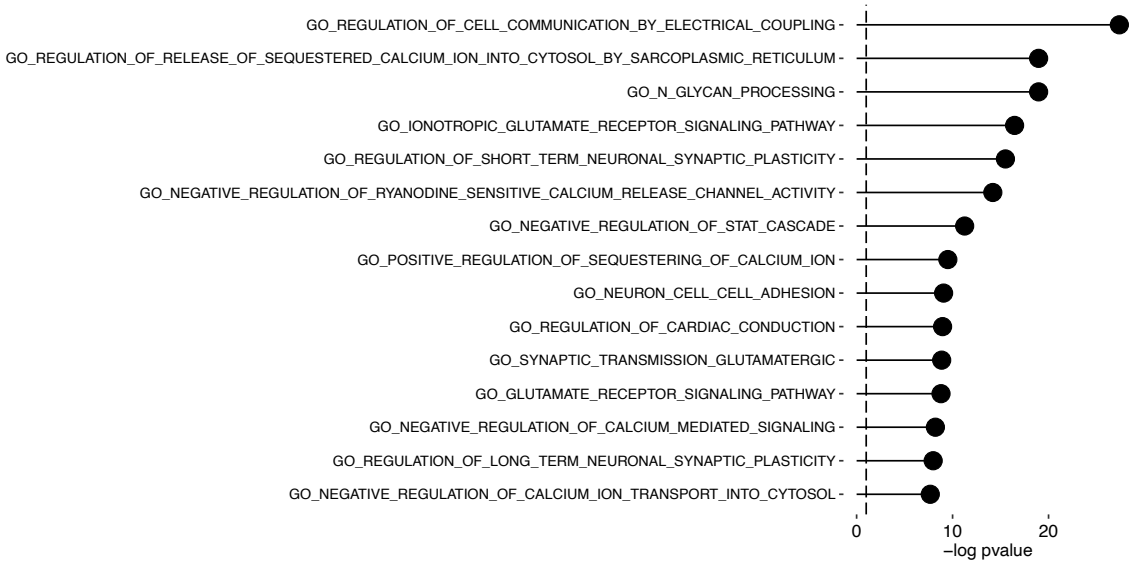
Data integration & feature selection

Genebody methylation markers

Data integration & feature selection

Genebody methylation markers (component 1)

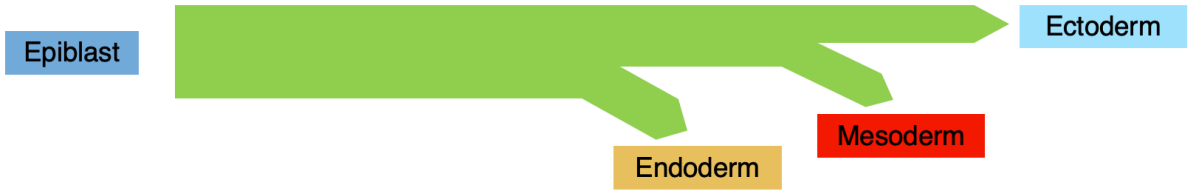




- Selected genebody regions follow the global methylation behaviour but more strongly
- Hypomethylation of selected regions in primitive endoderm cells

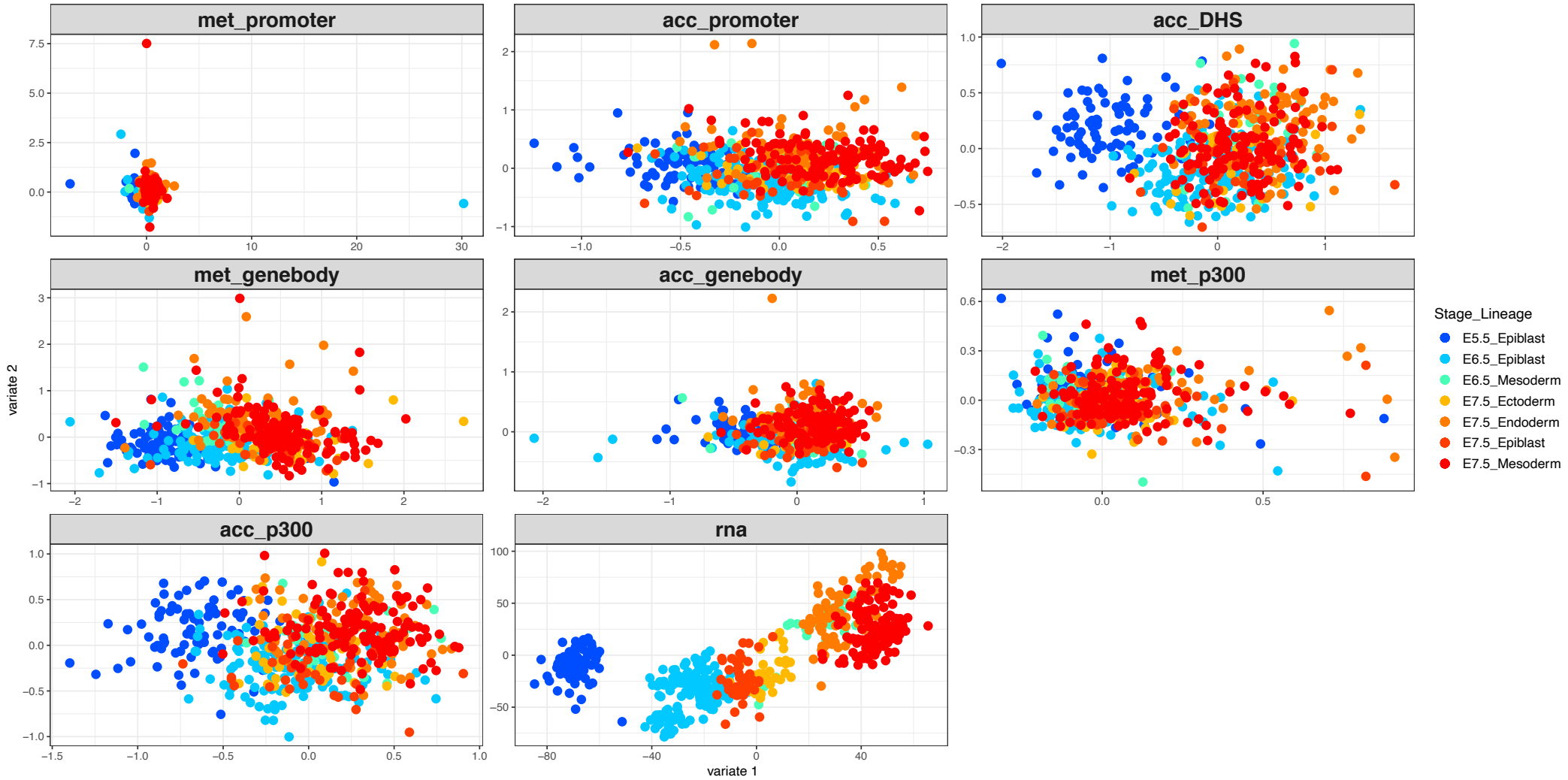
Focusing on lineage specification

	E4.5	E5.5	E6.5	E7.5
Ectoderm	0	0	0	43
Endoderm	0	0	0	81
Epiblast	60	84	146	44
Mesoderm	0	0	28	141
Primitive_endoderm	43	0	0	0
Primitive_Streak	0	0	43	33
Visceral_endoderm	0	24	45	0



- A total of 567 epiblast and germline cells

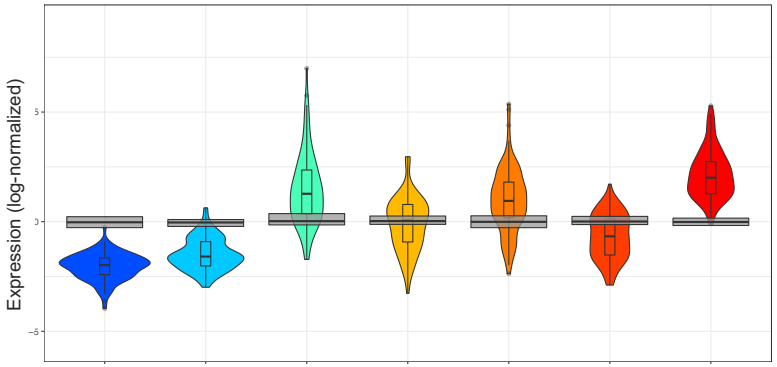
Focusing on lineage specification



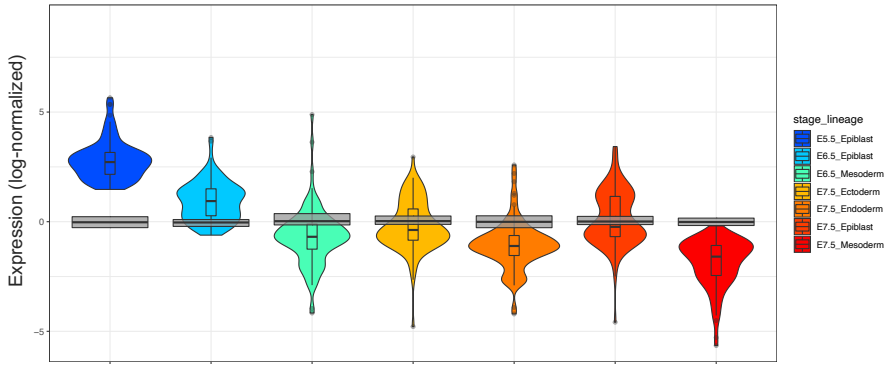
- Embryonic stages are still the main driver of shared variation
- Promoter methylation shows less coordinated variation when considering late-stage cells. It could be caused by abundant missing values.

Focusing on lineage specification

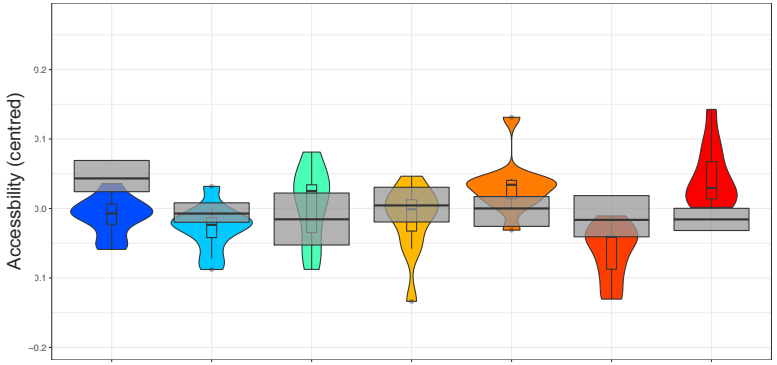
RNA markers (component 1 - positive loadings)



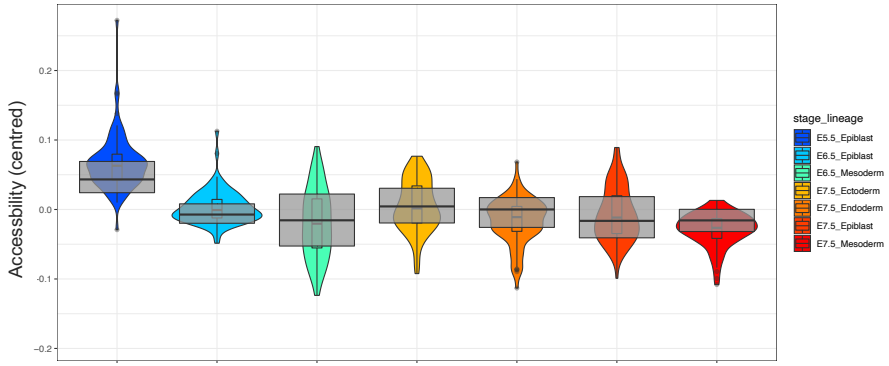
RNA markers (component 1 - negative loadings)



enhancer markers (component 1 - positive loadings)



enhancer markers (component 1 - negative loadings)



- Putative mesoderm enhancer markers are less accessible in early stage and more accessible in late stage

Summary

- Multi-modal sparse PLS approach integrates multiple modalities of various sizes
- The selected markers mainly characterise the embryonic stages
- Modalities differ widely in their level of covariance with the transcriptome and potentially capture different biological interactions
- So far most of integrative analyses focused on coordinated variation with respect to transcriptome, although it is possible to investigate the interaction between all modalities

Limitations & Challenges

- Only looks at shared axes of variation
- Needs observations on same set of cells
- Needs continuous variables as input
- Where to summarise the calls? Open question

Future work

- Performing supervised integration using the epigenetic data and the assigned lineages
- Investigate the inclusion of weights for each data set in the integrative approach
- Benchmarking against current methods
- Investigate manifold learning using the learned components

Acknowledgements

- **Ricard Argelaguet**, European Bioinformatics Institute; University of Cambridge
- All **LêCao lab** members
- Heather Lee, The University of Newcastle; Hunter Medical Research Institute
- Elizabeth Mason, Hogan lab, Peter MacCallum Cancer Centre
- Aleksander Dakic, LêCao lab, The University of Melbourne
- Attila Csala, Academic Medical Center, Amsterdam

THE UNIVERSITY OF MELBOURNE

Australian Government
Australian Research Council

CHAN ZUCKERBERG INITIATIVE