THE UNIVERSITY of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Probabilistic Modelling of Single Cell Multi-Omics Data

Christos Maniatis

Doctor of Philosophy

Institute of Adaptive and Neural Computation
School of Informatics
University of Edinburgh

September 2022

Dedicated to my brother Fivos
and
my loving parents

Declaration

I declare that this thesis has been composed by myself and that this work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated at the beginning of each chapter. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Edinburgh, 2022

Christos Maniatis
September 2022

Acknowledgements

I would like to express my immense gratitude to my supervisors, Guido Sanguinetti and Catalina Vallejos, for their guidance and support throughout the thesis. Guido's guidance and insights, especially in the early stages of thesis, helped me overcome many challenges of the academic landscape. Catalina has been a pillar of support and a beacon of guidance. Her experience has refined this work and shaped it to a comprehensive whole. Both my supervisors' patience and ability to explain in simple terms convoluted ideas has helped me understand core concepts vital to this work and shaped the way I approach research.

I would also like to thank Michael Gutmann for his constructive criticism during the annual reviews. Furthermore, I would like to extend my appreciation to Chantriolnt-Andreas Kapourani and Michalis Michaelides for the fruitful research discussion.

During the PhD, I had the privilege to be a member of CDT in Data Science vibrant research community along with groups that Catalina and Guido have assembled. I am grateful to the people populating these groups for their stimulating conversations broadening my understanding around a plethora of research topics. In particular, I would like to thank Allan O'Callaghan for his support around technical matters and Christine Simpson for the thought provoking discussions.

A special thanks goes to all people accompanying me in this journey. In particular, I am grateful to Christos and Takis for our political and philosophical debates, which kept me sane and shaped my opinion regarding contemporary problems of humanity. I am also grateful to Kostas for affording me every comfort during rough days. I am indebted to Maria for her love, understanding and her continuing patience. Finally, I would like to express my gratitude to my family for their unwavering mental and material support.

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

Abstract

Multicellular organisms possess a diverse set of cells exhibiting unique properties and function. Despite their physiology and role each cell owns the same copy of genetic instructions encoded in its DNA. The ability of cells to differentiate into various shapes and forms stems from a careful orchestration of gene expression through various regulatory mechanisms.

Recent developments in single cell multi-omics protocols offer unprecedented opportunities to simultaneously quantify phenomena in epigenome and gene expression at a single cell resolution. Advances in cell isolation and barcoding eliminated various confounding phenomena, shedding light into the regulatory role of epigenome in gene expression over diverse tissues and cells. Yet, combining omics modalities introduces serious statistical and computational challenges. Limitations of single-omics get exacerbated when combined into multi-modal assays, making result interpretation hard. In this thesis, we argue that inconsistent treatment of technical variability offered by classical statistical tools can corrupt statistical analyses and produce misleading results. In the Bayesian template, we introduce probabilistic models that explicitly and transparently decouple technical variability from biological signal. These methods are then used to investigate how epigenetic regulatory mechanisms interact with gene expression, both at genomic and at a cellular level. Single cell sequencing technologies are notoriously affected by high sparsity, leaving scientists to wonder if data are a product of sample handling or some genes are not expressed. As a result, even simple correlative tools (eg. Pearson's correlation) seeking to identify regions with strong regulatory patterns between molecular layers routinely pinpoint a handful of associations. To overcome some of these limitations we introduce SCRaPL (Single Cell Regulatory Pattern Learning), a Bayesian hierarchical model to infer correlation between different omics components. SCRaPL's uncertainty quantification allows for accurate results and good control over false positives, compared to its counterparts.

Existing limitations force practitioners to partially or fully discard molecular modalities from cell observations, significantly under-powering subsequent downstream analysis. An alternative solution for scaling datasets is to post-experimentally address protocol limitations using a generative model. We introduce single cell Multi View Inference (scMVI), a deep

learning model designed to accommodate analyses on both partially and fully observed data. Using jointly quantified data, scMVI builds a low-dimensional joint latent space by aligning omcis representations for each cell. In similar cells, scMVI can match individual modalities creating more complex sets. Subsequently, this manifold is used to approximate the data generating process. Hence, in partially quantified cells missing observations could be imputed getting the full potential of the data.

To summarize, this thesis proposes novel statistical tools to interpret the regulatory interactions between epigenome and gene expression using data from modern multi-omics sequencing experiments. Their flexible design along with robust uncertainty quantification, allow these methods to unlock the immense potential of existing and future sequencing protocols. We hope that with the increased adoption in these methods, SCRaPL and scMVI will become an integral part of downstream analysis.

Lay Summary

How is it that identical twins, who inherit the same set of genes end up with different physical characteristics, behaviour or predisposition to disease? Why are offspring of a 77 year old Dutch cohort more likely to suffer from diabetes, obesity and schizophrenia compared to the general Dutch population? The cause of these differences is rooted in heritable changes in gene expression along with other stable long term alternations rather than underlying DNA as originally hypothesized. Expressed genes ultimately affect protein production which enables a series of processes vital to cells.

To better understand the role of epigenome in gene expression scientist have pioneered technologies which enable to jointly measure various epigenomic marks like DNA methylation, chromatin accessibility with gene expression across a series of cells. These data have fueled research striving to shed light on how the epigenome gives rise to cells with various phenotypic properties, under what circumstances diseases can appear and how epigenome can be used to prevent or cure life-threatening diseases.

Modern data generation technologies have brought up a series of computational problems. Firstly, the low amount of source DNA combined combined with low sensitivity, often produce datasets with high levels of dropouts and large number of zeros. Secondly, data specific variability can mask DNA regions where epigenome strongly affects gene expression producing misleading results.

In this thesis we use the Bayesian framework to devise noise aware methods aimed at analyzing jointly profiled epigenomic and gene expression data for existing and future generating technologies. We begin by investigating the interplay between epigenome and gene expression in various regions of the DNA, to demonstrate that careful account of different noise sources can improve signal extraction compared to traditional statistical approaches. Then we introduce a method capable of analyzing both fully and partially available cellular data. By identifying likely candidates for missing values, this approach helps to post-experimentally address some of the outlined technology limitations present in data.

Overall, this thesis contributes novel statistical methods that emphasize the importance of accounting for different types of noise when examining the role of epigenome in gene

expression. Their flexible design along with robust noise quantification, allow these methods to unlock the immense potential of epigenomic and gene expression data. We hope that with the increased commercialization and adoption of modern data generating technologies the proposed methods will become an integral part of biological and biomedical analysis tools.

Ibis, redibis nunquam per bella peribis

Table of contents

List of figures **xix**

List of tables **li**

Nomenclature **liii**

1 Introduction **1**
1.1 Contributions 3
1.2 Thesis Layout 3

2 Molecular biology and epigenetics **5**
2.1 Foundations of molecular biology 5
2.2 Regulation of Gene Expression 6
2.2.1 Regulatory elements of gene expression 8
2.2.2 Epigenetic regulation of gene expression 9
2.2.2.1 Chromatin organisation and Histone modification 10
2.2.2.2 DNA Methylation 13
2.3 Sequencing technologies 14
2.3.1 Fundamentals of of single cell omics and multi-omics protocols . . 15
2.3.2 Gene expression sequencing 16
2.3.3 Genome accessibility sequencing 17
2.3.4 DNA methylation sequencing 18
2.3.5 Single cell multi-omics protocols 19

3 Mathematical Background **23**
3.1 Statistical Modeling - The Bayesian Perspective 23
3.1.1 Bayesian Inference 24
3.1.2 Hierarchical Models 26
3.1.3 Model Selection 27

3.2 Multiple Hypothesis Testing 29
3.3 Probabilistic graphical models 31
3.4 Approximate Bayesian Inference 33
3.4.1 Hamiltonian Monte Carlo 34
3.4.2 Assessing the quality of MCMC outputs 36
3.5 Function Approximation 39
3.5.1 Variational Autoencoders 40

4 Computational methods for single cell multi-omics data 47
4.1 Introduction 47
4.2 Integration 49
4.3 Downstream analysis 53
4.3.1 Downstream analysis in single cell single-omics 54
4.3.2 Downstream analysis in single cell multi-omics 54
4.3.3 Downstream Challenges 55

5 SCRaPL: A Bayesian hierarchical framework for detecting technical associates in single cell multiomics data 57
5.1 Datasets 58
5.2 Methods 60
5.2.1 SCRaPL model 60
5.2.2 Prior Specification 63
5.2.3 Implementation 67
5.2.4 A probabilistic rule to detect statistically significant across layers . 68
5.2.5 Current approach based on Pearson/Spearman correlation. 68
5.2.6 Comparing between alternative models 69
5.3 Experimental setup 70
5.3.1 Synthetic data experimental design 70
5.3.2 Comparisons on real data 73
5.4 Results 75
5.4.1 Benchmarking SCRaPL on synthetic data 75
5.4.2 SCRaPL improves the power to identify associations between molecular layers in mouse embryonic stem and brain cells 76
5.4.3 SCRaPL associations are influenced by data sparsity and are robust to outliers. 80
5.4.4 SCRaPL identifies biologically meaningful epigenetic regulation in early mouse gastrulation 81

5.5 Discussion 83

6 scMVI: A deep generative model for integrated analysis of single-cell multiomics data 85

6.1 Datasets 86
6.2 Methods 87
6.2.1 scMVI: Single Cell Multi View Inference 89
6.2.2 Related Methods 94
6.3 Experimental setup 96
6.4 Results 97
6.4.1 scMVI achieves effective integration accurate and probabilistic imputation. 97
6.4.2 scMVI leverages effectively multi-source information to build biologically meaningful latent spaces. 101
6.4.3 Exploratory analysis using Differential directions 107
6.5 Discussion 113

7 Discussion 115

7.1 Future Directions 116
7.1.1 Accounting for genomic region interaction 117
7.1.2 Transferring knowledge across protocols 118

Appendix A Supplementary information for SCRaPL 121

A.1 Synthetic Data 121
A.2 Negative control experiments 130
A.3 Extended comparisons between SCRaPL, Pearson and Spearman predictions. 133
A.3.1 Classified as significant by SCRaPL and Pearson/Spearman 133
A.3.2 Classified as insignificant by SCRaPL and significant by Pearson . 135
A.3.3 Classified as significant by EFDR and insignificant by FDR 138
A.3.4 Posterior correlation inference for common genes in bibliography . 139
A.4 Gene Set Enrichment Analysis. 141
A.5 Mathematical Derivations 142
A.6 Efficiency Analysis 146
A.7 Macroscopic analysis of SCRaPL's behavior compared to Pearson and Spearman Correlation. 147
A.8 Using SCRaPL as a data denoising tool 148
A.9 Convergence Analysis 154

A.9.1 SCRaPL with Inflation on mESC 155
A.9.2 SCRaPL without Inflation on mESC 159
A.9.3 SCRaPL with Inflation on mEBC 164
A.9.4 SCRaPL without Inflation on mEBC 168
A.10 Posterior predictive checks 173
A.10.1 Posterior predictive checks for mESC 173
A.10.2 Posterior predictive checks for mEBC 181
A.11 SCRaPL without over-dispersion 189
A.12 Prior Sensitivity 192
A.12.1 Prior Sensitivity for mESC data 192
A.12.2 Prior Sensitivity for mEBC data 201

Appendix B Supplementary information for scMVI 211
B.1 Imputing missing data 211
B.2 Comparison with related methods 223
B.3 Exploratory Analysis 230

Bibliography 253

List of figures

2.1 **Core concepts of molecular biology.** (2.1a) Comparison between DNA double helix (left) with RNA single strand; figure taken from Sponks/Wikimedia Commons/CC BY-SA 3.0. (2.1b) Central dogma of biology illustration explaining the flow of genetic information from nucleus to protein production; figure taken from Narayanese at English Wikipedia, Public domain, via Wikimedia Commons 7

2.2 **Regulation of transcription sketch in mammals.**Transcription factors interact with sequences of DNA to promote and enhance transcription. Activating transcription factors are epigenetic regulators of specific genes. DNA bending protein brings the activators into the proximity of the promoter where additional transcription factors are bound and interact with the activators. Transcription factors at the promoter recruit RNA polymerase and transcription can proceed (Mobley, 2019). Figure taken from Bernstein0275/Wikimedia Commons/CC BY-SA 4.0 10

2.3 **Waddington Epigenetic Landscape for human ESC.** A metaphor representing various differentiation pathways a cell might make during development. For cell to change phenotypic properties it has to traverse through valleys and ridges from one steady state to the next. Figure taken from Simon Elsasser/PD. 11

2.4 **Units of chromatin organization.** Histones (H2A, H2B, H3, H4) first bind to each other to form H3–H4 and H2A–H2B dimers, and the H3–H4 dimers combine to form tetramers. An H3–H4 tetramer then further combines with two H2A–H2B dimers to form the compact octamer core, around which the DNA folds (Alberts et al., 2007). Successive nucleosomes are connected together into higher order structure resembling the “beads of string" which are folded in 30*nm* compact fibre. Compact fibres containing tens of thousands of nucleotide pairs, are further organized into a series of looped domains (Alberts et al., 2007). Figure adapted from David O Morgan/Wikimedia Commons/CC BY-SA 3.0 12

2.5 **Sequencing technologies discussed .** Sequencing protocols used to interrogate epigenomic regulation of transcriptome at bulk and sigle cell resolution. 21

3.1 **D-separation summary for variables and sets of variables.** In figures 3.1a and 3.1b paths from z_1 to z_2 are blocked by z_3 using d-separation rule (1.), hence $z_1 \perp\!\!\!\perp z_2 | z_3$. Since path from z_1 to z_2 is not blocked by z_3 using d-separation rule (2.), $z_1 \not\perp\!\!\!\perp z_2 | z_3$. (3.1d) The example extends D-separation to sets of variables. Drawing paths from members of W_1 to members of W_2 we deduce that $W_1 \perp\!\!\!\perp W_2 | W_3$. 32

3.2 **HMC explanatory figures** (3.2a) For a target density $P(z)$ an effective proposal should generate many samples from the orange area and some from gray areas. (3.2b) Sketch of HMC accepted (green) or rejected (red) trajectories in the momentum-parameter space for a 1D Gaussian distribution. At the end of each trajectory momentum is negated, hence the vertical jump. If state is accepted, a new momentum value is sampled form a normal distribution and the leapfrog simulates the numerical trajectories while conserving the total energy. 37

3.3 **Variational Auto Encoder graphical model.** Combination of hierarchical graphical models and deep learning. Solid lines denote the generative model $\mathbb{P}_\phi(z, \mathscr{D})$, dashed lines the variational approximation $\mathbb{Q}_\theta(z|\mathscr{D})$ to the intractable posterior $\mathbb{P}_\phi(z|\mathscr{D})$. Variational parameters ϕ, θ are jointly learned during training. Figure and caption taken from Kingma and Welling (2014). 41

5.1 **Schematic representations of SCRaPL.** SCRaPL combines error models tailored to single cell multi-omics data with Bayesian hierarchical frameworks to discover correlative associations between epigenomic and transcriptomic modalities in different genomic regions. 60

5.2 **Graphical representations of SCRaPL fitted to each genomic region $j \in J$ with mESC and mEBC data.** Observed genomic data (Y_{ij1} is RNA expression; Y_{ij2} is Chromatin Accessibility), with their associated latent variables (Z_{ij1}, Z_{ij2}) and feature-specific model parameters $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ capturing feature specific epigenomic and transcriptomic marks. Additional inflation parameters π_j and $\tilde{\pi}_j$ are specific to Gene Expression and Chromatin accessibility error models. For the case of mESC data where epigenome is measured differently, parameter $\tilde{\pi}_j$ is dropped. 62

5.3 **Plots used to tune prior correlation parameters.** The decision if a stronger or a weaker prior better suit our data, is based on a set of three scatter plots drawn using the original and a negative control dataset. The first two look at posterior median correlation as a function of % of gene expression zeros in the mESC data (Figure 5.3a) and corresponding negative control (Figure 5.3b) data for strong prior $\mathrm{TB}(15,15)$ (left), weak prior $\mathrm{TB}(3,3)$ (middle). For comparison we also plot Pearson correlation as a function of % of zeros (right). Each dot represents a feature. Negative control dataset, where no strong correlations are expected goes against what we observe. Therefore, we chose the stronger prior. 65

5.4 **Distribution of WAIC and DIC difference between model with and without inflation for mESC and mEBC data across features. Evidence demonstrating the importance of zero inflation.** The more negative the difference, the stronger the evidence in favor of the model with zero inflation on the gene expression component for a particular feature. As a visual reference, zero is marked with dashed red line. 71

5.5 **Distribution of WAIC difference between models with and without dispersion for mESC and mEBC data across features. FIgures demonstrating the importance of dispersion.** The more negative the difference, the stronger the evidence in favor of the model with dispersion on the gene expression component for a particular feature. As a visual reference, zero is marked with dashed red line. (5.5a) WAIC distribution of difference between ZIFP and Normal likelihood in mESC data. 5.5b) WAIC distribution of difference between ZIFP and Normal likelihood in mEBC data.(5.5c) WAIC distribution of difference between Poisson and Normal likelihood in mESC data. 5.5d) WAIC distribution of difference between Poisson and Normal likelihood in mEBC data. 72

5.6 **Sketch of tensors storing methylation/accessibility(left) and expression(right) respectively. Visual explanation of the negative control experiments.**(5.6a) Black boxes represent missing methylation/accessibility data. (5.6b) Sketch of cell scrambling for a random feature j. It is important to note that despite cells are permuted at the end of scrambling, cell labels remain fixed. 74

5.7 **Plots summarizing differences in correlation estimation between SCRaPL, Spearman and Pearson in Experiment 1 with synthetic data across features**. (5.7a) Estimated correlation difference from true correlation as a function of cells for SCRaPL, Spearman and Pearson. (5.7b) Estimated correlation as a function of true correlation for SCRaPL, Spearman and Pearson in synthetic datasets with 300 genes and 1600 cells. Each dot represents a gene and is color-coded based inference approach. 77

5.8 **Summary of experiments on real data. Plots demonstrating that SCRaPL can significantly improve detection rates, by choosing features picked up by Pearson and Spearman correlation while avoiding spurious correlations.** (5.8a,5.8b) Bayesian volcano plots for mESC and mEBC data respectively. Scatter plot of posterior probability under the null hypothesis (in log scale) as a function of posterior median correlation. Each dot represents a feature and is marked with different color depending the method that labels it as a significant association. (5.8c,5.8d) Venn diagrams summarizing detection rates for SCRaPL, Pearson and Spearman in mESC and mEBC data. By accounting for different sources of noise it detects a large set of features identified by Pearson and Spearman correlation. SCRaPL also uncovers a additional large set that would be impossible for frequentist methods to identify in a robust way. 79

5.9 **SCRaPL's behavior compared to Pearson/Spearman correlation in micro and macro scale. Specific features demonstrate examples where frequentist approaches agree and disagree with SCRaPL's estimates. The scatter plots show the link between coverage and zero inflation and distance between frequentist and SCRaPL correlation.**. In all figures apart from 5.9d the scatter plot depicts raw data for chosen features color-coded by CpG coverage, and normalized expression plotted in the $log(1+x)$ scale. The violin plots depict the posterior correlation densities estimated by SCRaPL for the raw data in their left hand side. (5.9a) Example where both SCRaPL and Pearson/Spearman identify the feature's association as significant. (5.9b) Example were only Pearson/Spearman identifies the feature's association significant. (5.9c) Example were only SCRaPL identifies the feature's association significant. (5.9d) Scatter plots to demonstrate the negative/positive relationship between alternative correlation estimates and CpG coverage/% zeros in expression respectively. ($\bar{\rho}_j$ and ρ_{prs} in figure 5.9d are posterior mean and Pearson correlation for feature j.) 82

6.1 **scMVI schematic representation. This is to demonstrate the how the model looks like.** scMVI is a deep generative model designed to handle partially and fully observed single cell multi-omics data. DNAm, gene expression and chromatin accessibility are non-linearly encoded to modality dependent low dimensional manifolds. During training, the manifolds are aligned using Jeffrey's divergence. In the generative step, a modality specific latent representation and decoder is used to reconstruct observations. This graph is given, as the way scMVI is defined and trained has some differences to traditional VAEs. 88

6.2 **scMVI cross modal schematic representation. This is to demonstrate how scMVI imputes missing profiles.** Schematic depiction of cross modal imputation. Using relevant encoder(s), we project data to create a latent space representation, which is then used to impute missing data with the corresponding decoder. 89

6.3 **scMVI's graphical model for mESC data.** For each cell $i \in I$, scMVI learns a forward and reverse mapping between molecular layers ($\mathbf{Y}_{ri}$ is Gene expression; $\mathbf{Y}_{mi}$ is DNAm; $\mathbf{Y}_{ai}$ is Chromatin accessibility) and their associated latent parameters ($\mathbf{Z}_{ri}, \mathbf{Z}_{mi}, \mathbf{Z}_{ai}$). The flexible design assuming one VAE per molecular layer allows for accurate imputation and integration even in cases with partially observed data. Depending the sourcing protocol VAEs could be added or removed, from scMVI, to accommodate imputation and integration. 90

6.4 **Summary of imputed expression for data generated with commercial 10X and scNMT (Clark et al., 2018). Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots of imputed gene expression for one cell of the test set used when analysing the mESC_Pj (6.4a), PBMC (6.4b) and mESC_Ag (6.4c) datasets. For the multi-omics datasets (PBMC and mESC_Ag), gene expression is imputed using chromatin accessibility. For mESC_Pj we present the reconstruction of gene expression data since there are no other modalities to impute from. Each dot represents a gene and is color-coded using frequency of point occurrence. The red line is $y = x$ corresponds to the perfect match. Error bars are omitted for clarity of plots. 99

6.5 **Violin plots summarizing distribution of imputed and raw profile means for gene expression, chromatin accessibility and DNA methylation in mESC_Ag. Useful comparison between statistics of imputed and ground truth values.** (6.5a) Violin plots summarizing distribution of imputed and raw profile means for gene expression in mESC_Ag test cell data. (6.5b) Violin plots summarizing distribution of imputed and raw profile means for chromatin accessibility in mESC_Ag test cell data. (6.5c) Violin plots summarizing distribution of imputed and raw profile means for DNA methylation in mESC_Ag test cell data. 100

6.6 **Violin plots summarizing distribution of imputed and raw profile standard deviations for gene expression, chromatin accessibility and DNA methylation in mESC_Ag. Useful comparison between statistics of imputed and ground truth values.** (6.6a) Violin plots summarizing distribution of imputed and raw profile standard deviations for gene expression in mESC_Ag test cell data. (6.6b) Violin plots summarizing distribution of imputed and raw profile standard deviations for chromatin accessibility in mESC_Ag test cell data. (6.6c) Violin plots summarizing distribution of imputed and raw profile standard deviations for DNA methylation in mESC_Ag test cell data. 101

6.7 **Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression, chromatin accessibility and DNA methylation in mESC_Ag. Useful comparison between statistics of imputed and ground truth values.** (6.7a) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression in mESC_Ag test cell data. (6.7b) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for chromatin accessibility in mESC_Ag test cell data. (6.7c) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for DNA methylation in mESC_Ag test cell data. 102

6.8 **UMAP visualization of artificially unpaired data projected to the latent space for PBMC and mESC_Ag. Visualization qualitatively demonstrating that similar manifolds can be generated from partial representations.** UMAP plots summarizing latent space projections of artificially unpaired readings from jointly profiled PBMC (6.8a) and mESC_Ag (6.8b). Each dot represents a cell and is color-coded using its source modality. 103

6.9 **Violin plots summarizing latent representation euclidean distances in training set and artificially unpaired modalities in test sets cells, broken by technology and selected cell types. Figures used to asses how close in the latent manifold would previously unseen data would sit.** Violin plots summarizing latent representation euclidean distance in training set cells and artificially unpaired modalities in test set cells for droplet (6.9a) and plate based data (6.9b,6.9c,6.9d). In mESC_Ag the selected cells are Epiblast (Epi), Mesoderm (Mes), Endoderm (End) and Primitive Streak (PS). Similarly for PBMC the selected cells were CD14 Mono (CD14), CD8 T cells (CD8), Memory B cells (MemB) and Natural Killer cells (NK). Violin plots of euclidean distances on the training set are distinguished from the ones of the test set using labels "_tr" and "_tst" respectively. For example violin plot of euclidean distances between epiblast cells on the training set is denoted by Epi_tr. 104

6.10 **Summary of scMVI's latent space on mESC_Ag and PBMC. Figures demonstrating that scMVI learns latent representations correlated with cell phenotype or developmental state.** 2D UMAP projection of scMVI's, trained on mESC_Ag and PBMC with each dot representing a cell. Since scMVI estimates one latent representation per molecular layer these plot are generated by averaging across them. (6.10a) scMVI latent space on PBMC cells color-coded by cell type. (6.10b) scMVI latent space on mESC_Ag data color-coded by developmental stage. (6.10c) scMVI latent space on mESC_Ag color-coded by lineage. 105

6.11 **Summary of scMVI's latent space on mESC_Pj color-coded by lineage and developmental stage. Figures demonstrating that scMVI learns clear latent representations correlated with cell phenotype or developmental state.** 2D UMAP projection of scMVI's latent space, trained on mESC_Pj. Each dot represents a cell and is color-coded by lineage (6.11a) and developmental stage (6.11b). 108

6.12 **Median DDLFC and LFC comparison for mESC_Pj. Scatter plots used to asses how robust is DDLFC compared to LFC under various scenarios for specific developmental steps.** Scatter plots summarizing average and range of median DDLFC against average and range of median LFC across simulated scenarios in label pairs where it is less (Embryonic days $E7.0$ vs $E7.5$)(6.12a,6.12c) and most (Embryonic days $E8.0$ vs $E8.5$)(6.12b,6.12d) impacted from subsampling and label imbalance, for mESC_Pj. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range. 111

7.1 **Schematic summary of SCRaPL with more complex latent mean expression prior. Future work.** Schematic summary of SCRaPL with more complex latent mean expression prior. Dashed circle replaces standard plate notation implying that nodes on the circumference are independent unless explicitly mentioned. For a subset of genomic regions with correlated latent means and indices $\{e,u,k,t\}$, dependencies are marked with solid lines. . . 119

A.1 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 1. Figures used to demonstrate model behaviour under synthetic data generated from the model.**. Differences of posterior estimates from data generating parameters as functions of the number of cells across features for: (A.1a) methylation mean; (A.1b) expression mean; (A.1c) methylation standard deviation; (A.1d) expression standard deviation; (A.1e) correlation between expression and methylation. 122

A.2 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 2. Figures used to demonstrate model behaviour under synthetic data generated from the model.**. Differences of posterior estimates from data generating parameters as functions of average expression inflation across features for: (A.2a) methylation mean; (A.2b) expression mean; (A.2c) methylation standard deviation; (A.2d) expression standard deviation; (A.2e) correlation between expression and methylation. 123

A.3 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 3. Figures used to demonstrate model behaviour under synthetic data generated from the model.**. Differences of posterior estimates from data generating parameters as functions of average methylation coverage across features for: (A.3a) methylation mean; (A.3b) expression mean; (A.3c) methylation standard deviation; (A.3d) expression standard deviation; (A.3e) correlation between expression and methylation. 124

A.4 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 4. Figures used to demonstrate model behaviour under synthetic data generated from the model with different prior.**. Differences of posterior estimates from data generating parameters as functions of the number of cells across features for: (A.4a) methylation mean; (A.4b) expression mean; (A.4c) methylation standard deviation; (A.4d) expression standard deviation; (A.4e) correlation between expression and methylation. 125

A.5 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 5. Figures used to demonstrate model behaviour under synthetic data generated from the model with different prior.**. Differences of posterior estimates from data generating parameters as functions of average expression inflation across features for: (A.5a) methylation mean; (A.5b) expression mean; (A.5c) methylation standard deviation; (A.5d) expression standard deviation; (A.5e) correlation between expression and methylation. 126

A.6 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 6. Figures used to demonstrate model behaviour under synthetic data generated from the model with different prior.**. Differences of posterior estimates from data generating parameters as functions of average methylation coverage across features for: (A.6a) methylation mean; (A.6b) expression mean; (A.6c) methylation standard deviation; (A.6d) expression standard deviation; (A.6e) correlation between expression and methylation. 127

A.7 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 7. Figures used to demonstrate model behaviour under synthetic data generated under model mis-specification.** . Differences of posterior estimates from data generating parameters as functions of the number of cells across features for: (A.7a) methylation mean; (A.7b) methylation standard deviation; (A.7c) correlation between expression and methylation. 128

A.8 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 8. Figures used to demonstrate model behaviour under synthetic data generated under model mis-specification.** . Differences of posterior estimates from data generating parameters as functions of average expression inflation across features for: (A.8a) methylation mean; (A.8b) methylation standard deviation; (A.8c) correlation between expression and methylation. 129

A.9 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 9. Figures used to demonstrate model behaviour under synthetic data generated under model mis-specification.** . Differences of posterior estimates from data generating parameters as functions of average methylation coverage across features for: (A.6a) methylation mean; (A.9b) methylation standard deviation; (A.6e) correlation between expression and methylation. 130

A.10 **Bayesian and frequentist versions of volcano plots for the mESC and mEBC data. Plots demonstrating that SCRaPL can significantly improve detection rates, by choosing features picked up by Pearson and Spearman correlation while avoiding spurious correlations.** Each dot represents a feature and is marked with different color depending the method that labels it as a significant association. (A.10a) Bayesian volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mESC data. (A.10b) Bayesian volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mEBC data. (A.10c) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mESC data. (A.10d) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mEBC data. (A.10e) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Pearson correlation in mESC data. (A.10f) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Pearson correlation in mEBC data. 132

A.11 **Example where both SCRaPL and Pearson identify feature's association as significant in mESC and mEBC data. Examples used to identify common patterns between SCRaPL and Pearson correlation.** In all Figures the left part is a scatter plot of feature's raw data and the right part is the posterior correlation as inferred by SCRaPL. In features from mESC data (i.e. (A.11a), (A.11c) and (A.11e)) each dot of the scatter plot represents a cell reading, color-coded by CpG coverage and the expression axis is in $log(1+x)$ scale. In features from mEBC data (i.e. (A.11b), (A.11d) and (A.11f)) each dot of the scatter plot represents a cell reading, color-coded by space occupation and the accessibility/expression axis are in $log(1+x)$ scale. 134

A.12 **Example where only Pearson identifies the feature's association as significant in mESC and mEBC data. Examples used to understand why SCRaPL avoids Pearson correlation limitations.** In all Figures the left part is a scatter plot of feature's raw data and the right part is the posterior correlation as inferred by SCRaPL. In features from Argelaguet et al. (2019) data (i.e. (A.12a), (A.12c) and (A.12e)) each dot of the scatter plot represents a cell reading, color-coded by CpG coverage and the expression axis is in $log(1+x)$ scale. In features from mEBC data (i.e. (A.12b), (A.12d) and (A.12f)) each dot of the scatter plot represents a cell reading, color-coded by space occupation and the accessibility/expression axis are in $log(1+x)$ scale. 136
A.13 **Example where only SCRaPL identifies the feature's association as significant in mESC and mEBC data. Examples used to understand why SCRaPL avoids Pearson correlation limitations.** In all Figures the left part is a scatter plot of feature's raw data and the right part is the posterior correlation as inferred by SCRaPL. In features from Argelaguet et al. (2019) data (i.e. (A.13a), (A.13c) and (A.13e)) each dot of the scatter plot represents a cell reading, color-coded by CpG coverage and the expression axis is in $log(1+x)$ scale. In features from mEBC data (i.e. (A.13b), (A.13d) and (A.13f)) each dot of the scatter plot represents a cell reading, color-coded by space occupation and the accessibility/expression axis are in $log(1+x)$ scale. 138
A.14 **Raw data and posterior gene correlation for selected members of the Dppa gene family, inferred with SCRaPL. Examples used to demonstrate that SCRaPL identifies genes with known role in early development.** . . 139
A.15 **Raw data and posterior gene correlation for selected members of the Dnmt gene family, inferred with SCRaPL. Example used to demonstrate that SCRaPL identifies genes with known role in early development.** . . 140
A.16 **Raw data and posterior gene correlation for other genes in Argelaguet et al. (2019) that have been partially or not studied, inferred with SCRaPL. Examples where SCRaPL has identified genes of potential interest.** 140
A.17 **GO analysis with features detected by SCRaPL. Figures used to demonstrate early developmental stage biological processes identified by SCRaPL.** 141
A.18 **GO analysis with features detected by Spearman correlation. Figures used to demonstrate early developmental stage biological processes identified by Spearman correlation.** 142

A.19 **Execution times as a function (features/cell numbers). Graph used to asses how SCRaPL scales with dimension. For large datasets it scales almost linearly.** 147
A.20 **SCRaPL's behavior compared to Pearson/Spearman correlation in macro scale across. Scatter plots used to asses how Spearman/Pearson predictions vary from SCRaPL's with zero inflation and CpG coverage.**. (A.20a) Scatterplots demonstrating the positive relationship between alternative correlation estimates (Pearson,Spearman and SCRaPL) and % zeros in accessibility and expression across features. (A.20b) Scatterplots demonstrating the positive relationship between alternative correlation estimates (Spearman and SCRaPL) and % zeros in accessibility and expression across features. (A.20c) Scatter plots demonstrating the negative/positive relationship between alternative correlation estimates and CpG coverage/% zeros in expression across features. 148
A.21 **Cell label transfer from expression to accessibility data for raw data A.21a. (A.21b) Visualization of sc-RNA and scATAC data on the same plot for raw data. Serves as baseline to compare raw data integration with Seurat.** 150
A.22 **Cell label transfer from expression to accessibility data for SCRaPL preprocessed data A.22a. (A.22b)Visualization of sc-RNA and scATAC data on the same plot for SCRaPL preprocessed data. Used to compare SCRaPL preprocessed data and raw data when integration is done with Seurat.** 151
A.23 **Stability analysis of cell integration using SCRaPL as a denoising tool. Used to asses stability of Seurat's integration for SCRaPL preprocessed data.** Panel summarizing label transfer from expression to accessibility in trial 2(A.23a) along with respective latent plot visualization (A.23b) of sc-RNA and scATAC. 152
A.24 **Stability analysis of cell integration using SCRaPL as a denoising tool. Used to asses stability of Seurat's integration for SCRaPL preprocessed data.** Panel summarizing label transfer from expression to accessibility in trial 3 (A.24a) along with respective latent plot visualization (A.24b) of sc-RNA and scATAC. 153

A.25 **Stability analysis of cell integration using SCRaPL as a denoising tool. Used to asses stability of Seurat's integration for SCRaPL preprocessed data.** Panel summarizing label transfer from expression to accessibility in trial 4 (A.25a) along with respective latent plot visualization (A.25a) of sc-RNA and scATAC. 154
A.26 **Convergence analysis for feature 61.** Convergence analysis of SCRaPL latent parameters for feature 61 when SCRaPL is applied with inflation on mESC data. (A.26a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.26b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 155
A.27 **Convergence analysis for feature 189.** Convergence analysis of SCRaPL latent parameters for feature 189 when SCRaPL is applied with inflation on mESC data. (A.27a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.27b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 156
A.28 **Convergence analysis for feature 1053.** Convergence analysis of SCRaPL latent parameters for feature 1053 when SCRaPL is applied with inflation on mESC data. (A.28a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.28b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 156
A.29 **Convergence analysis for feature 1211.** Convergence analysis of SCRaPL latent parameters for feature 1211 when SCRaPL is applied with inflation on mESC data. (A.29a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.29b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 157
A.30 **Convergence analysis for feature 1950.** Convergence analysis of SCRaPL latent parameters for feature 1950 when SCRaPL is applied with inflation on mESC data. (A.30a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.30b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 157
A.31 **Convergence analysis for feature 2724.** Convergence analysis of SCRaPL latent parameters for feature 2724 when SCRaPL is applied with inflation on mESC data. (A.31a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.31b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 158

A.32 **Convergence analysis for feature 4139.** Convergence analysis of SCRaPL latent parameters for feature 4139 when SCRaPL is applied with inflation on mESC data. (A.32a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.32b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 158
A.33 **Convergence analysis for feature 4795.** Convergence analysis of SCRaPL latent parameters for feature 4795 when SCRaPL is applied with inflation on mESC data. (A.33a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.33b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 159
A.34 **Convergence analysis for feature 7591.** Convergence analysis of SCRaPL latent parameters for feature 7591 when SCRaPL is applied with inflation on mESC data. (A.34a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.34b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 159
A.35 **Convergence analysis for feature 61.** Convergence analysis of SCRaPL latent parameters for feature 61 when SCRaPL is applied without inflation on mESC data. (A.35a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.35b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 160
A.36 **Convergence analysis for feature 189.** Convergence analysis of SCRaPL latent parameters for feature 189 when SCRaPL is applied without inflation on mESC data. (A.36a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.36b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 160
A.37 **Convergence analysis for feature 1053.** Convergence analysis of SCRaPL latent parameters for feature 1053 when SCRaPL is applied without inflation on mESC data. (A.37a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.37b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 161
A.38 **Convergence analysis for feature 1211.** Convergence analysis of SCRaPL latent parameters for feature 1211 when SCRaPL is applied without inflation on mESC data. (A.38a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.38b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 161

A.39 **Convergence analysis for feature 1950** Convergence analysis of SCRaPL latent parameters for feature 1950 when SCRaPL is applied without inflation on mESC data. (A.39a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.39b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 162
A.40 **Convergence analysis for feature 2724.** Convergence analysis of SCRaPL latent parameters for feature 2724 when SCRaPL is applied without inflation on mESC data. (A.40a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.40b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 162
A.41 **Convergence analysis for feature 4139.** Convergence analysis of SCRaPL latent parameters for feature 4139 when SCRaPL is applied without inflation on mESC data. (A.41a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.41b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 163
A.42 **Convergence analysis for feature 4795.** Convergence analysis of SCRaPL latent parameters for feature 4795 when SCRaPL is applied without inflation on mESC data. (A.42a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.42b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 163
A.43 **Convergence analysis for feature 7591.** Convergence analysis of SCRaPL latent parameters for feature 7591 when SCRaPL is applied without inflation on mESC data. (A.43a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.43b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 164
A.44 **Convergence analysis for feature 1.** Convergence analysis of SCRaPL latent parameters for feature 1 when SCRaPL is applied with inflation on mEBC data. (A.44a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.44b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 164
A.45 **Convergence analysis for feature 48.** Convergence analysis of SCRaPL latent parameters for feature 48 when SCRaPL is applied with inflation on mEBC data. (A.45a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.45b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 165

A.46 **Convergence analysis for feature 371.** Convergence analysis of SCRaPL latent parameters for feature 371 when SCRaPL is applied with inflation on mEBC data. (A.46a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.46b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 165
A.47 **Convergence analysis for feature 477.** Convergence analysis of SCRaPL latent parameters for feature 477 when SCRaPL is applied with inflation on mEBC data. (A.47a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.47b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 166
A.48 **Convergence analysis for feature 731.** Convergence analysis of SCRaPL latent parameters for feature 731 when SCRaPL is applied with inflation on mEBC data. (A.48a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.48b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 166
A.49 **Convergence analysis for feature 1168.** Convergence analysis of SCRaPL latent parameters for feature 1168 when SCRaPL is applied with inflation on mEBC data. (A.49a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.49b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 167
A.50 **Convergence analysis for feature 1444.** Convergence analysis of SCRaPL latent parameters for feature 1444 when SCRaPL is applied with inflation on mEBC data. (A.50a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.50b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 167
A.51 **Convergence analysis for feature 3166.** Convergence analysis of SCRaPL latent parameters for feature 3166 when SCRaPL is applied with inflation on mEBC data. (A.51a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.51b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 168
A.52 **Convergence analysis for feature 4111.** Convergence analysis of SCRaPL latent parameters for feature 4111 when SCRaPL is applied with inflation on mEBC data. (A.52a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.52b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 168

A.53 **Convergence analysis for feature 1.** Convergence analysis of SCRaPL latent parameters for feature 1 when SCRaPL is applied without inflation on mEBC data. (A.53a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.53b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 169
A.54 **Convergence analysis for feature 48.** Convergence analysis of SCRaPL latent parameters for feature 48 when SCRaPL is applied without inflation on mEBC data. (A.54a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.54b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 169
A.55 **Convergence analysis for feature 371.** Convergence analysis of SCRaPL latent parameters for feature 371 when SCRaPL is applied without inflation on mEBC data. (A.55a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.55b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 170
A.56 **Convergence analysis for feature 477.** Convergence analysis of SCRaPL latent parameters for feature 477 when SCRaPL is applied without inflation on mEBC data. (A.56a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.56b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 170
A.57 **Convergence analysis for feature 731.** Convergence analysis of SCRaPL latent parameters for feature 731 when SCRaPL is applied without inflation on mEBC data. (A.57a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.57b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 171
A.58 **Convergence analysis for feature 1168.** Convergence analysis of SCRaPL latent parameters for feature 1168 when SCRaPL is applied without inflation on mEBC data. (A.58a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.58b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 171
A.59 **Convergence analysis for feature 1444.** Convergence analysis of SCRaPL latent parameters for feature 1444 when SCRaPL is applied without inflation on mEBC data. (A.59a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.59b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 172

A.60 **Convergence analysis for feature 3166.** Convergence analysis of SCRaPL latent parameters for feature 3166 when SCRaPL is applied without inflation on mEBC data. (A.60a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.60b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 172
A.61 **Convergence analysis for feature 4111.** Convergence analysis of SCRaPL latent parameters for feature 4111 when SCRaPL is applied without inflation on mEBC data. (A.61a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.61b) PSRF. distributions for epigenetic and transcriptomic latent parameters. 173
A.62 **Posterior predictive check for feature** 61 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 61 in mESC. Each dot represents a cell. 174
A.63 **Posterior predictive check for feature** 189 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 189 in mESC. Each dot represents a cell. 175
A.64 **Posterior predictive check for feature** 1053 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1053 in mESC. Each dot represents a cell. 176
A.65 **Posterior predictive check for feature** 1211 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1211 in mESC. Each dot represents a cell. 177
A.66 **Posterior predictive check for feature** 1950 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1950 in mESC. Each dot represents a cell. 178
A.67 **Posterior predictive check for feature** 2724 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 2724 in mESC. Each dot represents a cell. 179
A.68 **Posterior predictive check for feature** 4795 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 4795 in mESC. Each dot represents a cell. 180
A.69 **Posterior predictive check for feature** 7591 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 7591 in mESC. Each dot represents a cell. 181

A.70 **Posterior predictive check for feature** 1 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1 in mESC. Each dot represents a cell. 182
A.71 **Posterior predictive check for feature** 48 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 48 in mESC. Each dot represents a cell. 183
A.72 **Posterior predictive check for feature** 477 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 477 in mESC. Each dot represents a cell. 184
A.73 **Posterior predictive check for feature** 731 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 731 in mESC. Each dot represents a cell. 185
A.74 **Posterior predictive check for feature** 1168 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1168 in mESC. Each dot represents a cell. 186
A.75 **Posterior predictive check for feature** 1444 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1444 in mESC. Each dot represents a cell. 187
A.76 **Posterior predictive check for feature** 3166 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 3166 in mESC. Each dot represents a cell. 188
A.77 **Posterior predictive check for feature** 4111 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 4111 in mESC. Each dot represents a cell. 189
A.78 Graphical representations of model wihtout over-dispersion fitted to each genomic region $j \in J$ with mESC and mEBC data. Observed genomic data (Y_{ij1} is RNA expression; Y_{ij2} is Chromatin Accessibility), with their associated latent variables (Z_{ij1}, Z_{ij2}) and feature-specific model parameters (μ_j, Σ_j) capturing feature specific epigenomic and transcriptomic marks. . . 190
A.79 **Pearson correlation and Normal likelihood model correlation comparison.**(A.79a) Pearson correlation and Normal likelihood model correlation comparison in mESC. (A.79b) Pearson correlation and Normal likelihood model correlation comparison in mEBC. 192

A.80 **Posterior correlation/covariance for feature** 61 **with different correlation, standard deviation and mean priors in mESC data.**(A.80a) Posterior correlation distribution for feature 61 and different correlation. (A.80b) Posterior correlation distribution for feature 61 and different correlation. (A.80c) Posterior covariance distribution for feature 61 and different mean priors in mESC data. 193
A.81 **Posterior correlation/covariance for feature** 189 **with different correlation, standard deviation and mean priors in mESC data.**(A.81a) Posterior correlation distribution for feature 189 and different correlation. (A.81b) Posterior correlation distribution for feature 189 and different correlation. (A.81c) Posterior covariance distribution for feature 189 and different mean priors in mESC data. 194
A.82 **Posterior correlation/covariance for feature** 1053 **with different correlation, standard deviation and mean priors in mESC data.**(A.82a) Posterior correlation distribution for feature 1053 and different correlation. (A.82b) Posterior correlation distribution for feature 1053 and different correlation. (A.82c) Posterior covariance distribution for feature 1053 and different mean priors in mESC data. 195
A.83 **Posterior correlation/covariance for feature** 1211 **with different correlation, standard deviation and mean priors in mESC data.**(A.83a) Posterior correlation distribution for feature 1211 and different correlation. (A.83b) Posterior correlation distribution for feature 1211 and different correlation. (A.83c) Posterior covariance distribution for feature 1211 and different mean priors in mESC data. 196
A.84 **Posterior correlation/covariance for feature** 1950 **with different correlation, standard deviation and mean priors in mESC data.**(A.84a) Posterior correlation distribution for feature 1950 and different correlation. (A.84b) Posterior correlation distribution for feature 1950 and different correlation. (A.84c) Posterior covariance distribution for feature 1950 and different mean priors in mESC data. 197
A.85 **Posterior correlation/covariance for feature** 2724 **with different correlation, standard deviation and mean priors in mESC data.**(A.85a) Posterior correlation distribution for feature 2724 and different correlation. (A.85b) Posterior correlation distribution for feature 2724 and different correlation. (A.85c) Posterior covariance distribution for feature 2724 and different mean priors in mESC data. 198

A.86 **Posterior correlation/covariance for feature** 4139 **with different correlation, standard deviation and mean priors in mESC data.**(A.86a) Posterior correlation distribution for feature 4139 and different correlation. (A.86b) Posterior correlation distribution for feature 4139 and different correlation. (A.86c) Posterior covariance distribution for feature 4139 and different mean priors in mESC data. 199
A.87 **Posterior correlation/covariance for feature** 4795 **with different correlation, standard deviation and mean priors in mESC data.**(A.87a) Posterior correlation distribution for feature 4795 and different correlation. (A.87b) Posterior correlation distribution for feature 4795 and different correlation. (A.87c) Posterior covariance distribution for feature 4795 and different mean priors in mESC data. 200
A.88 **Posterior correlation/covariance for feature** 7591 **with different correlation, standard deviation and mean priors in mESC data.**(A.88a) Posterior correlation distribution for feature 7591 and different correlation. (A.88b) Posterior correlation distribution for feature 7591 and different correlation. (A.88c) Posterior covariance distribution for feature 7591 and different mean priors in mESC data. 201
A.89 **Posterior correlation/covariance for feature** 1 **with different correlation, standard deviation and mean priors in mEBC data.**(A.89a) Posterior correlation distribution for feature 1 and different correlation. (A.89b) Posterior correlation distribution for feature 1 and different correlation. (A.89c) Posterior covariance distribution for feature 1 and different mean priors in mEBC data. 202
A.90 **Posterior correlation/covariance for feature** 48 **with different correlation, standard deviation and mean priors in mEBC data.**(A.90a) Posterior correlation distribution for feature 48 and different correlation. (A.90b) Posterior correlation distribution for feature 48 and different correlation. (A.90c) Posterior covariance distribution for feature 48 and different mean priors in mEBC data. 203
A.91 **Posterior correlation/covariance for feature** 477 **with different correlation, standard deviation and mean priors in mEBC data.**(A.91a) Posterior correlation distribution for feature 477 and different correlation. (A.91b) Posterior correlation distribution for feature 477 and different correlation. (A.91c) Posterior covariance distribution for feature 477 and different mean priors in mEBC data. 204

A.92 **Posterior correlation/covariance for feature** 731 **with different correlation, standard deviation and mean priors in mEBC data.**(A.92a) Posterior correlation distribution for feature 731 and different correlation. (A.92b) Posterior correlation distribution for feature 731 and different correlation. (A.92c) Posterior covariance distribution for feature 731 and different mean priors in mEBC data. 205

A.93 **Posterior correlation/covariance for feature** 1168 **with different correlation, standard deviation and mean priors in mEBC data.**(A.93a) Posterior correlation distribution for feature 1168 and different correlation. (A.93b) Posterior correlation distribution for feature 1168 and different correlation. (A.93c) Posterior covariance distribution for feature 1168 and different mean priors in mEBC data. 206

A.94 **Posterior correlation/covariance for feature** 1444 **with different correlation, standard deviation and mean priors in mEBC data.**(A.94a) Posterior correlation distribution for feature 1444 and different correlation. (A.94b) Posterior correlation distribution for feature 1444 and different correlation. (A.94c) Posterior covariance distribution for feature 1444 and different mean priors in mEBC data. 207

A.95 **Posterior correlation/covariance for feature** 3166 **with different correlation, standard deviation and mean priors in mEBC data.**(A.95a) Posterior correlation distribution for feature 3166 and different correlation. (A.95b) Posterior correlation distribution for feature 3166 and different correlation. (A.95c) Posterior covariance distribution for feature 3166 and different mean priors in mEBC data. 208

A.96 **Posterior correlation/covariance for feature** 4111 **with different correlation, standard deviation and mean priors in mEBC data.**(A.96a) Posterior correlation distribution for feature 4111 and different correlation. (A.96b) Posterior correlation distribution for feature 4111 and different correlation. (A.96c) Posterior covariance distribution for feature 4111 and different mean priors in mEBC data. 209

B.1 **Imputation statistics for droplet and plate based protocols. Figures used to asses scMVI's imputation quality. Useful comparison between statistics of imputed and ground truth values.** Scatter plots summarizing average euclidean distance between imputed and observed value as a function of 95% confidence interval length for droplet (i.e. mESC_Pj and PBMC) (B.1a) and plate based (i.e. mESC_Ag)(B.1b) protocols. 211

B.2 **Violin plots summarizing distributions of imputed and raw profile means for gene expression and chromatin accessibility in mEBC. Useful comparison between statistics of imputed and ground truth values.** (B.2a) Violin plots summarizing distributions of imputed and raw profile means for gene expression in mEBC test cell data. (B.2b) Violin plots summarizing distributions of imputed and raw profile means for chromatin accessibility in mEBC test cell data. 212

B.3 **Violin plots summarizing distributions of imputed and raw profile standard deviations for gene expression and chromatin accessibility in mEBC. Useful comparison between statistics of imputed and ground truth values.** (B.3a) Violin plots summarizing distributions of imputed and raw profile standard deviations for gene expression in mEBC test cell data. (B.3b) Violin plots summarizing distributions of imputed and raw profile standard deviations for chromatin accessibility in mEBC test cell data. . . . 212

B.4 **Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression and chromatin accessibility in mEBC. Useful comparison between statistics of imputed and ground truth values.** (B.4a) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression in mEBC test cell data. (B.4b) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for chromatin accessibility in mEBC test cell data. 213

B.5 **Imputing chromatin accessibility with gene expression for PBMC data. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed chromatin accessibility from gene expression as a function of observed chromatin accessibility for cells from PBMC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match. 214

B.6 **Imputing expression with chromatin accessibility for PBMC. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed expression from chromatin accessibility as a function of observed expression for cells from PBMC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match. 215

B.7 **Imputing expression with chromatin accessibility for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed expression from chromatin accessibility as a function of observed gene expression for cells from mESC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match. 216

B.8 **Imputing expression with DNA methylation for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed expression from DNA methylation as a function of observed expression for cells from mESC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match. 217

B.9 **Imputing chromatin accessibility with gene expression for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed chromatin accessibility from gene expression as a function of observed chromatin accessibility for cells from mESC test set. Each dot represents a gene and it is color-coded by GpC coverage. The red line is $y = x$ and corresponds to perfect match. 218

B.10 **Imputing chromatin accessibility with DNA methylation for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed chromatin accessibility from DNA methylation as a function of observed chromatin accessibility for cells from mESC test set. Each dot represents a gene and it is color-coded by GpC coverage. The red line is $y = x$ and corresponds to perfect match. 219

B.11 **Imputing DNA methylation with gene expression for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed DNA methylation from gene expression as a function of observed DNA methylation for cells from mESC test set. Each dot represents a gene and it is color-coded by CpG coverage. The red line is $y = x$ and corresponds to perfect match. 220

B.12 **Imputing DNA methylation with chromatin accessibility for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed DNA methylation from chromatin accessibility as a function of observed DNA methylation for cells from mESC test set. Each dot represents a gene and it is color-coded by CpG coverage. The red line is $y = x$ and corresponds to perfect match. 221

B.13 **Summary of scVI's latent space on mESC_Pj color-coded by lineage and developmental stage. Figures demonstrating that scMVI learns latent representations correlated with cell phenotype or developmental state.** 2D UMAP projection of scVI's latent space, trained on mESC_Pj. Each dot represents a cell and is color-coded by lineage (B.13a) and developmental stage (B.13b). 223

B.14 **UMAP projection of scMVI and MultiVI latent manifolds on paired PBMC data. Figures used to qualitatively compare latent representations from related models to scMVI.** Two dimensional depiction of scMVI's (B.14a) and MultiVI's (B.14b) latent space trained on paired PBMC data. Each dot represents a cell and is color-coded by cell type. Since scMVI estimates one latent representation per molecular layer, Figure B.14a is generated by averaging across them before visualising with UMAP. 224

B.15 **UMAP projection of Cobolt and scMM latent manifolds on paired PBMC data. Figures used to qualitatively compare latent representations from related models to scMVI.** Two dimensional depiction of Cobolt's (B.15a) and scMM's (B.15b) latent space trained on paired PBMC data. Each dot represents a cell and is color-coded by cell type. 225

B.16 **UMAP projection of scMVAE and scVI latent manifolds on paired PBMC data. Figures used to qualitatively compare latent representations from related models to scMVI.** Two dimensional depiction of scMVAE's (B.16a) latent space trained on paired PBMC data. Each dot represents a cell and is color-coded by cell type. 226

B.17 **Heatmap summarizing proximal cell lineage in scMVI's latent space for each cell on paired mESC_Ag. Heatmap used to demonstrate which cell type would be picked in dataset integration with scMVI.**. For each latent representation in mESC_Ag, estimated with paired data, we track proximal cell lineage to understand how scMVI's latent space is structured. Since scMVI estimates one latent representation per molecular layer, the heatmap is generated after averaging across them. 227

B.18 **Heatmap summarizing proximal cell type in scMVI's and MultiVI's latent space for each cell on paired PBMC data. Heatmaps used to compare which cell type would be picked in dataset integration with alternative models.** For each PBMC latent representation we track proximal cell types to understand how scMVI's (B.14a) and MultiVI's (B.18b) latent space is structured. Since scMVI estimates one latent representation per molecular layer, Heatmap B.18a is generated after averaging across them. . 228

B.19 **Heatmap summarizing proximal cell type in Cobolt's and scMM's latent space for each cell on paired PBMC data. Heatmaps used to compare which cell type would be picked in dataset integration with alternative models.** For each PBMC latent representation we track proximal cell types to understand how Cobolt's(B.19a) and scMM's(B.19b) latent space is structured. 229

B.20 **Heatmap summarizing proximal cell type in scMVAE's latent space for each cell on paired PBMC data. Heatmap used to compare which cell type would be picked in dataset integration with alternative models.** For each PBMC latent representation we track proximal cell types to understand how scMVAE's(B.20a) latent space is structured. 230

B.21 **Median DDLFC and LFC_NN comparison for PBMC. Scatter plots used to asses how robust is DDLFC compared to LFC under various scenarios for specific cell types.** Scatter plots summarizing average and range of median DDLFC against average and range of median LFC_NN in pairs where it is less (CD14 vs CD4)(B.21a,B.21c) and most (Natural kille vs Memory B)(B.21b,B.21d) impacted from subsampling and label imbalance, for PBMC. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range. 232

B.22 **Median DDLFC and LFC_DT comparison for mESC_Pj. Scatter plots used to asses how robust is DDLFC compared to differential expression analysis under various scenarios for specific developmental steps.** Scatter plots summarizing average and range of median DDLFC against LFC_DT(and its range) in label pairs where it is less (Embryonic days $E7.0$ vs $E7.5$)(B.22a,B.22c) and most (Embryonic days $E8.0$ vs $E8.5$)(B.22b,B.22d) impacted from subsampling and label imbalance, for mESC_Pj. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range. 233

B.23 **Median DDLFC and LFC_DT comparison for PBMC. Scatter plots used to asses how robust is DDLFC compared to differential expression analysis under various scenarios for specific cell types.** Scatter plots summarizing average and range of median DDLFC against LFC_DT(and its range) in labels where it is less (CD14 vs CD4)(B.23a,B.23c) and most (Naturall Killer vs Memory B)(B.23b,B.23d) impacted from subsampling and label imbalance, for PBMC. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range. 234

B.24 **GSEA analysis on** 500 **genomic regions with the strongest negative DDLFC between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 235

B.25 **GSEA analysis on** 500 **genomic regions with the strongest positive DDLFC between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 235

B.26 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_NN between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.** 236

B.27 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_NN between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.** 236

B.28 **GSEA analysis on** 500 **genomic regions with the strongest negative DDLFC between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 237
B.29 **GSEA analysis on** 500 **genomic regions with the strongest positive DDLFC between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 237
B.30 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_NN between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.** 238
B.31 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_NN between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.** 238
B.32 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_DT between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.** 239
B.33 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_DT between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.** 239
B.34 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_DT between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.** 240
B.35 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_DT between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.** 241

B.36 **Gene set enrichment analysis for features with the strongest negative gene expression DDLFC between embryonic days $E6.5$ and $E7.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 242
B.37 **Gene set enrichment analysis for features with the strongest positive gene expression DDLFC between embryonic days $E6.5$ and $E7.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 242
B.38 **Gene set enrichment analysis for features with the strongest negative gene expression DDLFC between embryonic days $E5.5$ and $E6.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 243
B.39 **Gene set enrichment analysis for features with the strongest positive gene expression DDLFC between embryonic days $E5.5$ and $E6.5$ formESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 243
B.40 **Gene set enrichment analysis for features with the strongest positive gene expression DDLFC between embryonic days $E4.5$ and $E5.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 244
B.41 **Gene set enrichment analysis for features with the strongest negative gene expression DDLFC between embryonic days $E4.5$ and $E5.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 244
B.42 **Gene set enrichment analysis for features with the strongest positive DNA methylation DDLFC between embryonic days $E6.5$ and $E7.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 245
B.43 **Gene set enrichment analysis for features with the strongest positive chromatin accessibility DDLFC between embryonic days $E6.5$ and $E7.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.** 245

B.44 **Gene set enrichment analysis for features with the strongest negative gene expression LFC_DT between embryonic days $E6.5$ and $E7.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.** 246
B.45 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days $E6.5$ and $E7.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.** 247
B.46 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days $E5.5$ and $E6.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.** 248
B.47 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days $E5.5$ and $E6.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.** 249
B.48 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days $E4.5$ and $E5.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.** 250
B.49 **Gene set enrichment analysis for features with the strongest negative gene expression LFC_DT between embryonic days $E4.5$ and $E5.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.** 251

List of tables

3.1 Table summarizing Pearson correlation hypothesis testing for N genomic regions with $N_{0,R}$ denoting the number of regions for which H_0 is true but the test rejected H_0, $N_{0,A}$ denoting the number of regions for which H_0 is true and the test accepted H_0, $N_{a,R}$ denoting the number of regions for which H_a is true and the test rejected H_0 and $N_{a,A}$ denoting the number of regions for which H_a is true and the test accepted H_0. 30

4.1 **Computational single-cell multi-omics tools applying intermediate integration approaches and their applicable omic data types (Part1).** 53

4.2 **Computational single-cell multi-omics tools applying intermediate integration approaches and their applicable omic data types (Part2).** 53

4.3 **Computational single-cell multi-omics tools applying intermediate integration approaches and their applicable omic data types (Part3).** 53

5.1 **Table summarizing the SCRaPL default hyper-parameter values.** . . . 66

5.2 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in DNAm-expression modalities of mESC data when $\gamma = 0.29$** 66

5.3 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in accessibility-expression modalities of mEBC data when $\gamma = 0.155$** 67

5.4 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in methylation-expression modalities of mESC data when $\gamma = 0.205$** 78

5.5 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in accessibility-expression modalities of mEBC data when $\gamma = 0.145$** 78

6.1 **Table summarizing data used in Chapter 6.** Table giving a synopsis of the data used for this study. Columns 2-4 describe the sequencing protocol, the number of cells and the number of features per modality for each of the datasets used. Then columns 5-7 present the available data modalities and column 8 shows the coverage using average expression across genes and cells in each dataset. 87

6.2 **Related methods using VAEs to model partially or fully observed single cell multi-omics data**. Table summarizes key differences for each of the reviewed methods. 96

6.3 Level of cell subsampling in each of the 7 experiments created using mESC_Pj and PBMC to compare DDLFC, LFC_NN and LFC_DT behaviour under label imbalance in gene expression data. Set 1 and 2 are simulated datasets comprised of cells with the same developmental or phenotype label but distinct between datasets. After training scMVI, scVI and MultiVI are conditioned on simulated sets 1 and 2 to estimate median DDLFC (scMVI) and LFC_NN (scVI, MultiVI). Then the average along with the range of medians for DDLFC and LFC_NN across the simulated scenarios are compared. . . 110

A.1 **Number of detected features for original and negative control data with SCRaPL, Pearson and Spearman correlation in accessibility-expression modalities of mESC data data when $\gamma = 0.28$** 131

A.2 **Number of detected features for original and negative control data with SCRaPL, Pearson and Spearman correlation for synthetic data generated from SCRaPL when $\gamma = 0.115$** 131

Nomenclature

Mathematical Notation

$\mathbf{P}(z \mid \mathscr{D})$ Probability of z given $\mathscr{D}$

$\mathscr{D}$ Dataset

$D_{KL}(\mathbb{Q}, \mathbb{P})$ KL Divergence from distribution $\mathbb{Q}$ to $\mathbb{P}$

$\nabla_\theta f$ Gradient of f with respect to θ

$\mathbb{I}_{N \times N}$ The N $\times$ N identity matrix

$\in$ Belongs to

$Z_1 \perp\!\!\!\perp Z_2 \mid Z_3$ Variable Z_1 is independent of Z_2 conditioned on Z_3

$Z_1 \not\perp\!\!\!\perp Z_2 \mid Z_3$ Variable Z_1 is not independent of Z_2 conditioned on Z_3

$\mathbb{I}(\cdot)$ Indicator function

$\Phi(\cdot)$ The CDF of the standard normal distribution

$\mathbf{P}(z, y)$ Probability of z and y

z,Z scalars (lower or upper case)

$\{\ldots\}$ A set, of variables, indices samples or observations

$\mathbf{X}$ Vector or Matrix (uppercase bold letters

Acronyms / Abbreviations

ATAC Assay for Transposase-Accessible Chromatin

bindSC Bi-order INtegration of multi-omics Data from Single Cell sequencing technologies

BREM-SC Bayesian Random Effects Mixture Single Cell

BS-seq Bisulfite Sequencing

CCA Canonical Correlation Analysis

CDF Cumulative distribution function

CI Conditional Independence

scCITE-seq single cell cellular indexing of transcriptomes and epitopes sequencing

CPT Conditional Probability Tables

DAG Directed acyclic graph

DD Differential Direction

DDLFC Differential Direction Log Fold Change

DGM Directed graphical model

DIC Deviance Information Criterion

DNA Deoxyribonucleic Acid

DNMTs DNA Methyltransferases

Drop-Chip Droplet based Chromatin Immunoprecipitation

e.g. For example

EFDR Expected False Discovery Rate

ELBO Evidence Lower Bound

ENCODE Encyclopedia of DNA Elements

ESC Embryonic Stem Cells

FACS Fluorescence-activated cell sorting

FDR False Discovery Rate

GR Gelman Rubin

GSEA Gene Set Enrichment Analysis

HATs Histone Acetyltransferases

HDACs Histone Deacetylases

HMC Hamiltonian Monte Carlo

HTS High Throughput Sequencing

i.e. This is to say

IVT In Vitro Transcription

Jeff Jeffreys

KL Kullback-Leibler

LFC Log Fold Change

LIGER Linked Inference of Genomic Experimental Relationship

M.CviPI Methyltrasferase

MAGAN Manifold-Aligning Generative Adversarial Network

MATCHER Manifold Alignment to CHaracterize Experimental Relationships

MBD Methyl Binding Domain

MB Markov Blanket

MC Monte Carlo

MCMC Markov Chain Monte Carlo

mEBC mouse Embryonic Brain Cells

mESC mouse Embryonic Stem Cells

MMD-MA Maximum Mean Discrepancy-based Manifold Alignment

MoE Mixture of Experts

MOFA Multi-Omics Factor Analysis

mRNA messenger Ribonucleic Acid

NGS Next Generation Sequencing

NMF Non-negative matrix factorization

NUTS No-U-Turn Sampler

PBMC Peripheral Blood Mononuclear cell

PC-MSC Pair-wised Co-regularized Multimodal Spectral Clustering

PCR Polymerase Chain Reaction

PDF Probability Density Function

PoE Product of Experts

PSRF Potential Scale Reduction Factor

QC Quality Control

RNA Ribonucleic Acid

RRBS Reduced Representation Bisulfite Sequencing

rRNA ribosomal Ribonucleic Acid

SAME-clustering Single-cell RNA-seq Aggregated clustering via Mixture model Ensemble

scATAC Single Cell Assay for Transposase-Accessible Chromatin

scBS-seq Single Cell Bisulfite Sequencing

Sc-GPE Graph Partitioning Based Cluster Ensemble Method

scM&T-seq Single cell Methylome and Transcriptome sequencing

scMM single-cell Mixture-of-experts Multimodal variational autoencoder

scMoC Single-Cell Multi-omics Clustering

scMT-seq Single Cell Methylome and Transcriptome Sequencing

scMVAE single-cell Multimodal Variational Autoencoder

scMVI single cell Multi View Inference

scNMT-seq Single Cell Nucleosome, Methylation and Transcription Sequencing

scNOMe-seq Single Cell Nucleosome Occupancy and Methylation Sequencing

SCRaPL Single Cell Regulatory Pattern Learning

scRNA-seq Single Cell RNA Sequencing

scRRBS Single Cell Reduced Representation Bisulfite Sequencing

scWGBS Single Cell Whole Genome Bisulfite Sequencing

SNARE-seq Single-Nucleus chromatin Accessibility and mRNA Expression sequencing

tRNA transfer Ribonucleic Acid

TSS Transcription Start Site

UMIs Unique Molecular Identifiers

VAE Variational Autoencoder

VI Variational Inference

w.r.t. With respect to

WAIC Watanabe Information Criterion

WGBS Whole Genome Bisulfite Sequencing

WNN Weighted Nearest Neighbours

ZINB Zero inflated Negative Binomial

ZIP Zero inflated Poisson

ZI Zero inflatio

Chapter 1

Introduction

According to the Roadmap Epigenomics Project the epigenome consists of heritable changes in gene activity and expression along with stable long-term alternation in the transcription that are not necessarily heritable (Kundaje et al., 2015). The release from genetic determinism has fueled a drive to produce data that illustrate complex interactions between different molecular layers (e.g. DNA sequence and epigenome), with the ultimate goal of understanding how the epigenome gives rise to different phenotypes, under what circumstances diseases can appear and if modern medication targeting epigenetic mechanisms can be developed to cure life threatening diseases such as cancer.

Early efforts to generate the first sequence of the human genome like the Human Genome Project (Lander et al., 2001) were based on Sanger sequencing (Sanger and Coulson, 1975; Sanger et al., 1977). Despite being extremely laborious, Sanger sequencing allowed for the creation of an accurate reference genome. The advent of new sequencing protocols, the scaling and parallelization attempts as part of Encyclopedia of DNA elements (ENCODE)(Dunham et al., 2012) along with the accurate reference genome map produced large scale datasets. This has enabled researchers to explore the role of functional genomics in genetic variations and disease. Commercialization and lack of similar to ENCODE epigenomic studies has lead to the Roadmap Epigenomics Consortium (Kundaje et al., 2015), perhaps the most systematic study of the human epigenome for primary cells and tissue. An integrated analysis of 111 reference human epigenomes profiling DNA methylation, DNA accessibility and RNA expression have demonstrated the central role of epigenomic information for understanding gene regulation, cellular differentiation and human disease (Kundaje et al., 2015).

Historically, large scale profiling of genome and products like RNA transcripts, chromatin structure and metabolites (Lander, 1996) could only be applied to bulk tissue samples comprised of an ensemble of cells, providing an average representation which masked cell

heterogeneity (Regev et al., 2017). The introduction of massively parallel assays jointly quantifying thousands of cells' omics profiles at single cell resolution (i.e. epigenomics and transcriptomics (Argelaguet et al., 2019; Cao et al., 2018)), have among others allowed for a more thorough examination of the regulatory role the epigenetic landscape has on gene expression and lead to the creation of the most comprehensive reference maps of all human cells (Regev et al., 2017).

The emergence of novel single cell multi-omics protocols to study epigentic regulation of transcriptome has also brought up a series of computational problems. From integrating, clustering and visualizing to addressing high sparsity or multiple sources of technical noise. There are various factors confounding multi-omics data, leading to under-powered statistical analyses. A similar situation arose in the early days of transcriptomic studies when scientists were correlating the transcriptome and the proteome trying to detect post-transcriptional regulation. Using Pearson correlation they found that transcriptional regulation accounts for at most 50% of the variation (Brockmann et al., 2007; Csárdi et al., 2015; Maier et al., 2009; Schmidt et al., 2007; Schwanhäusser et al., 2011; Vogel and Marcotte, 2012) with strong correlation present only on a subset of genes (Csárdi et al., 2015; Schwanhäusser et al., 2011; Vogel and Marcotte, 2012). A systematic revisit of such studies by Csárdi et al. (2015) using noise-aware tools, revised the widely credited models of protein-level regulation demonstrating a much stronger dependency between mRNA transcription and protein levels. Multi-omics protocols operating at a single cell resolution tend to exacerbate sparsity compared to older transcriptome and proteome experiments as the starting amount of DNA is even less, leading prominent studies like Angermueller et al. (2016); Clark et al. (2018) and Hernando-Herraez et al. (2019) to identify only a handful significant associations between promoter DNA methylation and gene expression despite methylation's recorded silencing properties. Therefore, development of bespoke methods for addressing inherent single cell multi-omics protocol limitations are vital for accurately interrogating these data.

In this thesis, we use the Bayesian framework to develop probabilistic models that explicitly and transparently decouple technical variability from biological signal to decipher complex relationship between epigenome and transcriptome. Drawing inspiration from Csárdi et al. (2015) we introduce SCRaPL (Single Cell Regulatory Pattern Learning) (Maniatis et al., 2022), a hierarchical model to infer correlation between different omics components in various genomic regions. SCRaPL's tailored likelihoods and uncertainty quantification combined with Bayesian equivalent of type I errors conceptualization method, allows for accurate results and good control over false positives compared to statistical correlations metrics like Pearson and Spearman correlation. Then we introduce single cell Multi View Inference (scMVI), a multi-modal Variational Autoencoder trained on jointly

quantified epigenomic and transcriptomic data. scMVI's aim is to impute and integrate partially observed molecular layers of similar cells post-experimentally, addressing protocol limitations and improving downstream analysis.

1.1 Contributions

In this thesis we apply probabilistic models to data gathered from modern single cell sequencing protocols, jointly quantifying epigenome and transcriptome data, to shed light in their complex interactions. The contributions of this thesis can be summarized as follows:

1. The development of SCRaPL, a novel Bayesian hierarchical method, to overcome limitations of traditional tools for detecting strong regulatory patterns between epigenome and transcriptome, at a genomic level (Maniatis et al., 2022).

2. The development of scMVI, an extension to standard variational auto-encoders. In data with partial information about different molecular layers, it can accurately integrate or impute missing modalities. In multi-omics data, it can create biologically meaningful latent spaces aiding in downstream interpretative analysis.

1.2 Thesis Layout

We hope that researchers across various fields will take interest in this thesis. To make it accessible to general audience in molecular and computational biology, care has been taken to provide a concise introduction to the required background. In Chapter 2, we introduce the concepts from molecular biology, focusing on the regulatory role of epigenome in gene expression and the modern single cell multi-omics sequencing protocols for jointly assaying epigenomic and transcriptomic data. In Chapter 3, we introduce statistical modelling, demonstrate how the Bayesian paradigm provides a robust template for quantifying uncertainty and provide the necessary background methods employed during inference. In Chapter 4, we provide some background for the computational methods devised for single cell multi-omics. In Chapter 5, we introduce SCRaPL, a Bayesian model using novel single cell multi-omics data to quantify statistical associations between epigenome and transcriptome at a gene level. Chapter 6 introduces scMVI, a flexible method aimed at facilitating downstream analysis in fully and partially quantified data for existing and future single cell multi-omics protocols. In Chapter 7, we conclude the thesis, summarizing key contributions and discussing future paths, such as devising a prior which allows for information sharing between co-expressed genomic regions.

Chapter 2

Molecular biology and epigenetics

2.1 Foundations of molecular biology

Deoxyribonucleic Acid (DNA) was accidentally discovered in 1869 by Swiss chemist Friedrich Miescher when he was trying to characterize the protein components of white blood cells. Since the isolated molecule was from the cells' nuclei, he named it nuclein (Dahm, 2008). Forty years later, Phoebus Levene proposed the phosphate-sugar-base polynucleotide structure of a single nuclotide and the carbohydrate components of DNA (Levene, 1919). Important work of Erwin Chargaff (Chargaff, 1950)[1] on the DNA chemical properties along with X-ray crystallography work by Rosalind Franklin and Maurice Wilkins (Dahm, 2008; Franklin and Gosling, 1953) paved the way for the derivation of the three dimensional, double helical model for the DNA structure (Watson and Crick, 1953).

DNA is a double stranded helix carrying hereditary instructions vital for the development and function of all known organisms. Each strand has a backbone made of alternating sugar-phosphate groups. Attached to each sugar is one of four nitrogenous sidegroups: adenine (A), guanine (G), cytosine (C) and thymine (T). Bases of each strand protrude into the center of the double helix, where they are paired with the bases of the other strand with hydrogen bonds. A and C in one strand are always paired with T and G from the other strand (Figure 2.1a)(Alberts et al., 2007).

During DNA replication, the double stranded helix uncoils at specific locations and hydrogen bonds are broken using initiator protein helicase. As hydrogen bonds are weaker compared to sugar-phosphate links, DNA strands are pulled apart with the DNA backbone remaining intact. Each strand of the original DNA can serve as a template for the formation

[1] Chargaff introduced a "rule" stating that no matter the organism or tissue, the total amount of purines (adenine and guanine) is equal to the total amounts of pyrimidines (cytosine and thymine).

of an entirely new strand. Hence, each time a cell divides the two daughter cells will inherit the same information from the parent cell (Alberts et al., 2007).

When necessary for a cell to sustain and reproduce itself, instructions are sent from the nucleus to ribosomes for protein production. Protein production requires large number of instruction replicates and it is extremely inefficient to achieve by replicating DNA (Richards and Hawley, 2011). To transfer specific region information in large scale, the cell relies on a more flexible molecule called Ribonucleic Acid (RNA). The chemical structure of RNA presents many similarities to DNA. The two prominent differences can be spotted on the backbone and in the nucleobase. Instead of sugar deoxyribose, RNA contains the sugar ribose and instead of T it contains uracil (U) (Figure 2.1a). In terms of structure and length, RNA molecules are single stranded and shorter. They can adopt complex three dimensional structures, unlike DNA which often pairs in double strands (Alberts et al., 2007). For protein synthesis three main types of RNA are involved; messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). In this thesis we focus on mRNA, whose role is to carry instructions from the DNA to cell's cytoplasm for protein production.

2.2 Regulation of Gene Expression

In multicellular organisms, different cell types differ drastically in both structure and function. When comparing a mammalian neuron with a liver cell, the differences are so stark that make it hard to imagine the two cells differentiated from the same stem cell, let alone they posses the same DNA (Alberts et al., 2007). Cells orchestrate the synthesis and accumulation of functional gene products, such as proteins and non-coding RNA, through a process known as gene expression (Richards and Hawley, 2011)[2]. This process explaining the flow of genetic information in a biological system is referred as the central dogma of molecular biology (Crick, 1958) (Figure 2.1b).

Gene expression is used by eukaryotes and prokaryotes to interpret information encoded in the DNA. In a process known as transcription, different parts of the DNA are used to synthesize RNA (Crick, 1958). This process is performed by special enzymes called RNA polymerases and yield the RNA molecules described above[3]. In eukaryotes, RNA

[2]Depending on whether a DNA fragment is used in protein synthesis or to facilitate other biochemical activities, it is characterized as coding or non-coding DNA. In the mammalian genome, it has been estimated that approximately 2% is protein coding. Susumu Ohno popularized the term "junk DNA" which describes non-coding DNA regions (Meyer and de Peer, 2003), sparking controversy (Doolittle, 2013; Ponting and Hardison, 2011). Later, the international Encyclopedia of DNA Elements (ENCODE) indicated that large parts of the non-coding DNA is important in gene expression (Kellis et al., 2014).

[3]Transcription in eukaryotes relies on three RNA polymerase enzymes (RNA polymerase I-III). These enzymes are structurally similar and share some sub-units. RNA polymerase I and III transcribe the genes

polymerase II tightly binds to special cis-regulatory sequences called promoters, opening the double helix, and starts to transcribe DNA to mRNA (Watson et al., 2015). The promoter is usually upstream of Transcription Start Site (TSS) and its length can reach thousands of base pairs. To initiate transcription, special binding proteins called transcription factors are used. RNA polymerase II continues to transcribe mRNA until it reaches the terminator, a special sequence marking the end of gene. The mRNA along with other functional gene products is directed to the ribosomes for protein synthesis (Crick, 1958).

Life depends on the ability of cells to express genes in different combinations at different times and tissues. Some housekeeping genes, expressed all the time in all cells, are often required for the basal cellular function. Other genes aid development of the organism or are expressed in response to external stimuli (Richards and Hawley, 2011). In this chapter, we give a short review of the regulatory mechanisms relevant to the thesis. Starting with the ones taking place at the genome level, we give a gentle introduction into some aspects of gene expression regulation. This serves as a stepping stone to take a deep dive into the more complex and perhaps more fascinating role of epigenome.

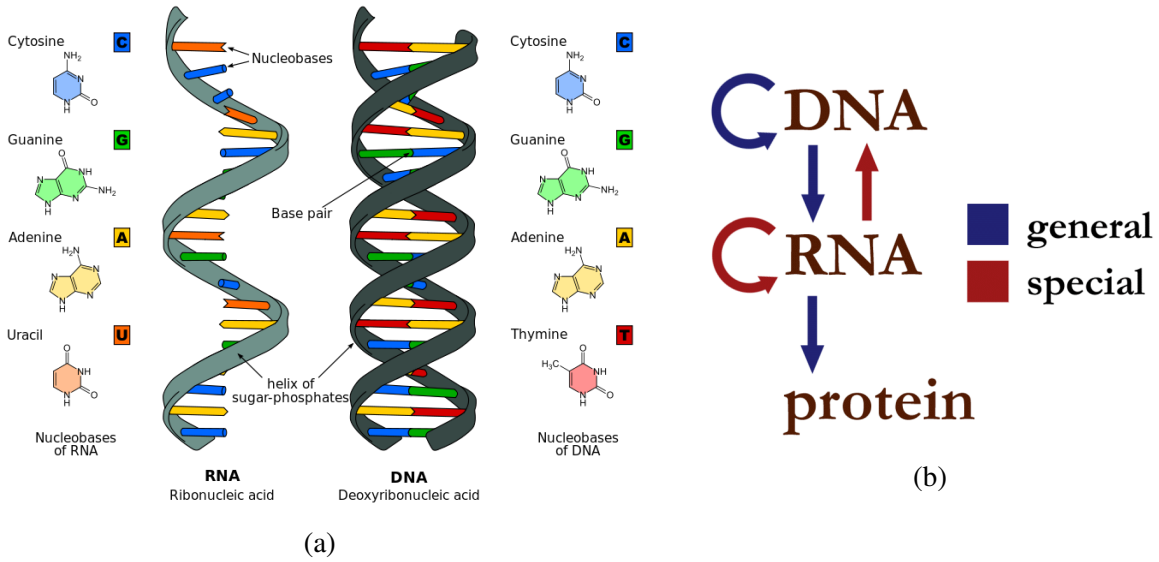


Fig. 2.1 **Core concepts of molecular biology.** (2.1a) Comparison between DNA double helix (left) with RNA single strand; figure taken from Sponks/Wikimedia Commons/CC BY-SA 3.0. (2.1b) Central dogma of biology illustration explaining the flow of genetic information from nucleus to protein production; figure taken from Narayanese at English Wikipedia, Public domain, via Wikimedia Commons

encoding tRNA, rRNA and various small RNAs. RNA polymerase II transcribes most genes including those that encode proteins. A detailed explanation regarding RNA polymerase enzymes in prokaryotes and eukaryotes can be found in Alberts et al. (2007).

2.2.1 Regulatory elements of gene expression

The pathway leading from DNA to protein includes many steps which can, in principle, be regulated. The ability of cells to coordinate gene expression is termed “regulation of gene expression" (Alberts et al., 2007). Inside the nucleus, cells in eukaryotes can regulate protein production by selecting which gene, in what order and at what rate are transcribed to mRNA, a process known as transcriptional regulation [4]. For the remaining of the thesis, regulation of transcription and regulation of gene expression are going to be used interchangeably.

Regulation of transcription is a complex phenomenon with many routes of influence. It relies on a combination of proteins and DNA sequences, called transcription regulators and cis-regulatory sequences (Watson et al., 2015). Transcription regulators bind to these sequences, which are dispersed throughout genome. This binding puts into motion a series of reactions that ultimately specify which genes are to be transcribed and at what rate (Watson et al., 2015). Depending on whether regulators encourage transcription or not, they are categorised as activators or repressors. In most cases, a given transcription regulator recognizes its own cis-regulatory sequence, which is different from those recognized by other regulators in the cell (Watson et al., 2015).

Transcription regulators operate in various binding sites, with the most obvious being promoters. Other sites that have come under scrutiny for their interesting properties are enhancers. Enhancers [5] are short segments of the DNA that can be bound by activators to increase the likelihood of transcription. Their function is both to attract and position RNA polymerase II at the promoter and to release it so that transcription can begin (Alberts et al., 2007). Enhancers bind different groups of transcription regulators and control expression of the same gene in response to different signals (Alberts et al., 2007). A gene might thus contain many such enhancers, each of which contributes in a cumulative manner to the overall regulation of the gene (Blackwood and Kadonaga, 1998). In the human genome, there are several thousands[6] enhancers located up to a million base pairs upstream or downstream away from the promoter they control (Pennacchio et al., 2013).

Transcriptional regulation can also by achieved through transcription factors. Transcription factors are special proteins that bind to specific DNA sequences (i.e. enhancers or

[4]Transcriptional regulation is followed by post transcriptional steps controlling the processing of RNA transcripts (Modrek and Lee, 2002), transportation and localization controls along with translation and protein activity control (Colgan and Manley, 1997; Mann and Jensen, 2003). More can be found in Alberts et al. (2007)

[5]Historically, the identification of enhancers has proved challenging. Enhancers are scattered across the human genome, at varying distances from the promoters they regulate. Furthermore, enhancers can regulate several promoters sometimes bypassing closest ones. Finally, enhancers’ activity can be restricted on specific tissues or cell types (Pennacchio et al., 2013)

[6]Surveys of tissues have indicated they vastly outnumber the protein encoding regions (Pennacchio et al., 2013)

promoters), to control the rate of transcription (Latchman, 1997). They are a subset of the DNA binding proteins and work alone or with other proteins in a complex, either by activating or by repressing the recruitment of RNA polymerase II (Nikolov and Burley, 1997). Transcription factors help to position the RNA polymerase at the promoter, aid in pulling apart the DNA strands to initiate transcription and release the RNA polymerase from the promoter (Alberts et al., 2007). During transcription in mammals, cell-function specific and general transcription factors bind to the enhancer and promoter respectively. Then, the DNA loops in order for the enhancer to come into close proximity with the promoter. The loop is stabilized by two connected proteins, one anchored in the enhancer and one in the promoter (Weintraub et al., 2017). Transcription factors are activated, triggering the activation of the enhancer and the target promoter. RNA polymerase II starts synthesizing the mRNA for the target gene, governed by a subset of enhancer-bound transcription factors (Mobley, 2019) (Figure 2.2).

Regulation of gene expression is an extremely complicated process with strong impact on the phenotype of cells. By adjusting differential gene spatiotemporal expression, cells can adapt to external stimuli and differentiate to serve a diverse set of purposes. In this chapter, we examined direct ways for regulating expression through DNA interactions and modulation of transcription factors. We intentionally omitted to discuss epigenetic modifications which we will introduce later.

2.2.2 Epigenetic regulation of gene expression

To explain aspects of development for which there was little understanding, Conrad Waddington defined the term “epigenetics" as changes in phenotype without changes in genotype (Waddington, 1942, 2012). Eighty years later, we know that epigenetic mechanisms control gene expression patterns without altering the underlying genome[7] (Allis and Jenuwein, 2016).

Later, Waddington proposed the concept of epigenetic landscape as a metaphor for representing various pathways a cell might take while developing (Creighton and Waddington, 1958). During differentiation, a cell would traverse through a series of ridges and valleys, from one steady state to the next (Navis, 2012) (Figure 2.3). Recently, different methods have been proposed for quantifying this landscape, in an attempt to get insights about the

[7]Epigetics affect the phenotype, in a similar way the absence or presence of commas can change the entire meaning of a sentence. An interesting example can be found in Chronicon of Aubry de Trois-Fontaines where a sibyl responded “Ibis, redibis nunquam per bella peribis" to a general consulting about his fate in an upcoming battle. If comma is placed after “redibis", the meaning is “you will go, you will return, never in war will you perish". If the comma is placed after “nunquam", the meaning of the sentence completely changes to “you will go, you will return never, in war you will perish".

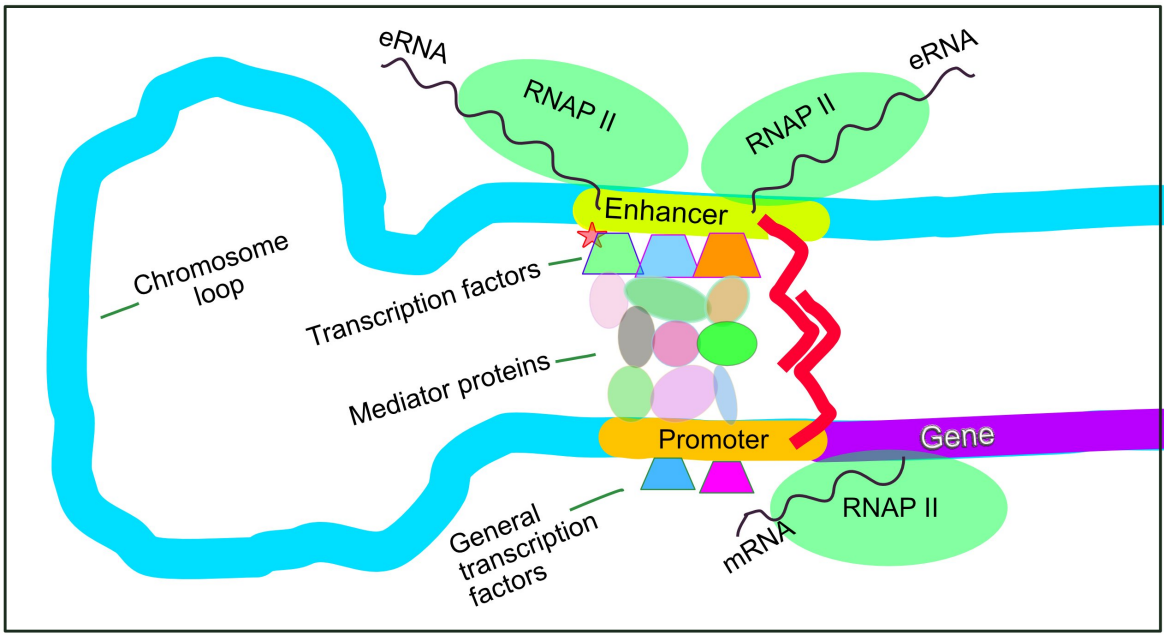


Fig. 2.2 **Regulation of transcription sketch in mammals.**Transcription factors interact with sequences of DNA to promote and enhance transcription. Activating transcription factors are epigenetic regulators of specific genes. DNA bending protein brings the activators into the proximity of the promoter where additional transcription factors are bound and interact with the activators. Transcription factors at the promoter recruit RNA polymerase and transcription can proceed (Mobley, 2019). Figure taken from Bernstein0275/Wikimedia Commons/CC BY-SA 4.0 .

optimal routes for cell reprogramming (Huang, 2009; Wang et al., 2010; Zhou and Huang, 2011)

Epigenetic processes are natural and essential to many functions, but if they occur improperly, there can be major adverse health effects. Hence, their study is vital for understanding and, wherever necessary, reversing them. In this chapter, we focus on perhaps the best known epigenetic processes, namely chromatin modification, histone modification and DNAm.

2.2.2.1 Chromatin organisation and Histone modification

In eukaryotic cells, chromatin is a highly organized complex of DNA and associated proteins used to tightly pack DNA into the cell's nucleus (Maloy and Hughes, 2013). Chromatin is comprised of nucleosomes, with their basic protein components known as histones. Histones are highly conserved proteins whose positive charge helps them bind to the negatively charged backbone of the DNA. Around the core, DNA is tightly wrapped in a left-handed coil of 1.7 turns. In assembling a nucleosome, the histones (H2A, H2B, H3, H4) first bind to each other to form H3–H4 and H2A–H2B dimers, and the H3–H4 dimers combine to form

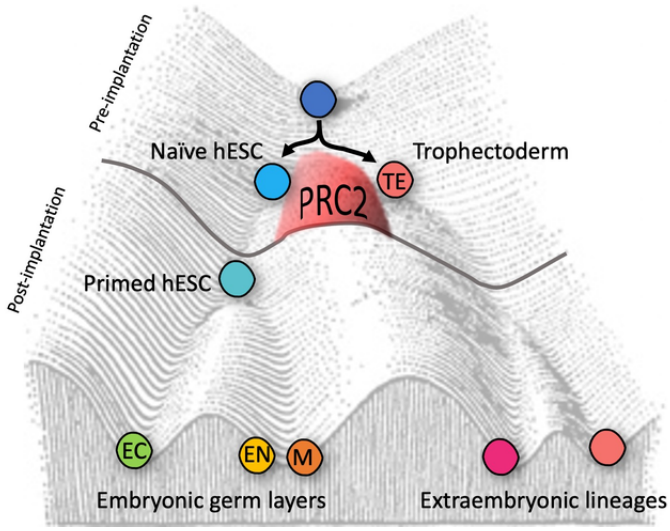


Fig. 2.3 **Waddington Epigenetic Landscape for human ESC.** A metaphor representing various differentiation pathways a cell might make during development. For cell to change phenotypic properties it has to traverse through valleys and ridges from one steady state to the next. Figure taken from Simon Elsasser/PD.

tetramers. An H3–H4 tetramer then further combines with two H2A–H2B dimers to form the compact octamer core, around which the DNA folds (Alberts et al., 2007). Successive nucleosomes are connected together into higher order structure resembling the “beads of string", which are folded in 30*nm* compact fibre and stabilized by H1 histone (Felsenfeld and Groudine, 2003). Compact fibres containing tens of thousands of nucleotide pairs are further organized into a series of looped domains (Alberts et al., 2007) (see Figure 2.4). These looped domains can limit enhancers’ ability to interact with genes outside the loop, regulating expression (Kellum and Schedl, 1991). In this thesis, we are mainly concerned with the impact of chromatin organization on gene expression. We briefly introduce histone modifications for comprehensiveness.

In the eukaryotic cells, studies have distinguished two types of chromatin, a highly condensed form called heterochromatin and a less condensed which goes by the name euchromatin (Alberts et al., 2007). Locally, the chromatin can become quite dynamic with euchromatin being converted to heterochromatin and backward, blocking or facilitating transcription (Alberts et al., 2007; Maloy and Hughes, 2013). Histone modifications and DNAm are deeply connected with chromatin organisation (Lee, 2011). However, it is still not clear how these configurations are established and stably inherited to daughter cells (Cedar and Bergman, 2009).

Histones are subject to a vast array of reversible post-translational modifications, promoting or inhibiting transcription (Cohen et al., 2011; Martin and Zhang, 2005). These modifications are vital for the structural stability of the nucleosome, with methylation and acetylation lysine residues often affecting transcription (Biswas et al., 2011; Sadakierska-Chudy and Filip, 2015). Histone methylation is associated with transcriptional regulation by influencing chromatin architecture, recruiting transcriptional factors and interacting with initiation (Greer and Shi, 2012; Zhao and Shilatifard, 2019). In mouse embryonic stem cells, histone methylation is important as it maintains pluripotency by gene silencing while it keeps cells poised for activation during subsequent developmental stages (Azuara et al., 2006; Bernstein et al., 2006; Hu et al., 2013; Zhao and Shilatifard, 2019). Histone acetylation is a reversible modification on lysine residues that can promote or obstruct open chromatin by being bound to a variety of transcription factors (Josling et al., 2012; Sawan and Herceg, 2010; Zhao and Shilatifard, 2019).

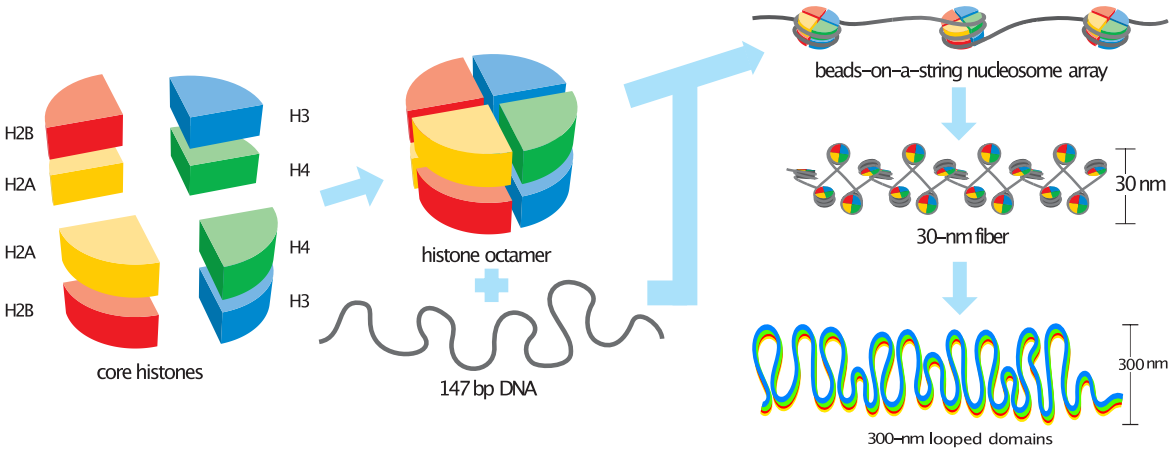


Fig. 2.4 **Units of chromatin organization.** Histones (H2A, H2B, H3, H4) first bind to each other to form H3–H4 and H2A–H2B dimers, and the H3–H4 dimers combine to form tetramers. An H3–H4 tetramer then further combines with two H2A–H2B dimers to form the compact octamer core, around which the DNA folds (Alberts et al., 2007). Successive nucleosomes are connected together into higher order structure resembling the “beads of string" which are folded in 30*nm* compact fibre. Compact fibres containing tens of thousands of nucleotide pairs, are further organized into a series of looped domains (Alberts et al., 2007). Figure adapted from David O Morgan/Wikimedia Commons/CC BY-SA 3.0 .

2.2.2.2 DNA Methylation

DNA methylation (DNAm) is a heritable epigenetic mark involving the transfer of a methyl group to the cytosine ring of DNA by DNA methyltransferases (DNMTs)[8] (Jin et al., 2011). In many cases, DNAm occurs in CpG dinucleotides (Lister et al., 2009)[9]. It was initially discovered in calf thymus cells (Hotchkiss, 1948) and its regulatory role in transcription was identified by Holliday and Pugh (1975) and Riggs (1975). Since then, much evidence has been obtained, recognising DNAm as the chief contributor to stability of gene expression (Jaenisch and Bird, 2003).

DNMT interactions with transcription factors at promoter regions have been suggested as a prominent way for DNA to silence genes (Hervouet et al., 2009). The dual ability of some transcription factors to attach to the genome and interact with DNMTs may enable widespread promoter DNAm. This site-specific methylation is responsible for the assembly, at these locations, of proteins that directly influence actions of transcriptional machinery (Kass et al., 1997; Wade, 2001), or cause alternations in chromatin structure which subsequently affects gene expression (Gibney and Nolan, 2010; Prokhortchouk and Defossez, 2008). Croce et al. provided initial evidence demonstrating that DNMT3a was recruited to the $RAR\beta 2$ promoter by the oncogenic transcription factor PML-RAR, leading to promoter methylation and silencing of the $RAR\beta 2$ gene. Hervouet et al. found direct interactions between 79 unknown transcription factors and enzymes of DNMT family in various promoters, with transcription acting as potential DNA anchors for DNMTs, aiding in site specific methylation (Gibney and Nolan, 2010).

DNAm can suppress transcription through the route of excluding proteins that can affect transcription (Jaenisch and Bird, 2003; Watt and Molloy, 1988). An example could be the exclusion of the CTCF protein, which prevents the interactions between an enhancer and a promoter (Bell et al., 1999; Ohlsson et al., 2001). DNAm blocks the binding of CTCF, allowing the enhancer to stimulate promoter activity (Jaenisch and Bird, 2003). This has implications in the imprinting of Igf2 gene which is exclusively expressed from the paternal allele during development (Hark et al., 2000; Ohlsson et al., 2001). There are also cases where DNAm attracts proteins, which go by the name methyl CpG biding. Example of these involve the methyl-CpG-binding domain (MBD): MBD1-MBD4 and MECP2 (Hendrich and

[8]Three DNMTs, namely DNMT1, DNMT3a and DNMT3b are orchestrated in catalysing DNAm. DNAm is classified as either *de novo* or maintained methylation. DNMT1 has a higher catalytic activity to preferentially methylate hemimethylated DNA during replication and is mostly responsible for maintaining the DNAm status. DNMT3 isoforms, namely DNMT3a and DNMT3b, are characterized as *de novo* since they prefer to bind to previously unmethylated DNA independently of replication (Lu et al., 2020).

[9]In mammal somatic cells, more than 98% of DNAm occurs in CpG dinucleotides, while as much as a quarter of all methylation appears in non-CpG context in ESC (Lister et al., 2009)

Bird, 1998). MECP2 is probably the most well characterized member of the MBD family. It binds with methylated CpG, via its MBD domain, and exerts its transcriptional repressive effects several hundred base pairs away (Newell-Price et al., 2000; Razin, 1998). Furthermore, MECP2 can alter chromatin compaction by binding to linker DNA and nucleosomes, resulting in a physical barrier to the transcriptional machinery (Gibney and Nolan, 2010).

After fertilization, there are many active parental genes involved in normal embryonic development. However, a subset of genes are selectively expressed from parental allele referred as imprinted (Reik et al., 2001; Tilghman, 1999). Imprinted genes located within GpG rich regions are mostly in clusters and are controlled through various epigenetic processes, including DNAm and histone modification (Branco et al., 2008; Golbabapour et al., 2011; Ikegami et al., 2009; Vaissière et al., 2008).

2.3 Sequencing technologies

DNA sequencing is a process used to read the sequence of nucleotides on a DNA strand. The founding methods for DNA sequencing were the Sanger dideoxy synthesis and the Maxam-Gilbert chemical cleavage methods (Maxam and Gilbert, 1980; Sanger and Coulson, 1975; Sanger et al., 1977; Slatko et al., 2018). Since its introduction, the Maxam-Gilbert method had undergone various chemistry modifications. Despite that, Sanger sequencing became the standard (Slatko et al., 2018) and was used to determine the base pairs that make up the human DNA (Lander et al., 2001; Venter et al., 2001). Even though Sanger sequencing is expensive to scale (Mardis, 2008), it is still preferred for its accuracy by many DNA core facilities and sequencing for profit companies (Slatko et al., 2018).

Next generation sequencing (NGS) or high throughput sequencing (HTS) protocols, were introduced at the onset of 21st century to improve cost along with speed and sequencing parallelization. In contrast to Sanger sequencing, NGS does not require physical separation of individual reactions into several tubes, capillaries or lanes (Alekseyev et al., 2018). Instead, the sequencing reactions occur in parallel on a solid surface and are only spatially separated (Alekseyev et al., 2018). Regardless of the platform, NGS involves several common steps [10]. Initially, DNA is broken into smaller fragments using ultrasonic frequencies. Sequencing libraries are then prepared by ligating specific enzymes to both ends of the fragments. DNA fragments in the sample are then amplified, sequenced and mapped to a reference genome or in cases where the genome is unavailable *de novo* assembly is performed (Alekseyev et al., 2018; Zerbino and Birney, 2008).

[10]Detailed reviews regarding the underlying steps of different platforms can be found in Heather and Chain (2016); Morey et al. (2013); Reuter et al. (2015).

Since their introduction, NGS technologies have served as powerful tools to study genomic traits and provide key insights in various research fields. To date, most methods used to quantify molecular state of a cell population, are based on estimating mean behaviours obtained for ensembles of millions of cells (Shapiro et al., 2013). This is typically referred to as "bulk" samples. Although useful, this is not ideal as it masks cell-specific features making the investigation of complex biological phenomena, such as early stem cell differentiation, impossible to study. Recently, developments in cell isolation and barcoding have enabled the development of methods for single cell analyses which help address these limitations (Dimitriu et al., 2022). Further breakthroughs in the processing of intracellular components have lead to the introduction of protocols that jointly quantify various omes (i.e. genome, epigenome, transcriptome and proteome) at a single cell resolution (Chappell et al., 2018).

Since the introduction of first single cell multi-omics protocol (Han et al., 2014), the field has seen massive development. Novel and diverse approaches that followed have improved throughput, coverage and sequencing depth, hinting promising future discoveries (Dimitriu et al., 2022). In this chapter we examine the key characteristics of single cell multi-omics protocols, mainly focusing on methods that simultaneously quantify epigenome and transcriptome (see Figure 2.5 for summary of disccused protocols.).

2.3.1 Fundamentals of of single cell omics and multi-omics protocols

Irrespective of the molecular layers quantified by a single cell (multi-)omics protocol, it is essential to isolate and barcode cells from the starting population. Cell isolation and barcoding is essential for post-sequential mapping of the reads to specific cells (Dimitriu et al., 2022). Depending on the isolation procedures, techniques are catalogued into three categories; well based, continuous flow microfluidics based and droplet microfluidics based techniques. In well based techniques, cells are distributed using fluorescence-activated cell sorting (FACS) into multi-well plates containing unique barcodes. Subsequently, these barcodes attach to the cell of interest. Continuous flow microfluidics use chip values to trap single cells in discrete reaction chambers where each step of the workflow occurs (Dimitriu et al., 2022; Philpott et al., 2020; Shim et al., 2013). Droplet microfluidics techniques, co-encapsulate single cells with barcoded beads into microdroplets (Macosko et al., 2015). Both well and microfluidics based approaches have their advantages and disadvantages. On the one hand, the well based is more concentrated, minimizing amplification bias, ultimately increasing precision (Ballantyne et al., 2006; Dimitriu et al., 2022). On the other hand, microfluidics reduce reagent use, minimize environmental contaminants and, in the case of droplet microfluidics based techniques, can achieve a magnitude higher throughput (Dimitriu et al., 2022). The data used for experiments in this thesis are gathered with well and droplet

microfluidics based protocols. In this document, we refer to well and droplet microfluidics based protocols as plate and droplet based protocols respectively.

Cell barcoding is the process of attaching a unique oligonucleotide, generating a unique cellular identifier. It allows libraries from multiple single cells to be jointly sequenced in the same pool, with cell specific data being deconvoluted post-experimentally (Chappell et al., 2018). In plate based protocols, barcoding is performed in the final Polymerase Chain Reaction (PCR) step before sequencing, with the combined pool of sequencing libraries generated in the final steps (Chappell et al., 2018). In microfluidics based techniques, barcodes are added near the beginning, enabling pool of libraries to be made earlier. A third type of barcoding called combinatorial indexing methods (Cao et al., 2017; Chappell et al., 2018; Cusanovich et al., 2015; Ramani et al., 2017) have also started gaining popularity. Groups of fixed and permeabilized cells are FACS sorted per well of a plate, with each well receiving a different short barcode that can be associated with the target well. Cells from all wells are then pooled and re-sorted into small groups per well, where additional barcodes are added to the wells(Chappell et al., 2018). With enough repetitions and sufficient barcodes, thousands of cells can be uniquely tagged.

2.3.2 Gene expression sequencing

Examining gene expression at the single-cell resolution allows the interrogation of inherent features of transcription, such as transcriptional noise and regulated cellular heterogeneity. Single cell resolution is essential in elucidating coexpression and mutually exclusive expression of genes within cells (Chappell et al., 2018).

There are various protocols for measuring transcription, under the umbrella term of single cell RNA sequencing (scRNA-seq). These have varying levels of sensitivity and accuracy [11] (Chappell et al., 2018). The steps required for single cell transcription quantification include; single cell isolation, cell lysis, reverse transcription (i.e. conversion of RNA to cDNA), cDNA amplification and library preparation (Jovic et al., 2022; Ziegenhain et al., 2017). Depending the protocol, isolation is either performed with FACS or using a microfluidic system and laser microdissection. Then transcripts from each cell are barcoded and are reversed transcribed to complementary DNA (cDNA) (Jovic et al., 2022). Subsequently cDNA fragments are amplified either using PCR or *in vitro* transcription (IVT) (Jovic et al., 2022; Kolodziejczyk et al., 2015). PCR amplification is commonly used in Smart-seq, Smart-seq2, 10X genomics among others (Goetz and Trimarchi, 2012; Jovic et al., 2022; Picelli et al., 2013; Zheng et al., 2017) and it is the amplification technique used to gather data of this thesis. To

[11] A comparative analysis can be found in (Ziegenhain et al., 2017).

overcome amplification biases, some protocols introduce unique molecular identifiers (UMIs) to barcode each individual mRNA molecule within a cell in the reverse transcription step (Jovic et al., 2022). Once the amplification is finished, cDNA is sequenced using NGS (Jovic et al., 2022).

In recent years, single cell RNA protocols have managed to exponentially increase the number of cells processed within a sample (Chappell et al., 2018; Rozenblatt-Rosen et al., 2017; Svensson et al., 2018). Making the most appropriate protocol choice is often a balance between sequencing depth and scale. Plate based methods like Smart-seq2 are able to process hundreds and in some cases thousands of cells with good sequencing depth. On the other hand, droplet based protocols like 10X Genomics platforms can scale up to millions of cells at low cost, at the cost of increased data sparsity (Chappell et al., 2018).

2.3.3 Genome accessibility sequencing

An assessment of how accessible certain parts of the genome are can be performed either by capturing information about the histone modifications, or by profiling the open chromatin. By looking at histone modifications, researchers could quantify changes in chromatin structure and eventually predict future transcriptional states. The droplet based chromatin immunoprecipitation (Drop-ChIP)(Rotem et al., 2015), which was designed for that purpose, has not been very popular due to a combination of low coverage and background noise caused by nonspecific antibody pull-down (Clark et al., 2016; Rotem et al., 2015). Alternatively, open chromatin can be used to infer the usage of different cis-regulatory sequences (i.e. enhancers or promoters), or transcription factors binding sites (Chappell et al., 2018).

The most popular protocol for quantifying open chromatin at cellular level is an adaptation of the assay for transposase-accessible chromatin (ATAC), namely scATAC-seq (Buenrostro et al., 2015). In scATAC-seq, cells are isolated using a microfluidics platform. In turn, accessible regions are tagged with hyperactive Tn5 transposase, barcoded using cell-dentifying primers, are amplified with PCR and are sequenced using NGS. The creation of barcode-carrying transposase has also enabled the development of a similar protocol that uses combinatorial indexing to barcode cells (Cusanovich et al., 2015). Other protocols for quantifying accessibility include an adaptation of DNase-seq (Cooper et al., 2017; Jin et al., 2015) and single cell nucleosome occupancy and methylation sequencing (scNOMe-seq) (Pott, 2017). scNOME-seq will be discussed in detail later.

2.3.4 DNA methylation sequencing

The standard approach for profiling DNAm on a genome wide scale is through bisulfite sequencing (BS-seq) (Frommer et al., 1992). Initially, bisulfite treatment of the sample converts unmethylated cytosines to uracils while keeping methylated cytosines intact. In subsequent steps, fragments are amplified using PCR, sequenced and mapped to a reference genome (Krueger and Andrews, 2011). The PCR amplification step serves two purposes. On the one hand, it allows NGS to sequence an adequate amount of fragments from a small sample. On the other hand, the DNA polymerase reads uracils as thymines allowing them to be discriminated from methylated cytosines which remain intact (Krueger et al., 2012). There are multiple methods for capturing DNAm at a single cell resolution, with most of them use BS-seq as their basis. Reduced representation bisulfite sequencing (RRBS) (Meissner et al., 2005) and whole genome bisulfite sequencing (WGBS) (Cokus et al., 2008; Lister et al., 2008) along with their single cell adaptations (Farlik et al., 2015; Guo et al., 2015; Kobayashi et al., 2016; Luo et al., 2018; Ma et al., 2018; Mulqueen et al., 2018; Wang et al., 2015) are popular genome wide methylation profiling methods (Ahn et al., 2021)[12].

Similarly to RRBS, scRRBS can be divided into five steps. The first two are to digest DNA using specific enzymes and adaptor ligation followed by preprocessing (Ahn et al., 2021). Then GC-rich sites are enriched using gel size selection followed by bisulfite treatment. Until bisulfite treatment, several purification steps are performed, to guarantee that the target DNA is maintained (Ahn et al., 2021). In turn, the sample is amplified and sequenced with NGS (Ahn et al., 2021). The scWGBS and single cell bisulfite sequencing (scBS-seq) (Clark et al., 2017) protocols are often modified versions of (Miura et al., 2012) and try to avoid bisulfite induced fragmentation of adaptor tagged DNAs. They consist of a bisulfite treatment step followed by adapter tagging and two rounds of random primer extensions before sequencing (Ahn et al., 2021). The most widely adopted are WGBS single cell variations along with scBS-seq, as they require fewer purification steps and can cover most CpGs in the genome (Ahn et al., 2021).

Single cell methylation sequencing remains technically challenging, as the starting DNA material is limited and the bisulfite treatment causes extensive DNA degradation (Chappell et al., 2018; Hayatsu, 2008; Leontiou et al., 2015). In addition, two types of error can also occur during bisulfite treatment; over-conversion (leading to false-negative CpGs) and under-conversion (leading to false-positive CpGs), but it is possible to control for them using methylated and unmethylated DNA spike-in controls (Chappell et al., 2018).

[12] A detailed review with their similarities and differences could be found in (Ahn et al., 2021).

2.3.5 Single cell multi-omics protocols

In recent years, a range of frameworks have been developed combing the cell isolation, barcoding and sequencing strategies outlined before. They go by the name single cell multi-omics protocols and provide researchers with direct ways to characterize the coordination between various omics layers (e.g. epigenome and transcriptome) in individual cells, allowing the study of complex and highly dynamical phenomena (Dimitriu et al., 2022). Depending on the information each protocols seeks to extract, different types of processing are necessary. Here we consider multi-omics technologies aimed at jointly quantifying DNAm, gene expression and chromatin accessibility along with some of their building blocks.

A number of plate based (Reyes et al., 2019; Rooijers et al., 2019) and droplet based (Cao et al., 2018; Chen et al., 2019; Zhu et al., 2019) methods for parallel profiling of chromatin accessibility and gene expression in single cells have recently been published. In this thesis, we look at data gathered with the commercial Chromium Next GEM Single Cell Multiome ATAC + Gene Expression from 10X. Initially, the protocol isolates nuclei from the cell population and sorts them using FACS. Nuclei are then treated in bulk with hyperactive transposase enzyme to cut and insert sequencing adapters into exposed DNA, as in scATAC-seq. Transposed nuclei are then partitioned into droplets, where barcodes are attached to mRNA and transposed fragments through incubation. Droplets are then broken to allow fragments pass through a purification step. Samples then pass through amplification and are sequenced with NGS [13].

Protocols jointly quantifying DNA methylation and transcription were presented by Angermueller et al. (2016) and Hu et al. (2016). In single cell methylome and transcriptome sequencing (scMT-seq) (Hu et al., 2016), mRNA is released from a cell, while leaving the nucleus intact, through a process known as cell lysis (Chappell et al., 2018). The mRNA is then converted into an scRNA-seq library and sequenced using Smart-seq2 (Picelli et al., 2013). The nuclei are in parallel sequenced using scRRBS, yielding a reduced representation overview of methylocytosine in the genome (Chappell et al., 2018). Single cell methylome and transcriptome sequencing (scM&T-seq)(Angermueller et al., 2016) is based on the scG&T-seq protocol (Macaulay et al., 2015, 2016). Instead of direct genome sequencing it applies scBS-seq following the separation of DNA and RNA (Chappell et al., 2018). The difference between scMT and scM&T-seq is in the methylation profiling approach. The scBS-seq employed by scM&T-seq gives it better coverage of the DNAm methylation landscape at

[13] A detailed description of the protocol can also be found here https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/10x-multiomics-tech-note-m-amr-00006/10x-multiomics-tech-note-m-amr-00006.pdf

the cost of loosing some nucleic acids during DNA separation from RNA (Chappell et al., 2018).

To jointly interrogate DNAm and chromatin accessibility at a single cell resolution, single cell nucleosome and methylome sequencing (scNOMe-seq) (Pott, 2017) was introduced by adapting conventional NOMe-seq (Kelly et al., 2012). Using methyltrasferase (M.CviPI), scNOMe-seq methylates cytosines in exposed GpC dinucleotides (Chappell et al., 2018; Kelly et al., 2012; Kilgore et al., 2007). Then, scBS-seq simultaneously encodes CpG methylation and GpC accessibility (Chappell et al., 2018). With this approach, inaccessible chromatin is better distinguished from missing data and resolution is determined by the frequency of GpC sites instead library size (Chappell et al., 2018).

In this thesis, we look at stem cells gathered with single cell nucleosome, methylation and transcription sequencing (scNMT-seq) (Clark et al., 2018). scNMT effectively combines scNOME-seq with scM&T-seq which were described before. Cells are sorted with FACS into tubes containing M.CviPI before their DNA is separated from their RNA. Then, using scNOMe-seq, DNAm and chromatin accessibility are jointly profiled. Transcription is measured using Smart-seq2.

Multi-omics protocols mentioned here are a small subset of a constantly increasing family. Existing trends of larger gene expression datasets are likely to be followed by larger multi-omics datasets. This will probably be driven by the much needed refinement of single-omics technologies. For example, due to small amounts of input DNA per cell, scBS-seq only captures between $5\% - 20\%$ of the genome coverage (Angermueller et al., 2016). In cases where chromatin accessibility is also quantified (i.e. scNOME-seq), the assayed cytosines are reduced by 50% as GpCpG and CpCpG positions are discarded due to inability to distinguish between endogenous methylation, *in vitro* methylation and off target effects (Kapourani and Sanguinetti, 2018). When looking at scATAC-seq and scRNA-seq, improvements in sensitivity are essential. Currently, most gene expression protocols require at least 10 mRNA molecules to detect gene expression (Svensson et al., 2017). This yields to assays with a large number of zeros, under-powering downstream analyses. Challenges also need to be addressed when measuring multiple molecular layers at the same time, as there is a tendency for weaknesses present in single omics to be exacerbated in multi-omics sequencing (Dimitriu et al., 2022). Another difficulty that multi-omics assays need to overcome is related to cost, particularly for those protocols targeting the whole genome (Chappell et al., 2018). Suppressing reaction volumes could potentially reduce library costs, but avoiding exponential cost increases may prove extremely ambitious (Chappell et al., 2018).

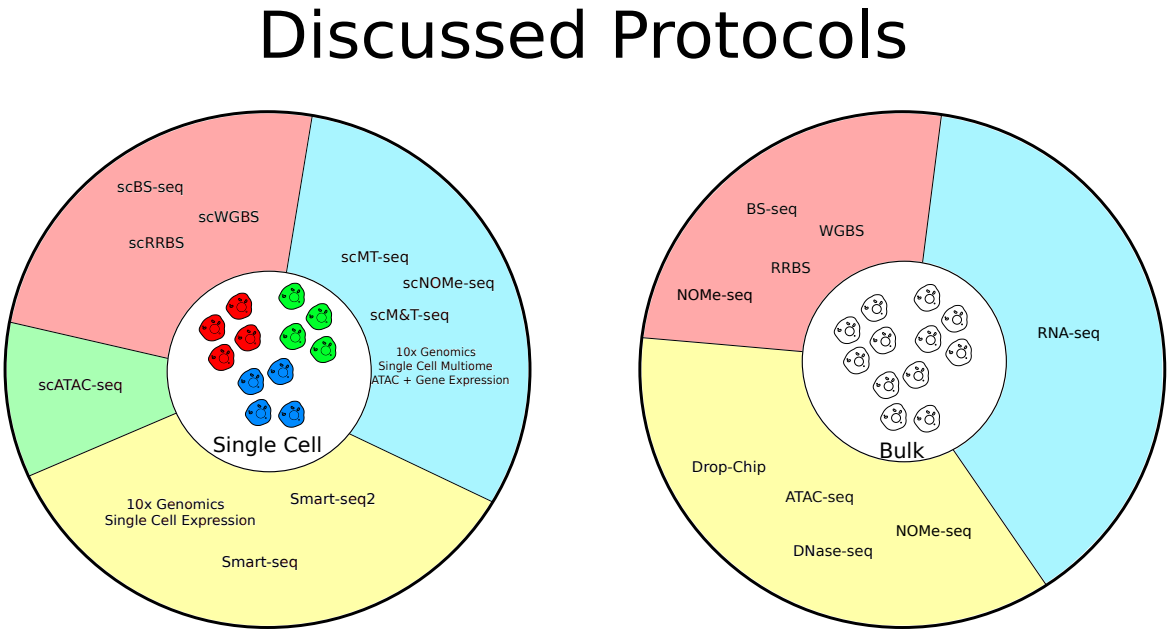


Fig. 2.5 **Sequencing technologies discussed .** Sequencing protocols used to interrogate epigenomic regulation of transcriptome at bulk and sigle cell resolution.

Chapter 3

Mathematical Background

3.1 Statistical Modeling - The Bayesian Perspective

Statistical modeling aims to reveal patterns in empirical data containing a combination of signal and noise. Assuming complex data generating processes, our aim is to approximate them with simpler yet realistic models (Lindsey, 1997). In order for these models to be effective, they must distinguish systematic variability related to the described phenomenon from random noise, possibly attributed to the limitations of data collection procedures. Depending on the phenomenon complexity, systematic variability could be quantified with models ranging from simple linear regression to large neural networks with specifically designed components. Hypothesized data generating mechanisms and their implied statistical models are scientific constructs aimed at providing important insights to their users[1]. As a result, statistical models resort to vast simplifications which in some cases diverge from reality. In a more descriptive context, modeling attempts to smooth stochastic data irregularities to detect patterns in data (Lindsey, 1997). Furthermore, identifying models that work well for one problem does not guarantee state of the art prediction levels in other tasks as effective modeling assumptions are not universal[2]. Hence, modeling could be thought as an effective tool in the hands of practitioners for interpreting rather than explaining problems whose results need careful handling (Robert, 2007).

The wealth of publicly available data (Dunham et al., 2012; Edgar et al., 2002; Sondka et al., 2018) has offered exciting opportunities to model complex biological systems and study life threatening diseases like cancer. We are in an era where high-throughput technologies can simultaneously measure tens of thousands of molecular targets per sample at

[1]This is nicely summarized in Box's popular aphorism "all models are wrong but some are useful" (Box, 1976).

[2]This observation is often called no free lunch theorem (Wolpert, 1996).

a steadily declining cost (Blair et al., 2012). Statistical modeling provides important tools and techniques for functionally organizing diverse piece of information stemming from such molecular data. In recent decades, following the automation of data generation techniques (Goel et al., 2006), novel statistical models have become an integral part of exploratory analysis (Colomé-Tatché and Theis, 2018).

In the field single cell omics field, statistical methods have followed a rapid growth with the scaling of modern data generation technologies. Initially, single cell multi-omics analysis and integration were performed with linear dimensionality reduction tools (e.g. Argelaguet et al., 2018; Butler et al., 2018) and simple correlative analyses (Angermueller et al., 2016). The introduction of deep probabilistic models (Lopez et al., 2019, 2018) has opened up new realms of possibilities. State of the art methods based on Variational Autoencoders (e.g. Gayoso et al., 2021; Minoura et al., 2021) perform extremely well in traditional multi-omics problems like integration, enabling the exploration of cell heterogeneity on epigenomic, transcriptomic and proteomic levels even with partially observed data. A thorough review of related methods is provided in Section 4.2. Although modern tools have demonstrated their capacity, they have largely stayed within the bioinformatics community. The lack of explainability and the need for large computational resources often prohibit the adoption by independent research groups (Ma and Xu, 2022).

3.1.1 Bayesian Inference

Bayesian inference is concerned with methods that make inference about observed and unobserved data using probability models. This reduction from unknown to the notion of random is an essential characteristic of Bayesian methods for quantifying uncertainty around inference and predictions (Robert, 2007).

According to Gelman et al. (2013) Bayesian inference could be idealized in three steps. Initially, available knowledge about the underlying problem and data collection procedures is utilized to reason about parameter interactions. Observable and unobservable parameters of a problem are summarized in a full joint probability model. Using observed data we can then adjust beliefs for unknown parameters. What is described here is an inversion process which aims to infer effects from causes by considering the probabilistic nature of the model and the unexplained factors, formalized using Bayes theorem (Robert, 2007). To mathematically formalize it, suppose we aim to model the behavior of a parameter z using a set of observations $\mathscr{D}$. The Bayes theorem that updates our beliefs about z using $\mathscr{D}$ is given by the following equation:

$$\mathbb{P}(z|\mathscr{D}) = \frac{\mathbb{P}(\mathscr{D}|z)\,\mathbb{P}(z)}{\mathbb{P}(\mathscr{D})} \tag{3.1}$$

Where $\mathbb{P}(z)$ is the prior which encodes some knowledge about z, $\mathbb{P}(\mathscr{D}|z)$ is the likelihood which models observed data generating mechanisms, $\mathbb{P}(z|\mathscr{D})$ is the posterior which expresses the adjusted beliefs for z based on the set of observations $\mathscr{D}$ and $\mathbb{P}(\mathscr{D})$ is the evidence which expresses the probability of observing $\mathscr{D}$ irrespective of the choice of z in the considered model. In the final step, we can evaluate a series of questions like how well the model fits the data, how sensitive are the results to changing assumptions in the first step or what are the implications for the problem we are trying to solve.

Likelihood

In Bayesian inference, the role of the likelihood is to quantify the strength of support the observed data lend to possible values for unknown parameters (van de Schoot et al., 2021). Suppose we have a set of N observations $y_i \in \mathscr{D}, i \in 1 \ldots N$. Likelihood is defined as the joint probability of observing $y_1 \ldots y_n$ given generating parameter z is denoted by $\mathbb{P}(y_1, \ldots, y_N; z)$. To simplify the joint distribution and make inference more scalable it is common to assume partial or complete independence. For example, in scMVI (Chapter 6) we assume that cell vectors $y_1 \ldots y_n$ are independent given an unknown latent state that we aim to infer. In Bayesian settings, observations in $\mathscr{D}$ are treated as fixed and latent parameter z is an unknown random quantity. Therefore, the likelihood can be understood as a function of z for fixed observations y_i. As data generating mechanisms are in many cases unknown, choosing a model that effectively approximates reality is not always straight forward. Often researchers resort to the exponential family [3] for its inference friendly properties. There are also cases where more complicated modeling assumptions are necessary. A good example is single cell RNA expression, where sparsity and variation cannot be explained by distributions in the exponential family (Svensson, 2020).

Prior

Prior selection is viewed as a crucial part of implementing Bayesian inference since it can substantially affect posterior. For that reason, they are considered a controversial aspect of Bayesian statistics. Priors should be considered as a tool with acceptable frequentist

[3] More info can be found on (Koopman, 1936). We are not necessarily advocating the use of exponential family. In fact, for one of our models we have used zero-inflated Poisson which is not in the exponential family. Any model used relies on a series of underlying assumptions with risk of model misspecification.

properties, or a way to summarize expert information with its surrounding uncertainty (Robert, 2007).

Informativeness of priors could be split in three categories. There are informative priors, which reflect researcher's certainty about parameter, weakly informative, which reflects lower degree of certainty and uninformative or diffuse priors, which take a completely agnostic stance, letting the likelihood to determine posterior uncertainty.

Choosing the right prior strictly depends on the problem. There are cases where biased results could be obtained with uninformative priors and examples where informative ones shift the posterior away from likelihood. In problems where few observations exist, weakly informative priors are preferred in Bayesian inference (van de Schoot et al., 2021).

3.1.2 Hierarchical Models

One criticism of the Bayesian paradigm, is that common priors are almost never rich enough to accurately reflects expert's information (Robert, 2007). Almost all priors commonly used in Bayesian inference have parameters controlling their shape and location. Hence, it is natural to ask how are these parameters set. In the Bayesian paradigm this could be answered by introducing hierarchies. Priors could be placed in parameters controlling priors. This choice decomposes the prior into several conditional levels, allowing for a rich enough prior to contain both structural and subjective items of information, improving robustness of resulting estimators (Robert, 2007). Going back to the example of equation 3.1, prior $\mathbb{P}(z)$ could have a parameter η controlling the amount of uncertainty about z. Hence it could be written as $\mathbb{P}(z|\eta)$ making the prior $\mathbb{P}(z|\eta)\,\mathbb{P}(\eta)$ and Bayes theorem :

$$\mathbb{P}(z,\eta|\mathscr{D}) = \frac{\mathbb{P}(\mathscr{D}|z)\,\mathbb{P}(z|\eta)\,\mathbb{P}(\eta)}{\mathbb{P}(\mathscr{D})} \tag{3.2}$$

In the context of omics data where there are many genes and observations are few, hierarchical models have attracted attention (Ji and Liu, 2010). Hierarchical models are very effective at describing data from heterogeneous sources and could be easily used to share information across genes to improve statistical predictions in missing data Argelaguet et al. (2020); Csárdi et al. (2015); Kapourani et al. (2021); Maniatis et al. (2022); Vallejos et al. (2015). The combination of hierarchical models with deep learning, a subfield of machine learning that involves the use of neural networks (Bishop, 2006) with multiple layers to extract complex features from data, has lead to modern methods like those proposed by Ashuach et al. (2021) and Zuo and Chen (2021). At the expense of explainability such approaches can model more complex biological patterns.

3.1.3 Model Selection

In various modeling problems, it is common to have alternative models trying to make predictions about the same data. Therefore, it is important to establish the foundations to formally compare different models. Unfortunately, building a strict decision theoretic framework where this could be achieved is impossible as it depends on the specific goals set by the researcher. For example, model selection could be used in very early steps to identify models that could better serve likelihoods (Robert, 2007).

Suppose we have N competing models models $M_1, \ldots, M_N$, with corresponding vector parameters $\mathbf{z}_1, \ldots, \mathbf{z}_N$ [4] and our task is to identify the model that better fits the data $\mathscr{D}$. The first step in the Bayesian setting, is to introduce a prior $\mathbb{P}(M_i)$ for each model reflecting our beliefs, the second step is to define the likelihood which will adjust our beliefs based on evidence, this is:

$$\mathbb{P}(\mathscr{D}|M_i) = \int \mathbb{P}(\mathscr{D}|\mathbf{z}_i, M_i)\,\mathbb{P}(\mathbf{z}_i|M_i)\,d\mathbf{z}_i \tag{3.3}$$

As z_i is marginalized rather than maximized, over-fitting is not an option [5]. Hence the posterior becomes:

$$\mathbb{P}(M_i|\mathscr{D}) = \frac{\mathbb{P}(\mathscr{D}|M_i)\,\mathbb{P}(M_i)}{\sum_{i=1}^{N}\mathbb{P}(\mathscr{D}|M_i)\,\mathbb{P}(M_i)} \tag{3.4}$$

Posterior $\mathbb{P}(M_i|\mathscr{D})$ could be interpreted as measure of how well model M_i fits the data compared to the competing models. Comparing alternative models with that approach is not always practical. The most obvious problem is posteriors' intractability, which precludes the integration in 3.3. To better understand this problem the reader is directed in section 2.1 of Blei et al. (2017), where the authors explain the issues with integrating a simple Gaussian mixture model. However, when considering the problem from a measure theoretic perspective, infinite (or very large) dimensional parameter spaces and model alternatives reveal the actual complications that could make exploration very challenging or impossible.

To simplify the problem alternatives are restricted to two, with no preference among them. The ratio of the models' likelihoods is known as Bayes factor (Kass and Raftery, 1995). Bayes factor does not solve the intractability of 3.3, hence popular heuristics like information criteria are used to approximate it [6]. Mathematically, information criteria yield solutions that balance between acceptable model complexity and goodness of fit (Clarke

[4] The dimensions of parameters z do not need to match.

[5] This implies that models with more parameters do not necessarily have higher marginal likelihood also known as Bayesian Occam's razor (Ghahramani, 2005).

[6] For more information as to how to estimate Bayes factor please look at (Michael and Timothy, 2000; Wei et al., 2022).

et al., 2009). In the Bayesian setting where posteriors are inferred with sampling, model selection is commonly done using Information Criteria like Deviance Information Criterion (DIC) and Watanabe Information Criterion (WAIC)(Spiegelhalter et al., 2002; Watanabe, 2010). DIC operates on goodness of fit versus complexity balance principle analogously to Akaike and Bayesian Information Criteria [7] (Spiegelhalter et al., 2014). For a set of posterior samples $\tilde{z}_1, \ldots, \tilde{z}_N$ on parameter z with average,

$$\bar{z} = \frac{1}{N} \sum_{n=1}^{N} \tilde{z}_n \tag{3.5}$$

and error model $\mathbb{P}(\mathscr{D}|z)$, DIC is expressed as follows:

$$DIC = D(\bar{z}) + 2p_D \tag{3.6}$$

where

$$D(z) = -2\log\left(\mathbb{P}(\mathscr{D}|z)\right)$$

and

$$p_D = \frac{1}{2}\mathbb{V}_{z|\mathscr{D}}\left[\log\left(\left[\mathbb{P}(\mathscr{D}|z)\right)\right]\right],$$

. where $D(z)$ is the Bayesian deviance[8] and p_D is the posterior mean deviance. In the limit of large sample size, the model with the lowest $D(z)$ will have the highest posterior probability. The posterior mean deviance p_D is used to measure the effective number of parameters of the model [9]. Hence, DIC can be thought as heuristic for choosing models that achieve high posterior probability with fewest possible parameters. Despite DIC being an important tool model selection heuristic, limitations outlined in Spiegelhalter et al. (2014) have lead to the introduction of novel information criteria. Perhaps the most promising is WAIC (Watanabe, 2010). Like DIC, WAIC estimates the effective number of parameters to adjust goodness of fit for over-fitting but it is often preferred as it resembles leave-one-out cross-validation (LOO-CV) (Gelman et al., 2013). For the same set of posterior samples $\tilde{z}_1, \ldots, \tilde{z}_N$ and error model $\mathbb{P}(\mathscr{D}|z)$, WAIC is expressed as follows:

$$WAIC = -2\sum_{j}\log\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}\left(\mathscr{D}_j|z_i\right)\right) + 2p_{WAIC} \tag{3.7}$$

.

[7] More information could be found in (Akaike, 1973) and (Schwarz, 1978).

[8] The Bayesian deviance has an important role in statistical comparison because of it connection to the Kullback-Leibler information measure(Gelman et al., 2004).

[9] More information can be found in (Gelman et al., 2004).

$$p_{WAIC} = \mathbb{V}_{z|\mathscr{D}}\left[\log\{\mathbb{P}\left(\mathscr{D}|z\right)\}\right],$$

.

WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate. This is especially relevant in a predictive context, as WAIC is evaluating the predictions that are actually being used for new data in a Bayesian context (Gelman et al., 2004).

In practice, to compare between models over a posterior sample set $\tilde{z}_1, \ldots, \tilde{z}_N$, we define $DIC_{M_1}/WAIC_{M_1}$ and $DIC_{M_2}/WAIC_{M_2}$ for M_1 and M_2 respectively. Then the model with the smallest $WAIC/DIC$ is chosen. In Maniatis et al. (2022) DIC is used to demonstrate that for most of genes in the dataset, the model with zero inflation better explains the expression data.

3.2 Multiple Hypothesis Testing

Multiple hypothesis testing is a common practice encountered when analysing high dimensional molecular data. To better motivate this section we are going to work with an example. Suppose we have N genomic regions, for which we would like to identify if they present any regulatory action between epigenomic and transcriptomic layers. A proxy for measuring regulation for a region i would be to estimate Pearson correlation (ρ_i) between epigenome and transcriptome. Then to identify if a correlation is statistically significant (i.e. there is indeed strong regulation), we perform hypothesis testing. On the one hand, we have null hypothesis H_0 saying that true correlation $|r_i| \leq \gamma$. On the other hand, there is alternative hypothesis H_a opposing H_0 (i.e. data support $|r_i| \geq \gamma$). To asses if H_a holds any merit, we evaluate the probability of observing strong correlation under H_0, namely the "p-value". If "p-value" is small (e.g. below 5%), then we reject H_0 for H_a. In the N regions, assume that for N_0 regions the null hypothesis holds true and in N_R the null hypothesis is rejected. Then we get the following table

	H_0 is true	H_a is true	Total
Test rejects H_0	$N_{0,R}$	$N_{a,R}$	N_R
Test does not reject H_a	$N_{0,A}$	$N_{a,A}$	N-N_R
Total	N_0	N-N_0	N

Table 3.1 Table summarizing Pearson correlation hypothesis testing for N genomic regions with $N_{0,R}$ denoting the number of regions for which H_0 is true but the test rejected H_0, $N_{0,A}$ denoting the number of regions for which H_0 is true and the test accepted H_0, $N_{a,R}$ denoting the number of regions for which H_a is true and the test rejected H_0 and $N_{a,A}$ denoting the number of regions for which H_a is true and the test accepted H_0.

In our example we are interested in regions that were falsely declared significant (i.e. H_0 but hypothesis testing indicates otherwise). Assuming test for each genomic region is independent with "p-value" p, the probability $\mathbb{P}(N_{0,R} > 1)$ is:

$$\mathbb{P}(N_{0,R} > 1) = 1 - (1-p)^N \tag{3.8}$$

Even if individual "p-values" for each test are set low, the probability of $N_{0,R} > 1$ is high for N test. To restrict the total number of false discoveries, adjustments for test multiplicity are necessary (Korthauer et al., 2019). For example, False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) was designed to adjust hypothesis testing "p-values" such that $\frac{N_{0,R}}{N}$ stays below a pre-specified threshold α.

In Bayesian paradigm, hypothesis testing for the outlined problem can be interpreted as a model comparison between $M_0 : \|z \leq \gamma\|$ and $M_1 \|z \geq \gamma\|$. Then posterior

$$p_{z_i} = \mathbb{P}(|z_i| \geq \gamma | \mathscr{D}) \tag{3.9}$$

can be viewed as the probability assigned to model M_1. When Bayesian hypothesis testing is performed, Expected False Discovery Rate (EFDR) (Newton et al., 2004) is used to suppress the prevalence of false positives. In our working example, suppose we have a Bayesian model quantifying posterior correlation z_i for model M_i, in genomic region i, given evidence $\mathscr{D}$. So EFDR for some evidence threshold α is

$$\text{EFDR}_\alpha = \frac{\sum_{i=1}^{N} (1 - p_{z_i}) \mathbb{I}(p_{z_i} > \alpha)}{\mathbb{I}(p_{z_i} > \alpha)} \tag{3.10}$$

where $\mathbb{I}(A) = 1$ if A is true and 0 otherwise [10].

[10] This the main idea behind SCRaPL (Maniatis et al., 2022), which we will encounter in Chapter 5.

In many practical applications (e.g. Kapourani et al., 2021; Maniatis et al., 2022; Vallejos et al., 2015). EFDR is used to tune parameters of probabilistic decision rules such that false positives are controlled.

3.3 Probabilistic graphical models

Modeling complex biological phenomena in the Bayesian paradigm, requires the definition of joint distributions that account for a series of non-trivial parameter interactions (Bishop, 2006). Probabilistic graphical models provide an intuitive machinery for encoding independence assumptions in the joint density enabling model motivation, computational efficiency in complex models and important insights to the user about variable dependencies (Bishop, 2006). In graphical representations, nodes denote random variables while edges direct probabilistic relationships (Koller and Friedman, 2009).

Depending the connections between nodes and their topological ordering, graphical representations form two families. In the first we find Markov Networks like the Boltzman Machine (Ackley et al., 1985). For this thesis we extensively use the second for modeling, namely Bayesian Networks. Bayesian Networks are acyclic graphs were nodes are connected in directed fashion (Bishop, 2006).

Theoretically inference in Bayesian Networks is possible algebraically for discrete variables via Conditional Probability Tables (CPTs) and chain rule (Murphy, 2012). However a CPT for N discrete random variables attaining two values would require $2^N - 1$ parameters which scales poorly. The solution for efficiently characterizing complex joint distributions while keeping them expressive is to partially simplify the topological ordering through Conditional Independence (CI). When modeling a problem, CI assumptions need to be carefully encoded in the model as there is not a "one to one" correspondence between topological ordering and CI rules. Carefully articulated graphical models respecting CI assumptions could help remove such ambiguities.

A popular way of detecting CI properties among variables in Bayesian Networks is D-separation. Consider three sets of non-intersecting nodes [11] W_1,W_2,W_3 in the directed graphical model of figure 3.1d. To deduce whether $W_1 \perp\!\!\!\perp W_2 | W_3$ is implied by the Bayesian Network, we consider all direct and indirect paths from nodes in W_1 to nodes in W_2. These paths are considered blocked if it includes a node such that either arrows meet:

1. head-to-head or tail-to-tail at the node in W_3, or

[11] The union of these sets is not prohibited to be smaller then the complete set of nodes in the graph (Bishop, 2006).

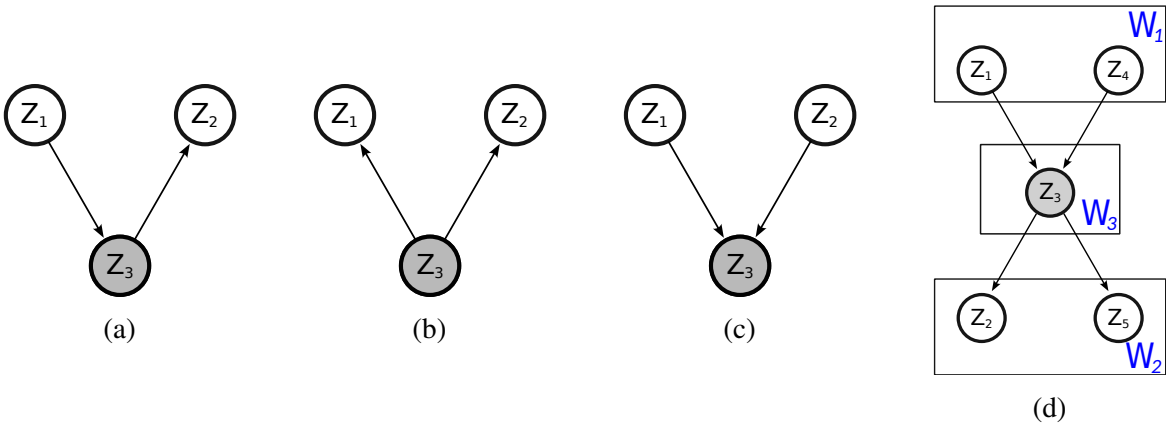


Fig. 3.1 **D-separation summary for variables and sets of variables.** In figures 3.1a and 3.1b paths from z_1 to z_2 are blocked by z_3 using d-separation rule (1.), hence $z_1 \perp\!\!\!\perp z_2 | z_3$. Since path from z_1 to z_2 is not blocked by z_3 using d-separation rule (2.), $z_1 \not\perp\!\!\!\perp z_2 | z_3$. (3.1d) The example extends D-separation to sets of variables. Drawing paths from members of W_1 to members of W_2 we deduce that $W_1 \perp\!\!\!\perp W_2 | W_3$.

2. head-to-head at the node, and neither the node, nor the any of its descendants are in Z_3

Since all paths are blocked, then W_1 is said to be D-separated from W_2 by W_3, and the joint distribution over all of the variables in the graph satisfies $W_1 \perp\!\!\!\perp W_2 | W_3$. In this example since W_3 consists of parents, co-parents and children nodes of W_2, rendering variables in W_2 conditionally independent of all other nodes in the Bayesian Network, we say that set W_3 is the Markov Blanket (MB) of W_2 ($\mathbf{MB}(W_2)$). The Markov Blanket is often useful in some inferential schemes requiring the evaluation of conditional distributions (e.g. Gibbs Sampling Gelfand and Smith, 1990).

When graphical models are used in practice two conventions are followed. If the model is applied over a set of independent identically distributed (iid.) observations, a box is drawn around repeated variables placing the number of copies in the bottom right corner (Murphy, 2012). Furthermore, observed variables are denoted with darker color. The former helps to avoid visual clutter while the latter distinguishes to be inferred parameters from the rest. Using these notations to compact information becomes extremely useful in large distributions. As we will see later, SCRaPL's scaling is achieved assuming independence across genomic regions and conditional independence across cells. In (Argelaguet et al., 2019), where there are 9500 genomic regions, graphical models can accurately encode these assumptions in a compact and digestible way.

3.4 Approximate Bayesian Inference

Bayesian inference has found applications from law and philosophy to engineering and science. After data are observed, beliefs around a parameter could be adjusted to obtain a posterior distribution. Contrary to popular perception, inference is concerned with more than finding the parameters that better explain a set of observations (MacKay, 2004). While identifying the most probable hypothesis might be important for some problems, quantifying the entire posterior distribution gives comparative advantage as it provides a measure of uncertainty, which can be particularly important in cases with few observations per parameter. Other problems might require tasks like variable marginalization or moment estimation. Estimating posterior distribution or some moment is often challenging (MacKay, 2004). This could be due to posterior's intractability or latent space's dimensionality (Bishop, 2006). To overcome that, we resort to approximate methods broadly falling into two classes depending on whether they rely on stochastic or deterministic approaches. In the class of stochastic methods we find sampling algorithms like Monte Carlo (MC), often adopted for their ease of implementation, scaling and theoretical guarantees. Class of deterministic methods include Variational Inference (VI) and other optimization based posterior approximation schemes which go beyond the scope of the thesis. In the next section we will consider VI in the context of Variational Autoencoders.

MC methods are a broad class of algorithms used to solve problems that have probabilistic interpretation. Despite their early use being traced back to 18th century, it was not until the $40s$ when a popular variation using Markov chains appeared (Hitchcock, 2003). Markov Chain Monte Carlo (MCMC) schemes rely on Markov chains to make exploration effective, offering much better scaling to conventional MC[12]. As MCMC schemes are constructed to satisfy the detailed balance equation, there are theoretical guarantees that their convergence is decoupled from "acceptable" initialization (Robert and Casella, 2004)[13]. Currently, there exist several flavors of MCMC whose main difference lies on the approach they take to propose new samples from the posterior distribution [14]. As many of the MCMC sampling algorithms do not take into account the geometry of the posterior's space when proposing samples, reaching new states and more importantly convergence is slow (Gelman et al., 2013). For continuous variables Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS) variation ameliorate this problem. In this thesis, we use them for effective posterior exploration, as SCRaPL's parameters live in high dimensional manifolds.

[12]A comprehensive introduction to fundamental notions of Markov chains and results required to establish MCMC convergence can be found in chapter 6 of Robert and Casella (2004).

[13]More on the detailed balance equation can be found in chapter 6 of (Robert and Casella, 2004).

[14]For a summary of MCMC flavors the reader is directed to chapters 7-11 of Robert and Casella (2004).

3.4.1 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) was introduced in 1987 (Duane et al., 1987) for the numerical simulation of lattice field theory. HMC utilizes techniques from differential geometry to generate transitions spanning the full marginal variance, circumventing the costly pilot runs required to tune proposal densities (Girolami and Calderhead, 2011) and eliminating the random walk behavior (Betancourt and Girolami, 2015).

An effective way of sampling from the target density $\mathbb{P}(z)$, presented in figure 3.2a, would be to construct a proposal that generates more samples from the orange region around the peak compared to the tails' gray region. Then accepting/rejecting step would make sure that each state is accepted in proportion to its height. To construct a proposal that achieves such properties the target density $\mathbb{P}(z)$ is log-transformed.

Using the physical analogy of a fictitious particle moving on friction-less terrain is an easy way to analyze its behavior. The fictitious particle moves in the terrain under the influence of gravity, propelled to random directions from time to time. The gravity encourages visiting high density regions while push to random direction ensures that low density regions are also surveyed. The combination of momentum built and gravity allows it to explore each state proportionally to the target density height (Lambert, 2017). State values should be carefully initialized as future proposed and accepted/rejected states depend on them. If initial parameter is close to the target density maximum and the momentum is negligible, effective exploration is impossible. In practice, HMC solves this problem by initializing a different momentum in each step. For a one dimensional normal distribution them HMC paths in the parameter-momentum space is summarized in figure 3.2b.

Analyzing the Hamiltonian of a particle moving at a friction-less terrain under the influence of gravity provides an effective way of translating the intuition behind HMC to equations. The particle has two types of energy, the potential (linked to location) and the kinetic (linked to momentum). As the particle moves through the terrain, potential and kinetic energies vary such that their sum remains constant. To formulate the Hamiltonian for the fictitious particle we start with a posterior density $\mathbb{P}(z)$ which will define the geometry of the friction-less terrain. Then we augment $\mathbb{P}(z)$ with an auxiliary variable momentum variable p creating the following joint density:

$$\mathbb{P}(z, p) = \mathbb{P}(p|z)\mathbb{P}(z) \tag{3.11}$$

The Hamiltonian of the system (equation 3.12) is derived by applying a negative log transformation to the joint density 3.11.

$$\mathbb{H}(z,p) = -\log[\mathbb{P}(z,p)] = -\log[\mathbb{P}(z)] - \log[\mathbb{P}(p|z)] = \mathbb{U}(z) + \mathbb{K}(p|z) \tag{3.12}$$

Where $\mathbb{U}(z)$ is the potential energy and $\mathbb{K}(p|z)$ the kinetic. By construction, $\mathbb{U}(z)$ is completely determined by $\mathbb{P}(z)$ and discourages particles from exploring low density areas, most of the time, by attaching a high potential value. On the other hand, kinetic energy must be specified during implementation and can take many forms (Betancourt, 2017). In many implementations, a multivariate standard normal that does not depend on particle's location is used. Equations governing particle's motion (equation 3.13) are derived from Hamiltonian (eq. 3.12) and Hamilton's equations (Betancourt, 2017). Equations 3.13 guarantee that in every state potential energy $\mathbb{U}(z)$ and kinetic energy $\mathbb{K}(p|z)$ change such that the total energy $\mathbb{H}(z,p)$ is conserved. This takes into account the space geometry of the posterior enabling for effective sampling.

$$\frac{dz}{dt} = \frac{\partial H}{\partial p}, \ \frac{dp}{dt} = -\frac{\partial H}{\partial z} \tag{3.13}$$

For most posteriors, Hamilton's equations of motion (equation 3.13) are sufficiently complicated to render analytic solutions infeasible; instead they are numerically integrated (Betancourt and Girolami, 2015). To minimize the accumulated error associated with long numerical trajectories, symplectic integrators are preferred.

Simplectic integrators preserve the linear symplectic structure inherent in the phase space representation of the Hamiltonian dynamics (Donnelly and Rogers, 2005), making them ideal for simulating large numerical trajectories [15]. To correct the bias introduced by the numerical approximation, endpoints of Hamiltonian trajectories are treated as proposals (Betancourt, 2017). These proposals are accepted with some probability stemming from the detailed balance equation. The most popular integrator used to solve HMC proposal paths is the leap frog algorithm (Bishop, 2006). Starting from location $z(0)$ with momentum $p(0)$ and running the leap frog algorithm for S steps with step size ε we end up in location $z(S\varepsilon)$ with momentum $p(S\varepsilon)$. To satisfy the detailed balance equation, the new state is accepted with probability

$$a = \min\left[1, \frac{e^{-\mathbb{H}(z(S\varepsilon), -p(S\varepsilon))}}{e^{-\mathbb{H}(z(0), p(0))}}\right] \tag{3.14}$$

[15] More information on simplectic integrators can be found in (Donnelly and Rogers, 2005).

[16]

The effectiveness of each HMC trajectory, exploring different energy levels, strongly depends on integration times across phase space of $\mathbb{H}(z, p)$. Integration over a short time span, could undermine the posterior space exploration. On the other hand, integrating for too long would probably push the particle to already explored regions wasting computational resources (Betancourt, 2017).

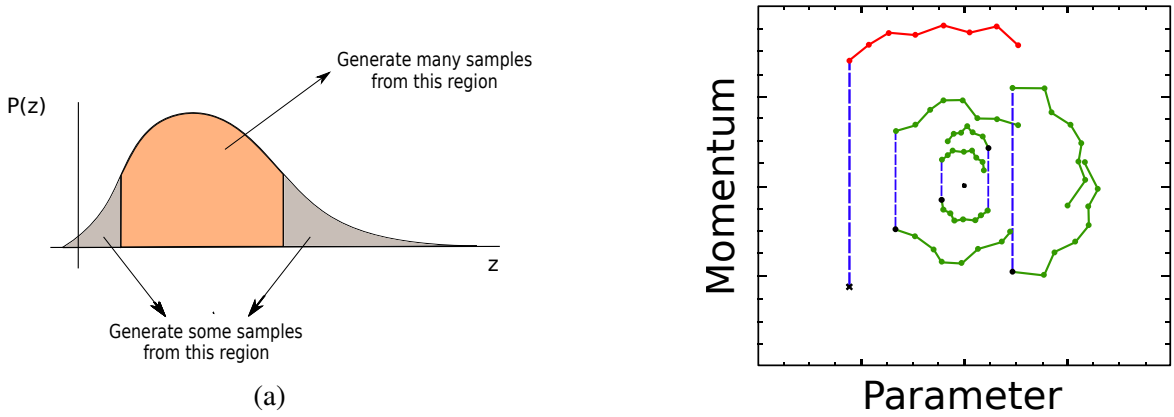
Dynamically determining optimal number of integration steps (S) and step size (ε) is a highly non-trivial task. Often practitioners rely on heuristics like auto-correlation or acceptance rate to adjust step-size and length (Hoffman and Gelman, 2014) with optimal acceptance rate being around 0.65 (Beskos et al., 2013). NUTS proposal step circumvents that problem using a tree of Hamiltonian paths obtained by forwardly and backwardly integrating with respect to time (Hoffman and Gelman, 2014). While the tree is expanding, the algorithm keeps track of the distance between discovered states and the original one. As long as the new states increase the distance to the original state the tree expands. At the end of the tree construction, one of the states is randomly chosen. Hence, NUTS can work with any reasonable step size and automatically detect number of steps to integrate without tuning. However, this is achieved at the cost of computationally expensive trees.

3.4.2 Assessing the quality of MCMC outputs

Irrespective of the MCMC variation, convergence could be split in three phases. In the first phase, the Markov chain moves from initialized location to the equilibrium distribution. Then it starts exploring the equilibrium distribution focusing more on high density regions. In this phase, estimators improve as biases from previous phase are eliminated. In the final phase, it refines the exploration which improves estimators precision (Betancourt, 2017). When exploration of the target density reaches the final phase, samples from previous explorations are discarded (typically known as burn-in).

Assessing the phase MCMC has reached is challenging as only asymptotic guarantees exist. After a finite number of iterations, the distribution used to draw samples lies between the starting and the target distributions (Gelman and Rubin, 1992). Hence, practitioners use a series of heuristics and plots to assess convergence. Sometimes the Markov chain can remain for a substantial amount of time in regions heavily influenced by initial distribution

[16] Keen-eyed readers will notice that in the proposed state, momentum is negated. This is done to deal with the asymmetry in proposal distributions. Since the leap frog is deterministic, the particle will almost surely (i.e. with probability 1) reach the proposed state. If the momentum is not negated, almost surely the particle will not return in the original position after S steps with step size ε, rejecting every sample (i.e. acceptance probability $a = 0$). If, however, the particle starts from the proposed location with negated momentum, then it is guaranteed to reach initial state.



(a) (b)

Fig. 3.2 **HMC explanatory figures** (3.2a) For a target density $P(z)$ an effective proposal should generate many samples from the orange area and some from gray areas. (3.2b) Sketch of HMC accepted (green) or rejected (red) trajectories in the momentum-parameter space for a 1D Gaussian distribution. At the end of each trajectory momentum is negated, hence the vertical jump. If state is accepted, a new momentum value is sampled form a normal distribution and the leapfrog simulates the numerical trajectories while conserving the total energy.

(Gelman and Rubin, 1992). Sampling methods like HMC suffer less from that behavior as it takes advantage of the space geometry, but convergence assessment is still important. There various of ways assessing convergence [17]. Most popular strategies estimate some statistic across chains initialized from various locations. To asses convergence in this thesis we rely on Gelman-Rubin (GR) (Gelman and Rubin, 1992) criterion. GR compares the variance between the means of parallel chains with the within chain variance to estimate the factor by which the target variance might shrink if sampling where to continue (Cowles and Carlin, 1996). For N chains the GR criterion $\hat{R}$ is estimated as follows:

$$\hat{R} = \frac{\hat{V}}{W}\frac{df}{df-2} \tag{3.15}$$

where W is the average within sequence variance, $\hat{V}$ is the across variance chains and $df = \frac{2\hat{V}}{\text{var}(\hat{V})}$, with var $(\hat{V})$ being the variance of $\hat{V}$ [18].

Convergence does not always guarantees that collected samples are enough to effectively estimate a parameter of interest. In the extreme scenario where all collected samples are perfectly correlated, we are effectively using only one sample in the estimation. Hence, to understand how many samples we are effectively using during estimation we have to look at

[17] A comparative review can be found in (Cowles and Carlin, 1996).
[18] For a step by step explanation on how to estimate $\hat{R}$ please refer to (Gelman and Rubin, 1992).

the dependence between parts of the chain. This gives rise to the notion of effective sample size (ESS) (Liu, 2004), where the number of collected samples after burn-in is adjusted to account for auto-correlation. For N samples of a variable z obtained with MCMC, ESS is estimated as follows:

$$\text{ESS} = \frac{N}{1 + 2\sum_{i=1}^{\infty} \rho_i(z)} \tag{3.16}$$

where ρ_i is the i step auto-correlation lag for samples of z. If ESS is small sampling might continue until an adequate effective size is reached.

Algorithm 1 No-U-Turn Sampler (Hoffman and Gelman, 2014)

Input: Given initial state $\bar{z}_0$, step size ε, Hamiltonian $\mathbb{H}$, number of samples N and a non-negative Δ_{max}.

Output: N samples from posterior $\mathbb{P}(z)$

for $ii = 1, 2, \ldots, N$ **do**
  Sample $p_0 \sim \mathbb{N}(\mathbf{0}, \mathbb{I})$.
  Sample $u \sim \mathrm{U}([0, \exp(\mathbb{H}(\bar{z}_{ii-1}, p_0))])$
  Set $\bar{z}^+ = \bar{z}_{ii-1}$, $\bar{z}^- = \bar{z}_{ii-1}$, $p^+ = p_0$, $p^- = p_0$, $C = \{(\bar{z}_{ii-1}, p_0)\}$, $s = 1$, $j = 1$
  while $s = 1$ **do**
    Sample integration direction $v_j \sim \mathrm{U}(\{-1, 1\})$.
    if $v_j = -1$ **then**
      $\bar{z}^-, p^-, -, -, C', s' \leftarrow \text{BUILDTREE}\left(\bar{z}^-, p^-, u, v_j, j, \varepsilon\right)$.
    else
      $-, -, \bar{z}^+, p^+, C', s' \leftarrow \text{BUILDTREE}\left(\bar{z}^+, p^+, u, v_j, j, \varepsilon\right)$.
    if $s' = 1$ **then**
      $C \rightarrow C \cup C'$.
    $s \leftarrow s' \mathbb{I}[(\bar{z}^+ - \bar{z}^-)\, p^- \geq 0]\, \mathbb{I}[(\bar{z}^+ - \bar{z}^-)\, p^+ \geq 0]$.
    $j \leftarrow j + 1$.
  Sample uniformly one of the pairs $(\bar{z}^{ii}, p)$ from C.
function BUILDTREE$(\bar{z}, p, u, v, j, \varepsilon)$
  if $j = 0$ **then**
    Base case—take one leapfrog step in the direction v.
    $\bar{z}', p' \leftarrow \text{LEAPFROG}(\bar{z}, p, v\varepsilon)$.
    $C' \leftarrow \begin{cases} \left\{\left(\bar{z}', p'\right)\right\} & \text{if } u \leq \exp(\mathbb{H}(\bar{z}_{ii-1}, p_0)), \\ \emptyset & \text{else.} \end{cases}$

$s^{'} \leftarrow \mathbb{I}\left[u < \exp\left(\Delta_{max} + \mathbb{H}\left(\bar{z}_{ii-1}, p_0\right)\right)\right].$
return $\bar{z}^{'}, p^{'}, C^{'}, s^{'}$
else
Recursion—build the left and right subtrees.
$\bar{z}^{-}, p^{-}, \bar{z}^{+}, p^{+}, C^{'}, s^{'} \leftarrow$ BUILDTREE $\left(\bar{z}, p, u, v_j, j-1, \varepsilon\right).$
if $v = -1$ **then**
$\bar{z}^{-}, p^{-}, -, -, C^{'}, s^{'} \leftarrow$ BUILDTREE $\left(\bar{z}^{-}, p^{-}, u, v_j, j-1, \varepsilon\right).$
else
$-, -, \bar{z}^{+}, p^{+}, C^{'}, s^{'} \leftarrow$ BUILDTREE $\left(\bar{z}^{+}, p^{+}, u, v_j, j-1, \varepsilon\right).$
$s^{'} \leftarrow s^{'} s^{''} \mathbb{I}\left[\left(\bar{z}^{+} - \bar{z}^{-}\right) p^{-} \geq 0\right] \mathbb{I}\left[\left(\bar{z}^{+} - \bar{z}^{-}\right) p^{+} \geq 0\right].$
$C^{'} \rightarrow C^{'} \cup C^{''}.$
return $\bar{z}^{+}, p^{+}, \bar{z}^{-}, p^{-}, C^{'}, s^{'}$
function LEAPFROG$(\bar{z}, p, \varepsilon)$
$\bar{p} \leftarrow p + \frac{\varepsilon}{2} \nabla_{\bar{z}} \mathbb{H}\left(\bar{z}, p\right).$
$z^{'} \leftarrow \bar{z} + \varepsilon \bar{p}.$
$\bar{p} \leftarrow \bar{p} + \frac{\varepsilon}{2} \nabla_{\bar{z}} \mathbb{H}\left(z^{'}, p\right).$

3.5 Function Approximation

The computational capabilities of modern equipment along with relatively recent storing improvements have lead to the popularization of deep learning posterior approximations based on VI. Deep learning VI models are often preferred for their scaling and flexibility to describe arbitrarily complex distributions.

VI is a class of deterministic approximation schemes inspired by calculus of variations (Jordan et al., 1999). VI fits an analytic approximation to the intractable posterior using a statistical "distance" called divergence. Divergence metrics are used to measure separation between probability distributions and in general are not symmetric. Let $\mathbb{Q}$ and $\mathbb{P}$ be two distributions in some probability space. Then $D\left(\mathbb{Q}, \mathbb{P}\right)$ is called a divergence if it satisfies three properties(Amari, 2016).

1. $D\left(\mathbb{Q}, \mathbb{P}\right) \geq 0$
2. $D\left(\mathbb{Q}, \mathbb{P}\right) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$
3. For every probability $\mathbb{P}$ and infinitesimal displacement $d\mathbb{P}$, $D\left(\mathbb{P}, d\mathbb{P}\right)$ is a positive-definite quadratic form.

The most popular divergence in the field of machine learning is the Kullback–Leibler (KL) and is computed as follows:

$$D_{KL}(\mathbb{Q},\mathbb{P}) = \int \mathbb{Q} \ln\left(\frac{\mathbb{Q}}{\mathbb{P}}\right) d\mathbb{Q} \tag{3.17}$$

If $\mathbb{P}$ is an intractable distribution which we wish to approximate with an analytic distribution $\mathbb{Q}$, inference boils down to optimizing $\mathbb{Q}$'s parameters such that D_{KL} is minimized. The reduction of inference to optimization means that VI can achieve statistical benefits of the Bayesian approach at the speed of optimization (Murphy, 2012). These methods are increasingly popular in the machine learning community, due to their efficiency and scalability to large applications (Blei et al., 2003). However, the strong structural assumptions made about posterior's shape along with the limited performance evaluations tools might lead to low performing approximations (Yao et al., 2018). In this thesis we focus on a subset of VI aimed at approximating intractable posterior with a parametric model in Variational Autoencoders.

3.5.1 Variational Autoencoders

Variational Autoencoders (VAEs) combine hierarchical models, deep learning and VI to create flexible models that can describe complex phenomena. VAEs, were introduced in 2014 (Kingma and Welling, 2014) as an unsupervised method mapping data distribution to a different multivariate distribution. The multivariate distribution is often low dimensional, simpler in form and the mapping is done using artificial neural networks. VAEs resemble the Helmholtz Machines' (Dayan et al., 1995) structure, but changes in their training objective have allowed for seamless integration with modern optimization tools (Kingma and Welling, 2019).

VAEs combine the hidden hierarchies found in latent variables models with deep neural networks to model arbitrarily complex variable dependencies. To better motivate this point we are going to work with an example. Suppose we have single cell RNA data of Memory B cells and we would like to understand how they are generated. This task involves modeling the interactions of numerous genomic regions and biological processes not fully observed or understood. VAEs circumvent these problems by assuming a generative process and an expressive enough map to capture arbitrarily complex intricacies. Its structure incorporates two independently parameterized functions. One which approximates the generative process from the unknown to the observed state and is known as decoder or generative model. The second one reverses the generative process to the unknown state and is known as encoder. The unknown state is also known as latent space and it is fully described by a posterior

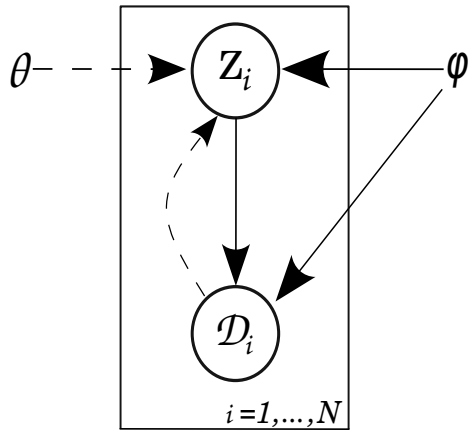


Fig. 3.3 **Variational Auto Encoder graphical model.** Combination of hierarchical graphical models and deep learning. Solid lines denote the generative model $\mathbb{P}_\phi(z, \mathscr{D})$, dashed lines the variational approximation $\mathbb{Q}_\theta(z|\mathscr{D})$ to the intractable posterior $\mathbb{P}_\phi(z|\mathscr{D})$. Variational parameters ϕ, θ are jointly learned during training. Figure and caption taken from Kingma and Welling (2014).

distribution conditioned on gene expression. The encoder delivers an approximation of the unknown state and generative model adapts to better capture underlying patters given its input and observations. In many cases, this procedure forces the encoder to create compact yet meaningful latent representations and the generative model to accurately simulate data generating mechanisms (Kingma and Welling, 2019).

To mathematically formalize this phenomenon we assume a set of observed variables $\mathscr{D}$ generated from an unknown generating process whose distribution is well approximated by $\mathbb{P}_\phi(\mathscr{D})$ with parameters ϕ. To fully leverage hierarchical models, density $\mathbb{P}_\phi(\mathscr{D})$ must be extended by a latent variable z to create a joint representation $\mathbb{P}_\phi(\mathscr{D}, z)$ of the observed and unknown spaces. By construction latent variables are not part of the observed data (Kingma and Welling, 2019). In this formulation, getting the generative model or its reverse ($\mathbb{P}_\phi(\mathscr{D}|z)$ and $\mathbb{P}_\phi(z|\mathscr{D})$ respectively) becomes a matter of applying Bayes rule.

The latent variable model described above is expressive enough to approximate complicated distributions $\mathbb{P}_\phi(\mathscr{D})$. However, this flexibility comes at the cost of intractable $\mathbb{P}_\phi(\mathscr{D})$. Using Bayes rule we can write the conditional density as:

$$\mathbb{P}_\phi(z|\mathscr{D}) = \frac{\mathbb{P}_\phi(z, \mathscr{D})}{\mathbb{P}_\phi(\mathscr{D})} \tag{3.18}$$

As the denominator contains the intractable marginal density $\mathbb{P}_\phi(\mathscr{D})$, the conditional $\mathbb{P}_\phi(z|\mathscr{D})$ is also intractable. Hence, inference in such models is hard (Blei et al., 2017).

To turn the intractable posterior and learning problems into tractable, we introduce a parametric model $\mathbb{Q}_\theta(z|\mathscr{D})$ whose aim is to approximate $\mathbb{P}_\phi(z|\mathscr{D})$. This model is also called encoder and our task is to tune variational parameters θ such that

$$\mathbb{Q}_\theta(z|\mathscr{D}) \approx \mathbb{P}_\phi(z|\mathscr{D}). \tag{3.19}$$

We further assume that

$$\mathbb{Q}_\theta(z|\mathscr{D}) = \mathrm{N}(z;\mu,\mathrm{diag}(\sigma)) \tag{3.20}$$

and parameters μ, $\ln[\sigma]$ are fully characterized by a non-linear functions

$$\mu = f_\theta^\mu(\mathscr{D}), \sigma = \exp(f_\theta^\sigma(\mathscr{D})) \tag{3.21}$$

where f_θ^μ and f_θ^σ share all the hidden layers apart from the final linear layer [19]. To make a more effective use of data, variational parameters θ are shared across samples. This strategy allows the use of gradient optimization schemes avoiding the costly per data point optimization and goes by the name amortized inference (Gershman and Goodman, 2014; Kingma and Welling, 2019).

At the end of training, we are looking to have a model that maximizes the log-likelihood $\ln\left[\mathbb{P}_\phi(\mathscr{D})\right]$ while accurately approximating $\mathbb{P}_\phi(z|\mathscr{D})$. The optimization objective that tries to achieves that is known as Evidence Lower Bound (ELBO) and is summarized below:

$$\mathscr{L}_{\phi,\theta}(\mathscr{D}) = \ln\left[\mathbb{P}_\phi(\mathscr{D})\right] - D_{KL}\left[\mathbb{Q}_\theta(z|\mathscr{D}) \parallel \mathbb{P}_\phi(z|\mathscr{D})\right] \tag{3.22}$$

Since $D_{KL}[\mathbb{Q}(z|\mathscr{D}) \parallel \mathbb{P}(z)] \geq 0$ the ELBO is a lower bound to log-likelihood of the data. Hence by optimizing the ELBO, we are improving the accuracy of the generative process as we are approximately maximizing the marginal log-likelihood $\ln\left[\mathbb{P}_\phi(\mathscr{D})\right]$.

Equation 3.22 is not possible to optimize as it requires knowledge of distributions $\mathbb{P}_\phi(\mathscr{D})$ and $\mathbb{P}_\phi(z|\mathscr{D})$, which are intractable. However, using Bayes rule and some algebraic manipulation we obtain equation 3.23.

[19] Modeling $\mathbb{Q}_\theta(z|\mathscr{D})$ with a conditional isotropic normal is one of the many choices made to remove intractability. Later it will become evident that normal distribution is often selected because the stochastic component (ie. multivariate normal number generator) is independent from the deterministic ones (ie. mean and standard deviation), with the reparametrization trick. However nothing stops us from picking alternative models. In fact there have been numerous extensions that go beyond the scope of the thesis. Some of them are described in chapter 3 of (Kingma and Welling, 2019).

$$\underbrace{\mathscr{L}_{\phi,\theta}(\mathscr{D})}_{\text{ELBO}} = -\underbrace{D_{KL}\left[\mathbb{Q}_\theta(z|\mathscr{D}) \parallel \mathbb{P}_\phi(z)\right]}_{\text{Regularizes Latent Space}} + \underbrace{\mathbb{E}_{\mathbb{Q}_\theta(z|\mathscr{D})}\left[\ln\left[\mathbb{P}_\phi(\mathscr{D}|z)\right]\right]}_{\text{Maximizes Conditional Log-Likelihood}} \tag{3.23}$$

Equation 3.23 is much easier to tune and requires less assumptions. Given that we have modeled $\mathbb{Q}_\theta(z|\mathscr{D})$ as a conditional Gaussian distribution, we set the prior $\mathbb{P}_\phi(z)$ equal to an isotropic normal with unit diagonal variance [20]. At its optimum, the ELBO in equation 3.23 strikes a balance between an approximate posterior $\mathbb{Q}_\theta(z|\mathscr{D})$ close to prior $\mathbb{P}_\phi(z)$ and a mapping producing parameters that maximize the conditional log-likelihood $\ln\left[\mathbb{P}_\phi(\mathscr{D}|z)\right]$.

The ELBO formulation presented in equation 3.23 allows for joint optimization with respect to parameters θ and ϕ using Stochastic Gradient Descent (SGD) (Kingma and Welling, 2014). Hence randomly initialized parameters θ and ϕ are stochastically updated until convergence.

Given iid. samples $\mathscr{D}_i \in \mathscr{D}, i = 1, \ldots, N$, the ELBO is the sum of ELBOs computed for individual data points.

$$\mathscr{L}_{\phi,\theta}(\mathscr{D}) = \sum_{i=1}^{N} \mathscr{L}_{\phi,\theta}(\mathscr{D}_i) \tag{3.24}$$

In large datasets we can randomly sample indices from $1, \ldots, N$ and get an unbiased estimate of $\mathscr{L}_{\phi,\theta}(\mathscr{D})$. Individual ELBO estimates and the respective gradients are in general intractable. However, unbiased estimates $\nabla_{\phi,\theta}\mathscr{L}_{\phi,\theta}(\mathscr{D}_i)$ exist, making minibatch SGD possible.

When attempting to compute gradients it quickly becomes evident that differentiating with respect to θ is challenging, since equation 3.23 involves and expectation with respect to $\mathbb{Q}_\theta(z|\mathscr{D}_i)$. This makes swapping gradient with expectation impossible, prohibiting cheap Monte Carlo gradient estimates. Under continuous latent variables and differentiable function approximations, we can differentiate ELBO with respect to both ϕ and θ through reparameterization trick (Kingma and Welling, 2014). By writing latent samples $z \sim \mathbb{Q}_\theta(z|\mathscr{D}_i)$ as

$$z = \mu + \sigma\varepsilon \text{ where } \varepsilon \sim \mathrm{N}(0, I) \tag{3.25}$$

the expectation in equation 3.23 can be rewritten as:

$$\mathscr{L}_{\phi,\theta} = -D_{KL}\left[\mathbb{Q}_\theta(z|\mathscr{D}) \parallel \mathbb{P}_\phi(z)\right] + \mathbb{E}_{\varepsilon\sim\mathrm{N}(0,I)}\left[\ln\left[\mathbb{P}_\phi(\mathscr{D}|z)\right]\right] \tag{3.26}$$

[20] Giving $\mathbb{P}_\phi(z)$ this form and shape is done to simplify computations. We could chose a different prior at the expense of closed form solution. Furthermore, we chose prior parameters to be as simple as possible because any choice will be absorbed by the generative model.

Hence Monte Carlo estimates for ϕ and θ are given by:

$$\nabla_{\phi,\theta}\mathscr{L}_{\phi,\theta}(\mathscr{D}) = -\nabla_{\phi,\theta}D_{KL}\left[\mathbb{Q}_\theta(z|\mathscr{D}) \parallel \mathbb{P}_\phi(z)\right] + \nabla_{\phi,\theta}\mathbb{E}_{\varepsilon\sim\mathrm{N}(0,I)}\left[\ln\left[\mathbb{P}_\phi(\mathscr{D}|z)\right]\right] \quad (3.27)$$

which simplify to:

$$\nabla_\phi\mathscr{L}_{\phi,\theta}(\mathscr{D}) = \nabla_\phi \ln\left[\mathbb{P}_\phi(\mathscr{D},z)\right] \quad (3.28)$$

$$\nabla_\theta\mathscr{L}_{\phi,\theta}(\mathscr{D}) = -\nabla_\theta \ln\left[\mathbb{Q}_\theta(z|\mathscr{D})\right] \quad (3.29)$$

Challenges

VAEs have provided a very flexible framework for quantifying complex phenomena. This flexibility, often comes at various costs. In this subsection we mainly focused on intractability and how to overcome it. However, challenges do not stop there. Independent researchers (Bowman et al., 2016; Sønderby et al., 2016) have reported that stochastic optimization with unmodified ELBO can get trapped to undesirable local minima [21].

Furthermore, pairs (D,z) which are likely by $\mathbb{Q}_\theta(z|\mathscr{D})$ but not $\mathbb{P}_\phi(\mathscr{D})$ are prohibited by term $\mathbb{E}_{\varepsilon\sim\mathrm{N}(0,I)}\left[\ln\left[\mathbb{P}_\phi(\mathscr{D}|z)\right]\right]$ in the ELBO. However, the reverse is not true essentially allowing unknown states z with little support over $\mathbb{Q}_\theta(z|\mathscr{D})$ (Kingma and Welling, 2019). This is apparent in cases where matching $\mathbb{Q}_\theta(z|\mathscr{D})$ with $\mathbb{P}_\phi(z|\mathscr{D})$ is not possible, so $\mathbb{Q}_\theta(z|\mathscr{D})$'s variance is reduced. In VAEs trained on images this is encountered as blurriness in the generated data.

Interpretability of VAEs and deep neural networks (DNNs) in general is essential, since it enables users to understand the overall strengths and weaknesses of the models and how to diagnose and correct potential problems (Dong et al., 2017). However, it is challenging to reason about what a VAE actually learns due to its opaque or black-box nature. Techniques like maximizing activation (Bengio et al., 2009), or Deep Taylor expansion (Montavon et al., 2017) can help improve interpretability by shedding light on the signals picked up by the network. However, VAE and more broadly deep neural network interpretability is still an active are of research.

Lack of large datasets describing to extract useful biological patterns. Many of the available multi-omics datasets are published as proof of concept. This means they rely on a relatively small number of cells and the underlying biology tends to be simple. We touched upon that issue in Chapter 2 There are exceptions where they capture interesting aspects of early developmental stages like (Argelaguet et al., 2019) but cells count is low. As VAEs tend to have parameters in the order of hundreds of thousands or million, the ratio of observations

[21] More information about the problems and suggested solutions can be found in (Kingma and Welling, 2019).

to number of parameters is incredibly low. This increases the likelihood of optimization converging into a local minimum or the model over-fitting the data. Furthermore, there are very few single cell multi-omics datasets which means that biologically interesting patters extracted from one dataset are easily verified to others.

In the context of single cell multi-omics, parameter tuning is a non-trivial yet aspect of developing accurate and robust VAEs. It often requires trying various combinations in parameter space. A careful selection of model parameters can substantially improve model performance. (Hu and Greene, 2019) that methods like UMAP and ZIFA can outperform poorly calibrated VAEs while taking a fraction of the time to train. They also found substantial performance differences with hyper-parameter tuning. Even though this is not entirely unexpected, the sensitivity of this class of methods under various parameter settings is not widely reported in the literature.

Chapter 4

Computational methods for single cell multi-omics data

4.1 Introduction

Single cell multi-omics technologies have emerged as a powerful approach to scrutinize the basic entities of life at a cellular resolution. By enabling the systematic interrogation of complex relationships between various omic layers at the single cell level, they have profoundly impacted several fields such as developmental biology, immunology and tumor development (Colomé-Tatché and Theis, 2018; Linnarsson and Teichmann, 2016). Among the most fundamental questions that could be solved by these data is how molecular states lead to different cell fates or disease. Key steps in answering such question would require to address a series of open computational challenges such as, devising efficient methods for integrating and aligning cells or introducing algorithms for robust and accurate downstream analysis.

Single cell multi-omics datasets can be created by jointly measuring different molecular layers, for example using scNMT-seq (Clark et al., 2018) (see Section 2.3), or by measuring single omic layers in different groups of cells. The former are known as matched while the latter as unmatched datasets (Adossa et al., 2021; Macaulay et al., 2017). Each approach has its own merits and limitations. For example, unmatched datasets generally require additional preprocessing steps, as different molecular layers originate from different cells and potentially different experimental setups Adossa et al. (2021); Ma et al. (2020a). Generating unmatched data on the other hand, has been a cheaper and more scalable solution compared to matched data. Depending on the data generation method different techniques have been developed.

The integration methods for single cell multi-omics data could be further divided into early, intermediate and late. In early integration, multiple molecular layers are concatenated into a single data matrix Adossa et al. (2021). This allows the use of any method designed to work with feature-cell matrix, simplifying further downstream analysis. However, this comes at an increased complexity cost requiring automatic feature learning approaches like dimensionality reduction Adossa et al. (2021). Moreover, each modality often presents unique sparsity and technical variability patterns which can affect the weight the algorithm places in each feature during early integration. Intermediate data integration typically includes a range of techniques to jointly analyse single cell multi-omics datasets. This is done by merging datasets into a single matrix using similarity-based integration, joint dimensionality reduction and statistical modelling approaches (Adossa et al., 2021). For late integration strategy, a consensus result is achieved by integrating results independently inferred from separately analysing each molecular layer. Here we find many methods that use clustering outputs usually generated from low-dimensional representations of single-omics data.

Downstream analysis is a critical step in extracting biological insights proceeding integration. In single cell multi-omics data it encompasses a wide range of techniques like differential analysis, identifying regions with strong regulatory action along with cell clustering and annotation. These methods enable the identification of cell disease-associated pathways, regulatory mechanisms, and subpopulations among others. Depending the task molecular layers are independently or jointly analysed. Differential analysis is a method that has been applied to single omics data. To interrogate the interactions between different molecular layers, methods rely on direct correlation estimates or modelling regulatory processes. Hence they rely on multi-omics data. Clustering and cell annotation have allowed to cluster cells and biologically relevant entities(Yang et al., 2017) fueled by the creation of novel atlas projects like (Davie et al., 2018; Han et al., 2018). Despite being initially applied to transcription, it has been applied to other molecular layers and even more recently to multi-omics data (Eltager et al., 2022; Yuan et al., 2022).

In this chapter, we provide an overview of the computational approaches related to integration and downstream analysis used in this thesis. More precisely we look at methods addressing intermediate integration also focusing our attention on methods that could handle partially observed matched data. For downstream analysis, emphasis is placed on methods that investigate regulatory patterns, especially the one relate to epigenome and transcriptome interactions. Furthermore, we look into different methods for performing differential analysis. By providing a comprehensive guide for readers interested in tools designed for researching molecular interactions between epignenome and transcriptome, we hope to enable them to gain deeper insights into computational tools for single-cell multi-omics.

4.2 Integration

Accompanying the thrust of technological advancement, an increasing body of computational methods to handle multi-omics data integration and alignment have been proposed [1]. In this section, we thoroughly consider some of the most popular dimensionality reduction-based approaches, for matched and unmatched data, giving short method descriptions and from time to time demonstrating examples where they have been used. More precisely, we look at Canonical correlation analysis, Matrix factorization-based, Manifold alignment, Network and VAE based integration methods. Furthermore we briefly consider similarity-based methods and statistical modelling techniques.

Canonical correlation analysis (CCA) is a correlation-based multivariate analysis method to examine the linear relationship between two unmatched datasets (Braak, 1986). CCA aims to find a latent representation that maximizes the covariation between two data modalities, thus neglecting variation that is not shared by the data modalities. Seurat3 (Cao et al., 2017) implements CCA in order to integrate two single-cell omics datasets (Adossa et al., 2021). It first jointly reduces the dimensionality of two datasets using CCA, followed by a search for a mutual nearest neighbor in lower dimensional space, and then establishes the cellular relationship across the datasets as an anchor. This has also been used, to integrate scRNA-seq and scATAC-seq data from the mouse visual cortex and scRNA-seq and surface protein expression from bone marrow (Cao et al., 2017). A relatively recent adaptation of CCA for matched single cell multi-omics clustering is bindSC (Dou et al., 2020). BindSC utilizes bi-order canonical correlation analysis (bi-CCA) that captures the correlated variables from both cells and features between two modalities to formulate the canonical correlation vectors in a latent space (Adossa et al., 2021).

Matrix factorization-based methods aim to describe each cell as the product between a vector that describes each omics element (genes, epigenetic loci, and proteins) and a vector of reduced and common features (factors) capturing its basic properties (Stanojevic et al., 2022). In single cell multi-omics field where raw readings are positive Non-negative matrix factorization (NMF) techniques are used. For unmatched single cell multi-omics data Linked Inference of Genomic Experimental Relationships (LIGER)(Welch et al., 2018) is a tool based on integrative NMF (iNMF) designed to identify the shared and dataset specific factors across datasets (Adossa et al., 2021). iNMF has been applied to epigenomic and transcriptomic data from mouse cortical (Welch et al., 2018) and human frontal cortex tissue (Luo et al., 2019). To perform integration, data have to be summarized such that the samples

[1] For more information regarding integration, reviews Adossa et al. (2021); Argelaguet et al. (2021); Colomé-Tatché and Theis (2018); Forcato et al. (2021); Ma et al. (2020a); Stanojevic et al. (2022) provide a good understanding of the filed

unambiguously match between assays. This can be achieved by aggregating cells into cell types that are confidently identified in both assays (Argelaguet et al., 2021). However, this strategy makes strong assumptions on the definition of cell types, and single-cell resolution is in part lost (Argelaguet et al., 2021). Furthermore, this strategy relies on fragile biological assumptions that can fail in scenarios where when the relationship between the molecular layers is complex. A good example is early embryonic development, where gene body DNA methylation or chromatin accessibility are not good predictors of gene expression (Argelaguet et al., 2021). For matched data methods like MOFA have been developed for integration. MOFA can be regarded as a generalization of CCA that builds on the Bayesian group factor analysis framework (Granja et al., 2019) to handle an arbitrary number of data modalities. By using structured sparsity priors, MOFA is capable of detecting sources of covariation between different sets of data modalities as well as sources of variation that are only present in a single data modality (Argelaguet et al., 2021). In a dataset of mouse post-implantation embryos where scNMT-seq was used to simultaneously profile RNA expression, DNA methylation and chromatin accessibility, an integrative analysis using MOFA revealed the existence of lineage-specific enhancers that are associated with germ layer commitment (Argelaguet et al., 2019, 2021).

Manifold alignment is a class of machine learning algorithms that aim to infer a lower-dimensional structure between multiple complex datasets (Jang et al., 2003). The idea is to create a low-dimensional representation for each dataset and then align these representations in a common space by matching points across datasets (Adossa et al., 2021). Perhaps the most known unsupervised learning tool that has been applied to matched and unmatched data is MATCHER (Welch et al., 2017). MATCHER constructs pseudotime trajectories, capturing the biological process, using Gaussian process and aligns them onto a reference line so that different cells can be matched across omics (Stanojevic et al., 2022). A potential limitation is the strong underlying model assumption there is only one common biological process modelled by a non-branching pseudotime trajectory. Another unsupervised method is Maximum mean discrepancy-based manifold alignment (MMD-MA) (Liu et al., 2019). The alignment is performed by matching low-dimensional representations of different omics, constructed through a kernel-based technique that minimizes the maximum mean discrepancy(MMD) between the two datasets (Stanojevic et al., 2022). Other tools that have been used for single cell multi-omics analysis are; Manifold-Aligning GAN (MAGAN) (Amodio and Krishnaswamy, 2018), a generative adversarial network (GAN) based manifold alignment tool and UNION-Com (Cao et al., 2020), a distance based matching method based on generalized unsupervised manifold alignment (Cui et al., 2014).

Network-based approaches seek to represent cellular dependency with the help of a weighted graph. This approach pools information between neighbors, leading to additional robustness against the noise. Among the most popular tools in this category is Seurat's (Hao et al., 2021) weighted-nearest neighbor(WNN) graph-based integration which assumes that cells which are similar according to the consensus are connected. Seurat' integration algorithm was applied to a CITE-seq-based transcriptomic and proteomic dataset of blood cells to improve the discovery of cell states and cell types, and several other datasets involving mRNA, proteins, and chromatin accessibility (Hao et al., 2021).

Variational AutoEncoders (VAEs) are a class of neural networks commonly used for integration and dimensionality reduction [2]. Single-cell multimodal variational autoencoder (scMVAE) (Zuo and Chen, 2021) was designed to integrate transcriptomic and chromatin accessibility data, using a flavor of a VAEs. Using Product of Experts (Hinton, 2002) along with latent representations obtained by separate encoders jointly acting on each molecular layer and concatenated layers, scMVAE creates a complex shared low-dimensional representation which can be used for visualization, clustering and imputation. The input data are reconstructed by processing the latent representations via decoder neural networks, which calculate the probabilities of gene dropouts and predict the expression of measured genes modeled as a negative binomial distribution. At the same time, cell-specific scales used to normalize expression across cells are learned (Zuo and Chen, 2021). scMVAE was applied to integrate two real datasets assaying mRNA and chromatin accessibility using SNARE-seq method, as well as simulated data generated by Splatter (Stanojevic et al., 2022; Zappia et al., 2017). totalVI (Gayoso et al., 2019) is a deep VAE for matched gene expression and protein data integration. Joint latent space representations are learned via an encoder network and used to reconstruct the original data while accounting for the difference between the original data modalities. The gene expression data are sampled from a negative binomial distribution whose parameters are obtained as outputs of a decoder neural network. The protein data are sampled from a mixture model with two negative binomial distributions simulating the experimental background and the actual signal, respectively. The model was applied to two datasets containing transcriptomic and proteomic measurements and generated shared representations of cells with interpretable components (Stanojevic et al., 2022). Another model introduced after scMVAE is scMM Minoura et al. (2021). It is a deep generative model-based framework for the extraction of interpretable joint representations and cross-modal predictions. Inspired by Shi et al. (2019), scMM utilizes a mixture-of-experts (MoE) (Jacobs et al., 1991) multi-modal deep generative model and achieves end-to-end learning by modeling raw count data in each modality based on tailored probability distributions.

[2]For additional details please look in Section 4.2.

To test scMM authors relied on a matched single cell CITE-seq dataset of peripheral blood mononuclear cells (PBMCs) from vaccinated patients (Hao et al., 2021), and mouse skin SHARE-seq data (Ma et al., 2020b), demonstrating that their model can extract biologically meaningful latent variables. The closest model to the contribution made in this thesis is MultiVI. MultiVI (Ashuach et al., 2021) is a deep generative model for the integrated analysis of single cell gene expression and chromatin accessibility data. MultiVI uses jointly profiled data to learn a multi-modal model of the data and to relate measurements of individual modalities on the same population of cells (Ashuach et al., 2021). It assumes independent VAEs for each modality, whose latent manifolds are aligned during training using Jeffrey's divergence. Similarly to scMM, MultiVI was tested on PBMCs. Previously discussed models, while not directly targeting integration, they allow the use of matched data to predict one single-modality dataset into another type of modality (Gong et al., 2021). Cobolt (Gong et al., 2021), is a framework for a full integrative analysis of multi-modality and single-modality platforms. Cobolt estimates this joint representation via a novel application of Multimodal Variational Autoencoder (MVAE) (Wu and Goodman, 2018) to a hierarchical generative model Gong et al. (2021). Cobolt was tested on adult mouse cerebral cortices data (Chen et al., 2019) and PBMCs yielding a robust and meaningful representation of the cells' state.

In similarity-based methods we typically find spectral clustering techniques such as SPECTRUM (John et al., 2020), Pairwised Co-regularized Multimodal Spectral Clustering (PC-MSC) (Adossa et al., 2021; Kumar et al., 2011) and graph fusion algorithms like Weighted Nearest Neighbours (WNN)(Hao et al., 2021). The joint dimensionality reduction include matrix factorization along with covariance techniques like Canonical Correlation Analysis (CCA)(Loan, 1976). The statistical modelling techniques operate in the Bayesian setting to determine cluster probabilities of cells in multiple layers Adossa et al. (2021). An example would be BREM-SC(Wang et al., 2020), a Bayesian Random Effects Mixture model that jointly clusters paired single cell transcriptomic and proteomic data.

Despite the immense progress integration methods have made, there are still some open challenges. Many of these strategies make implicit assumptions regarding the similarity of cellular states captured by each molecular layer (Vandereyken et al., 2023). This problem becomes more apparent on matched and unmatched data integration, where correspondence between features is not immediately obvious or when non-linearities between the different molecular layers might be biologically relevant (Rautenstrauch et al., 2022; Vandereyken et al., 2023). For example, gene expression changes have been shown to be foreshadowed by chromatin priming, as seen in cell cycling via SHARE-seq (Ma et al., 2020b). Hence novel integration methods have to both identify the common anchors between molecular layers but

also to account for sample-specific and modality-specific variation in a readily interpretable way.

Here we provided a brief yet thorough summary of integration methods. In Chapter 6 we build upon these methods to develop single cell Multi View Inference (scMVI). There we delve deeper into modelling differences between various deep learning approaches and compare their integration performance on PBMC dataset. In the next section we review tools for downstream analysis.

	CCA	BindSC	LIGER	MOFA	MATCHER	MMD-MA	UNION-Com
Matched		✓		✓	✓	✓	
Unmatched	✓		✓		✓		✓

Table 4.1 **Computational single-cell multi-omics tools applying intermediate integration approaches and their applicable omic data types (Part1).**

	MAGAN	WNN	scMVAE	scMM	totalVI	MultiVI	Cobolt
Matched		✓	✓	✓	✓	✓	✓
Unmatched	✓						

Table 4.2 **Computational single-cell multi-omics tools applying intermediate integration approaches and their applicable omic data types (Part2).**

	SPECTRUM	BREM-SC
Matched		✓
Unmatched	✓	

Table 4.3 **Computational single-cell multi-omics tools applying intermediate integration approaches and their applicable omic data types (Part3).**

4.3 Downstream analysis

Downstream analysis plays a crucial role in extracting meaningful biological insights from complex molecular profiles of individual cells. Depending on the task, downstream analysis can be applied to single-omics, multi-omics or both. In this section we provide an overview of the downstream analysis methods related to the thesis and touch upon challenges.

4.3.1 Downstream analysis in single cell single-omics

Differential analysis is a common method for assessing if difference of readings between two conditions is statistically significant. It has been used for the detection of biomarkers for novel cell types or gene signatures for cellular heterogeneity, and also provides inputs for other secondary analyses including gene set or pathway, and network analysis. First they were developed for bulk RNA data, with methods like DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2009) and limma (Ritchie et al., 2015) being still in use, but later expanded to other molecular layers as well. The advent of single cell gene expression protocols with Tang et al. (2009) created datasets with novel challenges like high sparsity. This lead to the creation of novel differential analysis methods capable of addressing some of the protocol limitations. Among the most popular has been the Wilcoxon rank sum test (Zhang et al., 2021) used as default in many popular platforms like Scanpy (Wolf et al., 2018) and Seurat(Hao et al., 2021). Early differential expression analysis tools for single cell differential expression like (Kharchenko et al., 2014) used mixture of Negative Binomial and Poisson distributions for non-zero and zero readings respectively. Another popular tool uses beta-Poisson model (Vu et al., 2016) to capture bi-modality of single cell data. Additional methods for performing differential analysis on gene expression data are Negative Binomial model with Independent Dispersions (NBID) (Chen et al., 2018), ZINB-WaVE (Risso et al., 2018) and BASiCS (Vallejos et al., 2015). Similarly, for chromatin accessibility and DNA methylation we find methods like PeakVI (Ashuach et al., 2022; Muto et al., 2021), scMET (Kapourani et al., 2021) and DeepCpG (Angermueller et al., 2017). In the same category we also have models like MultiVI (Ashuach et al., 2021) which can perform differential analysis on multiple molecular layers. PeakVI and MultiVI are part of a relatively novel trend which uses VAE models to approximate the posterior probability of the alternative and null hypotheses through repeated sampling (Ashuach et al., 2021). Although many approaches exist for differential analysis, benchmarks like for gene expression data like (Soneson and Robinson, 2018) indicate that many popular methods generally perform well.

4.3.2 Downstream analysis in single cell multi-omics

To improve our understanding regarding the cellular state, it is vital to analyse single cell multi-omics data and infer regulatory relationship between multiple molecular layers (Colomé-Tatché and Theis, 2018). A popular approach used in bulk would be to use mechanism-based integration strategies that link molecular layers on individual units using approximations to assumed regulation levels (Colomé-Tatché and Theis, 2018). An example linking multiple omic levels like mRNA and protein or mRNA and microRNA is MONA

(Sass et al., 2013). To analyse regulation between various molecular layers in single cell multi-omics data, existing approached have mainly relied on computing correlations (Colomé-Tatché and Theis, 2018). At the genomic level, correlations between copy number variation and transcription levels have been studied to delineate the effect of genomic duplications and deletions on gene expression (Dey et al., 2015; Macaulay et al., 2015). Correlations between DNA methylation and transcription levels have been computed to investigate the direct role of DNA methylation on transcription and to study the effect of heterogeneous DNA methylation in the cell population on gene expression (Angermueller et al., 2016; Argelaguet et al., 2019; Clark et al., 2018; Hu et al., 2016). A meta-analysis from (Csárdi et al., 2015) on post-translational regulation clearly indicated the catastrophic effects of uncorrelated noise. Due to the single cell multi-omics protocol limitations, the phenomenon described in Csárdi et al. (2015) is exacerbated. There are two approaches to address the effects of protocol limitations. On the one hand, there are methods like SCRaPL (Maniatis et al., 2022) where data generation mechanisms are carefully modelled for each individual feature with tailored made likelihoods, hoping to uncover the underlying biological variability. On the other hand, methods like CCA remove noise by averaging noise in each molecular layer prior to the creation of a correlated latent embedding.

4.3.3 Downstream Challenges

Downstream analysis also faces its own challenges. Data sparsity and noisy nature of single cell multi-omics data data pose fundamental statistical challenges which hinder our ability to extract biological knowledge from biologically rich datasets. Noise aware methods based on Bayesian hierarchical modelling have proven to be a valuable tool for several tasks (Kapourani et al., 2021; Kazwini and Sanguinetti, 2023; Maniatis et al., 2022; Vallejos et al., 2015).

Chapter 5

SCRaPL: A Bayesian hierarchical framework for detecting technical associates in single cell multiomics data

As outlined in Section 2.3 high throughput single cell assays based on next generation sequencing are revolutionizing our understanding of biology, with profound implications both fundamental and translational (Tang et al., 2009). While single-cell transcriptomic technologies are rapidly reaching maturity, more recent platforms have emerged that enable simultaneous large scale measurements of multiple molecular layers within the same cell.

Despite the huge potential for breakthroughs, technical limitations in multi-omics technologies create formidable statistical challenges in the interpretations of their results. Single-cell multi-omics sequencing protocols tend to multiply independent sources of noise present in single-cell omics technologies (Section 2.3). As a result, direct adoption of classical statistical tools to assess associations between different molecular layers (e.g. Pearson or Spearman correlation) routinely leads to underpowered analyses, which are only able to identify a handful of significant associations (Angermueller et al., 2016; Clark et al., 2018; Hernando-Herraez et al., 2019)[1].

In this chapter, we argue that proper treatment of noise is essential in order to robustly retrieve significant statistical associations. To do so, we introduce SCRaPL (**S**ingle **C**ell **R**egul**a**tory **P**attern **L**earning), a Bayesian hierarchical model to infer associations between different omics components. The Bayesian hierarchical framework, which has already been extensively used in single-omics single-cell analyses (e.g. Kapourani et al. (2021); Vallejos et al. (2015)), explicitly and transparently decomposes noise in the data, enabling efficient

[1] In this chapter we are interchangably going to use correlations and associations to denote linear correlations between molecular layers.

extraction of biological signals from technical noise. We demonstrate on both synthetic and real data sets that SCRaPL is both highly accurate and sensitive, identifying much larger numbers of statistically significant associations than standard correlation analyses while retaining a good control on false positives.

5.1 Datasets

We applied SCRaPL in the context of two single cell multi-omic datasets, generated by scNMT-seq (Clark et al., 2018) and the 10x Genomics Multiome ATAC plus Gene Expression platform. Samples correspond to mouse embryonics stem cells (mESC) and brain cells (fresh cortex, hippocampus, and ventricular zone) (mEBC) at various developmental stages (embryonic days $4.5, 5.5, 6.5$,7.5 for mESC and 18 for mEBC), which comprise the exit from pluripotency and primary germ layer (Argelaguet et al., 2019) and the end of retinal ganglion cell generation (Robinson and Dreher, 1990). For mESC data, our analysis focuses on the correlation between gene expression and DNAm. We also experimented with chromatin accessibility and gene expression pair with not much success due to high sparsity in accessibility data. Our second case study considers mEBC data generated using the 10X Genomics Multiome ATAC plus gene expression platform and focuses on correlation between gene expression and chromatin accessibility. To quantify the levels of false positives among significant associations, we create negative control datasets by randomly permuting observations in epigenomic and transcriptomic layers of mESC and mEBC data. This shatters any correlation structure present in data, yielding important insights.

Data preprocessing is an important step prior to inference, split into two parts; integration and quality control (QC). Integration looks at aggregating single cell sequencing data stored into different files and matching them all together in a reference genome. Quality control ensures that individual components and joint data satisfy a minimum set of requirements before inference. In our preprocessing pipelines, depending the type of data we use, integration steps precede quality control and vise-versa.

Sequencing epigenomic layers of mESC data is performed using sc-NOME-seq (Pott, 2017). When sequenced fragments are mapped to the reference genome we get multiple readings CpG/GpC methylation readings for the same region in the same cell. The first step is to average and binarise them in order to get an index for each cytosine in each region and cell. For the binarization step we assume that CpG/GpC readings with 50% methylation rate or more are considered methylated. In scRRBS the coverage is sparse, which implies that the probability of capturing a CpG in one allele is very low (ie. some of the CpGs are hemi-methylated). As fragments are PCR amplified, it is common to capture several

methylated and a few unmethylated CpGs, all origniating from the same allele. To address this problem we round the results (Kapourani and Sanguinetti, 2018). To aggregate DNAm data from different mESC and link open DNA chromatin from mESC to nearby genes we follow a window based approach. Reads are mapped using the GRCm38 mouse genome (accession number GSE56879). When looking at methylation/gene expression of promoter regions in mESC data, a window of $\pm$2.5kbp was used. This window size was chosen based on the analysis of (Kapourani and Sanguinetti, 2018) following findings similar to (Jeong et al., 2014), which demonstrated the hypo-methylation in the near proximity of TSS. Hence, we set the window size to 5kbp in order to create features with better coverage. During the model development we also considered windows up to 10kbp without substantial difference in outcomes. Following (Argelaguet et al., 2019; Clark et al., 2018) analysis for chromatin accessibility/gene expression the window was set to $\pm$0.25kbp, as chromatin accessibility presents more localized patterns. Genomic regions with coverage less than 4 are discarded in the quality control step. To perform integration and binarisation, we use Bioconductor package BPRMeth (Kapourani and Sanguinetti, 2018). Similarly, for accessibility and expression in the mECB data we map enhancers to genes at most $\pm$12.5 kbp away. To control how our window choices affect results, we experimented with multiple window sizes, noticing minimal impact on the results.

Single cell RNA expression is initially filtered based on the total number of counts per cell and minimum number of expressed genes per cell. For mESC and mEBC the library sizes were 100000 and 10000 and minimum number of expressed genes 1000. Using scran (Lun et al., 2016) we then estimate a unique normalization constant for each cell. In this step we take raw expression data, filter low quality cells and estimate per cell normalization constants.

Subsequently, a quality control step was applied to both datasets. For mESC data we removed features with zero variance in each modality and for which the percentage of expression zeros was above 80%. This resulted in a dataset with 9480 features and 679 cells. For DNAm data, we have the number of captured CpGs in the 5 kbp window and how many of them were methylated in gene promoter region. Similarly for chromatin accessibility data we have the number of captured GpCs in the 0.5 kbp window and how many of them were methylated in gene promoter region. For gene expression we have the count of captured mRNAs for each genomic region. Similarly, for mEBC data we removed features with more than 80% of zeroes in accessibility or expression, leading to a dataset with 4249 features and 4052 cells. For chromatin accessibility data, we have the number of accessible peaks in enhancers at most 12.5 kbp away from nearby gene promoter. For gene expression we have the count of captured mRNAs for each genomic region.

5.2 Methods

SCRaPL is a tool for exploratory analysis of high-throughput, single cell data, which aims to establish more robust linear associations between different molecular layers. The example we will focus on here is relating the expression of a specific gene with its epigenetic state, measured either by DNA methylation (DNAm) or chromatin accessibility, however different types of associations might be considered, by introducing alternative noise models within our framework.

Our starting point is the observation that correlations, while invaluable tools to generate hypotheses, are critically sensitive to noise. In particular, it is well known that adding uncorrelated noise to correlated random variables reduces the estimates of correlation, thus weakening statistical power in any analysis. To obviate this problem, SCRaPL introduces a hierarchical model, schematically described in Figure 5.1.

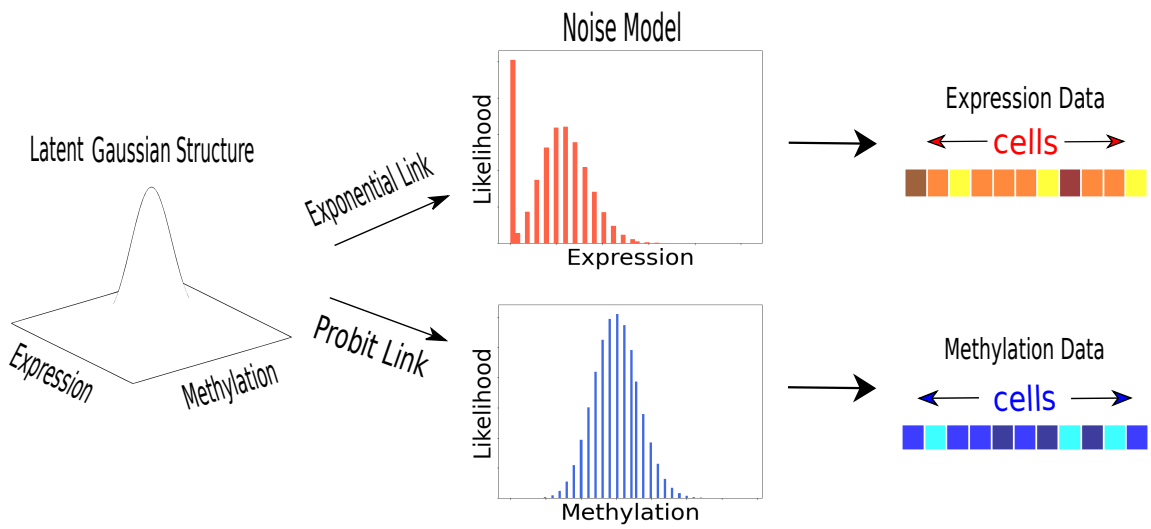


Fig. 5.1 **Schematic representations of SCRaPL.** SCRaPL combines error models tailored to single cell multi-omics data with Bayesian hierarchical frameworks to discover correlative associations between epigenomic and transcriptomic modalities in different genomic regions.

5.2.1 SCRaPL model

SCRaPL implements a Bayesian hierarchical approach that is tailored to the data generated by single cell multi-omic assays. Here, we assume that matched data is available for two molecular phenotypes, but our formulation is flexible and can in principle be expanded to include additional layers.

For each cell i ($\in \{1,\ldots I\}$) and genomic region j ($\in \{1,\ldots J\}$), observed transcriptomic and epigenomic marks $\mathbf{Y}_{ij} = \begin{pmatrix} Y_{ij1} \\ Y_{ij2} \end{pmatrix}$ are modeled using modality specific likelihoods parameterized to depend on latent bivariate Gaussian vector $\mathbf{Z}_{ij}$ with unknown latent mean $\boldsymbol{\mu}_j$ and covariance $\boldsymbol{\Sigma}_j$.

$$\mathbf{Z}_{ij} = \begin{pmatrix} Z_{ij1} \\ Z_{ij2} \end{pmatrix} \,|\, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \overset{ind}{\sim} \mathrm{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right), \tag{5.1}$$

where

$$\boldsymbol{\mu}_j = \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma_{j1}^2 & \rho_j \sigma_{j1} \sigma_{j2} \\ \rho_j \sigma_{j1} \sigma_{j2} & \sigma_{j2}^2 \end{pmatrix}. \tag{5.2}$$

In this formulation, we assume independence across all features. This enables trivial parallelization across features which guarantees both speed and effective posterior space exploration. For example in the mESC dataset, the 9480 features imply a model with ~ 13 million parameters. Such space is impossible to effectively explore with state of the art methods. With feature independence we can break the large model of ~ 13 million parameters into several smaller ones (say ~ 1500 by fitting 11 features at a time) and guarantee the same results. The latent Gaussian variables are then passed through a suitable nonlinear link function to generate the expected value of the observation. The observation noise model, as well as the nonlinear link function, are tailored to the type of assay being analysed (and can also be designed in a data-driven fashion by using model selection techniques). For count data (i.e. gene expression in mESC/mEBC and chromatin accessibility in mEBC) we use a zero-inflated Poisson noise and for the rest (i.e. DNAm and accessibility in mESC) we use a Binomial distribution. Graphical representations for the model implemented in SCRaPL addressing mESC and mEBC data is provided in Figure 5.2.

Parameter interpretation. To aid the interpretation of each model parameter, mean and variance expressions are derived for the noise models introduced above, after integrating out the distribution of the latent vector $\mathbf{Z}_{ij}$. In both cases, μ_{j1} and μ_{j2} control the overall RNA expression and DNAm values for the population of cells under study. Moreover, σ_{j1} and σ_{j2} capture the excess of variability (overdispersion) that is observed with respect to the baseline noise model. Finally, ρ_j captures the latent correlation between molecular layers.

RNA expression noise model. Let Y_{ij1} be a random variable representing the number of raw read-counts observed for each cell i and feature j. Conditional on the value of the latent

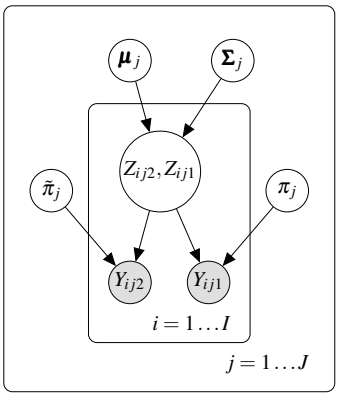


Fig. 5.2 **Graphical representations of SCRaPL fitted to each genomic region $j \in J$ with mESC and mEBC data.** Observed genomic data (Y_{ij1} is RNA expression; Y_{ij2} is Chromatin Accessibility), with their associated latent variables (Z_{ij1}, Z_{ij2}) and feature-specific model parameters ($\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$) capturing feature specific epigenomic and transcriptomic marks. Additional inflation parameters π_j and $\tilde{\pi}_j$ are specific to Gene Expression and Chromatin accessibility error models. For the case of mESC data where epigenome is measured differently, parameter $\tilde{\pi}_j$ is dropped.

variable X_{ij1}, we use an exponential link function and assume that

$$\mathbb{P}\left(Y_{ij1} = y_{ij1} \,|\, Z_{ij1} = z_{ij1}, s_i, \pi_j\right) = \begin{cases} \left(1-\pi_j\right) \frac{\left(s_i e^{z_{ij1}}\right)^{y_{ij1}} \exp\left(-s_i e^{z_{ij1}}\right)}{y_{ij1}!} \textbf{ if } y_{ij1} > 0, \\ \pi_j + \left(1-\pi_j\right) \exp\left(-s_i e^{z_{ij1}}\right) \textbf{ if } y_{ij1} = 0. \end{cases} \tag{5.3}$$

The latter corresponds to a zero-inflated Poisson (ZIP) model with an exponential link, where s_i (> 0) is a cell-specific scaling factor that accounts for global differences across cells (e.g. due to sequencing depth) and π_j ($\in [0,1]$) represents a zero-inflation parameter (if $\pi_j = 0$, Eq (5.3) reduces to a Poisson model). The exponential link function leads to a zero-inflated Poisson-lognormal model, whose variations have been previously used for single cell RNA sequencing (scRNAseq) data (Gu et al., 2014; Townes et al., 2019). In practice, we infer scaling factors s_i using scran (Lun et al., 2016) and use them as known model offsets. As scRNA-seq data count the number of mRNAs for each genomic region, the most natural distributions choice would be Poisson or Negative Binomial. The Negative Binomial can be viewed as a Poisson distribution with Gamma distribution rate. Hence our choice of Poisson with log-normal rate is similar and guarantees the correct handling of count data.

The need for a zero-inflation component is a matter of debate for scRNA-seq data (Svensson, 2020) and may depend on the experimental protocol used to generate the data. See

Comparing between alternative models (Section 5.2.6) later in this Section for a quantitative approach to evaluate the need for zero-inflation in specific datasets.

DNAm noise model. For each cell i and feature j, let n_{ij} be the number of CpG sites within a pre-specified genomic region (e.g. gene promoter) for which DNAm reads were obtained. These capture differences in coverage[2] across cells and features. The conditional model for the number of methylated CpG sites Y_{ij2} is then assumed to follow a binomial distribution such that

$$\mathbb{P}\left(Y_{ij2} = y_{ij2} \,|\, Z_{ij2} = z_{ij2}, n_{ij}\right) = \binom{n_{ij}}{y_{ij2}} \left(\Phi\left(z_{ij2}\right)\right)^{y_{ij2}} \left(1 - \Phi\left(z_{ij2}\right)\right)^{n_{ij} - y_{ij2}}, \tag{5.4}$$

where $\Phi(\cdot)$ denotes a probit link function. We choose the conditional binomial error model with probit link following the example of previously introduced work like (Kapourani and Sanguinetti, 2016, 2019) as it yields a computationally tractable model.

Chromatin accessibility noise model. The choice of noise model depends on the format of the input data. If the data consists of the number of features Y_{ij2} in a genomic region (e.g. as in Buenrostro et al., 2015), our modelling approach analysis follow the one described for RNA expression data. If the data consists of open peaks within a genomic region (e.g. as for the sc-NOME-seq (Pott, 2017) protocol), the same binomial noise model used for DNAm data can be applied.

5.2.2 Prior Specification

A popular prior choice for covariance matrices is the inverse Wishart distribution. However, this has been shown to bias correlation coefficients depending whether marginal variances are small or large (Liu et al., 2016). Instead, Barnard et al. (2000) used a separation strategy to decouple correlation from marginal variances. Our prior specification for $\mathbf{\Sigma}_j$ is based on the parametrization introduced in Eq (5.2), with independent priors assigned to all

[2]Contrary to the standard genomics coverage introduced in Chapter 2, we define CpG coverage as the number of CpG loci captured within a predefined region. In single cell methylation a specific CpG locus can be captured zero, once or twice, with twice being vanishingly improbable.

feature-specific parameters. Our prior specification is given by

$$\pi_j \overset{ind}{\sim} \text{Beta}(a_j, b_j), \tag{5.5}$$

$$\tilde{\pi}_j \overset{ind}{\sim} \text{Beta}(a_j, b_j), \tag{5.6}$$

$$\boldsymbol{\mu}_j \overset{iid}{\sim} \mathbf{N}(\mathbf{m}, \mathbf{H}), \tag{5.7}$$

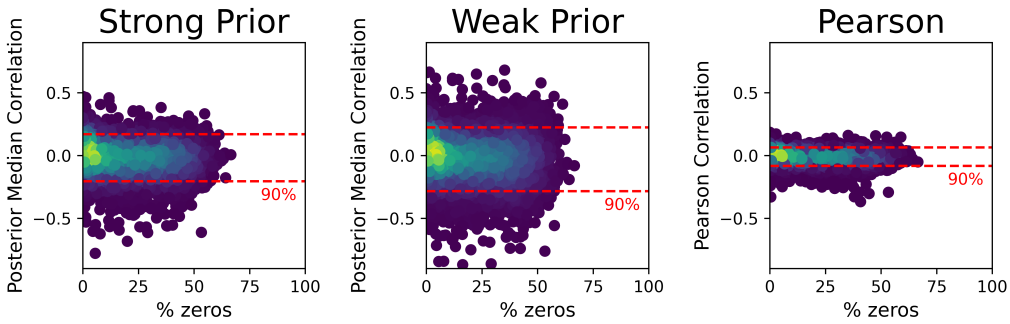
$$\sigma_{j1}, \sigma_{j2} \overset{iid}{\sim} \text{Inv-Gamma}(c_1, c_2), \tag{5.8}$$

$$\rho_j \overset{iid}{\sim} \text{TB}(d_1, d_2) = \frac{1}{\text{B}(d_1, d_2)} \frac{1}{2^{d_1+d_2-1}} \left(1+\rho_j\right)^{d_1-1} \left(1-\rho_j\right)^{d_2-1} \tag{5.9}$$

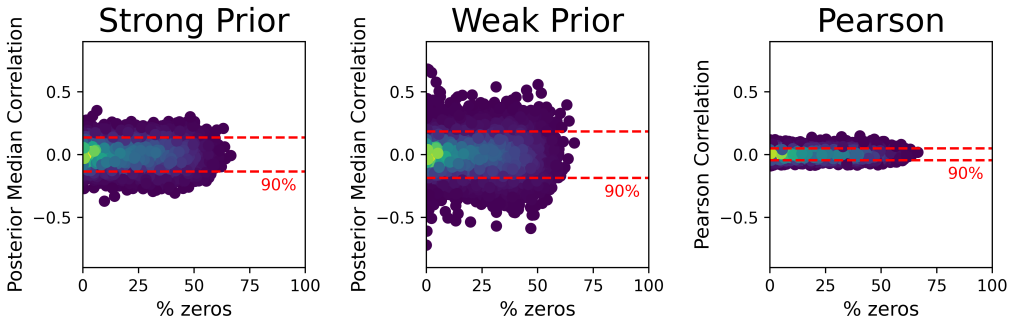
A summary of the default hyper-parameter values used in SCRaPL is presented in Table 5.1.

In order to avoid systematically favoring positive or negative correlations, we centered the prior at 0 by setting $d_1 = d_2$. Then to choose between weak (small d_1) and strong (large d_1) priors experiments were carried out on negative control data (see Section 5.1 and Figure 5.6). More specifically we inferred posterior correlation distributions for mESC data along with a corresponding negative control dataset using a weak (i.e. $\rho_j \sim \text{TB}(3,3)$) and a strong prior (i.e. $\rho_j \sim \text{TB}(15,15)$) [3]. To help identify the most suitable prior we drew scatter plots of posterior median correlation as a function of percentage of zeros in each feature. To complete the comparison we also included a similar scatter plot containing Pearson correlation estimates for each feature.

[3]We use the term weak to refer to a prior with higher entropy which tends to be broader (ie. TB(3,3)). Similarly we reserve the term strong to denote $\text{TB}(15,15)$, which is a prior with lower entropy (ie. More concentrated)



(a)



(b)

Fig. 5.3 **Plots used to tune prior correlation parameters.** The decision if a stronger or a weaker prior better suit our data, is based on a set of three scatter plots drawn using the original and a negative control dataset. The first two look at posterior median correlation as a function of % of gene expression zeros in the mESC data (Figure 5.3a) and corresponding negative control (Figure 5.3b) data for strong prior $\mathrm{TB}(15,15)$ (left), weak prior $\mathrm{TB}(3,3)$ (middle). For comparison we also plot Pearson correlation as a function of % of zeros (right). Each dot represents a feature. Negative control dataset, where no strong correlations are expected goes against what we observe. Therefore, we chose the stronger prior.

As evident from Figures 5.3a and 5.3b the strong prior ($\rho_j \sim \mathrm{TB}(15,15)$) tends to push correlations towards zero in all features, unless dictated otherwise by the data. In contrast, the weak prior ($\rho_j \sim \mathrm{TB}(3,3)$) allows greater correlation values across genes. The number of strong posterior correlations identified in experiment with the weak prior on negative control data is a deterring factor as it implies elevated false positive rates. As indicated by Figure 5.3b, Pearson correlation values tend to be more conservative across genes. This is not the case for the weak prior, where we can clearly see several confident estimates despite the lack of data support. Since strong correlation values goes against our expectations for negative control data, we chose the $\rho_j \sim \mathrm{TB}(15,15)$ which will suppress false positives and present closer detection patterns to Pearson correlation in both true and negative control data. A full sensitivity analysis on posterior correlation from the prior for mESC and mEBC data

can be also found in Section A.12. In the examined features any prior TB with parameters $d_1 = d_2 \geq 3$ and $d_1 = d_2 \geq 8$ will yield fairly similar posterior correlation values in mESC and mEBC datasets. Despite the stable results for d_1, d_2 values greater than 8, detection rates for $d_1 = d_3 = 3$ stay relatively robust as evident from Tables 5.2 and 5.3. The use of negative control data, the 90% quantile used to estimate γ and EFDR are filter layers which increase prediction robustness.

Apart from the correlation priors, we also investigated the impact of latent standard deviation and mean parameters in posterior covariance. More precisely, for latent standard deviation we varied (c_1, c_2) to attain $(2.5, 2.5), (2.5, 3.5), (2.5, 3.5), (3.5, 4.5)$ and $(4.5, 4.5)$. For mESC data posterior standard deviations and correlation were hardly impacted from these change as evident from Section A.12.1. A few exceptions can be found in mEBC like features 477 and 731 as evident from Fig. A.91 and A.92. However these are probably the exceptions as for every other example we could find results tend to be stable. Negligible impact on the covariance was observed by altering latent mean prior values. In particular $\boldsymbol{\mu}_j$ was varied between $(-1, 4), (0, 4), (1, 4), (1, 2), (1, 3)$ and $(2, 4), (3, 3), (3, 4), (3, 5), (4, 4)$ from mESC and mEBC data.

d_1	d_2	a_j	b_j	c_1	c_2	$\boldsymbol{\mu}_j$(mESC/mEBC)	**H**
15	15	2	8	2.5	4.5	$(4,1)'/(4,3)'$	$I_{2\times 2}$

Table 5.1 **Table summarizing the SCRaPL default hyper-parameter values.**

	Orig.	Neg. 1	Neg. 2	Neg. 3	Neg. 4	Neg. 5
SCRaPL	330	13	12	20	14	10
Pearson	49	0	0	0	0	0
Spearman	69	0	0	0	0	0

Table 5.2 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in DNAm-expression modalities of mESC data when** $\gamma = 0.29$

	Orig.	Neg. 1	Neg. 2	Neg. 3	Neg. 4	Neg. 5
SCRaPL	4200	70	55	59	60	56
Pearson	542	0	0	0	0	0
Spearman	0	0	0	0	0	0

Table 5.3 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in accessibility-expression modalities of mEBC data when** $\gamma = 0.155$

5.2.3 Implementation

SCRaPL uses Bayesian inference to reconstruct the latent mean values and covariance from independent observations over many cells. As the posterior distribution associated to the model above does not have a closed analytical form, inference is implemented using No-U-Turn Sampler (Hoffman and Gelman, 2014), a state of the art variation of Hamiltonian Monte Carlo (Duane et al., 1987). In particular we have used the Tensorflow probability (V et al., 2017) implementation with dual averaging step to sample from the joint distribution of each feature.

For all the analyses shown in this article, we obtained 10000 samples from this algorithm and discarded the first 5000 iterations (burn-in) before estimating model parameters. NUTS hyperparameters were optimized during burn-in to an acceptance ratio of 0.65 (Beskos et al., 2013), using Tensorflow probability (V et al., 2017) black-box parameter tuning. Convergence is monitored using the Gelman-Rubin criterion (Gelman and Rubin, 1992).

A vital step in inference with MCMC is to asses whether MCMC chains for each parameter have converged. Such analysis is presented in Section A.9 for several features for models with and without inflation in mESC and mEBC data. The majority of chains have converged to the same posterior. There are some features which could potentially benefit from further sampling such as Figures A.31,A.38,A.40. However, further sampling would benefit parameters other than correlation such as latent mean and standard deviation. So we can proceed with the correlation analysis. In models without inflation mean and standard deviation tend to be under and over estimated respectively. Moreover, it seems that convergence is harder to achieve. This also demonstrates additional benefits from choosing the inflation parameter apart from better WAIC and DIC fit.

5.2.4 A probabilistic rule to detect statistically significant across layers

SCRaPL identifies features with statistically significant correlation across multi-omics layers (e.g. RNA expression and promoter DNAm) based on the posterior distribution of feature-specific latent correlation parameters ρ_j. Our decision rule depends on whether the posterior mass for $|\rho_j|$ is concentrated around high values. As in Bochkina and Richardson (2007), this is quantified by the following tail posterior probabilities

$$p_j(\gamma) = \mathbb{P}\left(\left|\rho_j\right| \geq \gamma\right), \tag{5.10}$$

where γ (> 0) denotes a minimum correlation threshold. If $p_j(\gamma)$ is greater than a probability threshold α, a statistically significant correlation is reported for feature j.

Suitable values for γ and α could be chosen using different approaches. In principle, γ can be fixed a priori by the user [4]. Instead, we adopt a data-driven approach based on the distribution of feature-specific posterior estimates obtained for $|\rho_j|$ using negative control datasets. Such distribution can be used to quantify the strength of correlation estimates that can be expected by chance for a given sample size and sequencing depth. As a default choice, we select γ to match the 90% quantile of the distribution described above. For a fixed value of γ, a grid search is used to select α according to a target (Expected False Discovery Rate, EFDF; Newton et al., 2004). The latter is defined as

$$\text{EFDR}_\alpha = \frac{\sum_{j=1}^{J}\left(1 - p_j(\gamma)\right)\mathbb{I}\left(p_j(\gamma) \geq \alpha\right)}{\sum_{j=1}^{J}\mathbb{I}\left(p_j(\gamma) \geq \alpha\right)}, \tag{5.11}$$

where $\mathbb{I}(A) = 1$ if A is true, 0 otherwise. Our default target EFDR is equal to 10%.

5.2.5 Current approach based on Pearson/Spearman correlation.

To date, single cell multi-omics analyses have primarily used the Pearson/Spearman correlation coefficient r to quantify associations between different types of molecular data (e.g. Angermueller et al., 2016; Clark et al., 2018). These estimates are directly derived from the observed data and do not assume a specific noise model. As the input for this calculation, gene expression data is typically normalised (e.g. using scran; Lun et al., 2016) and subsequently log-transformed after adding a pseudocount, while DNAm is normalised by

[4] The edge case where γ is a priori set to 0 is problematic as Equation 5.10 is always 1. In case of symmetric around 0 posteriors the rule from (Bochkina and Richardson, 2007) based on the maximum posterior probabilities associated with the one-sided hypothesis: $2\max(\pi_j, 1-\pi_j) - 1$, with $\pi_j = \mathbb{P}(\rho_j \geq 0)$ can be used. For the non-symmetric examples the same rule is used as an approximation.

coverage (note that the addition of a pseudocount is arbitrary and has been shown to distort variance estimates; Townes et al., 2019).

Based on these estimates, statistically significant correlations are selected by contrasting the hypotheses $H_0 : \|r\| \leq u$ and $H_1 : \|r\| \geq u$, for some threshold u [5]. To the best of our knowledge the generalized case of $u > 0$ has not been systematically treated. By numerically marginalizing the joint density $f(r, \tilde{\rho}_j)$ of the population ($\tilde{\rho}_j$ and sample correlation (r) coefficients (Hotelling, 1953) we derive a similar expression to Equation 5.10 for the considered frequentist alternatives:

$$\tilde{p}_j(\tilde{\gamma}) = 2 \int_{-1}^{-|\tilde{\gamma}|} \int_{-u}^{u} \frac{1}{\mathbb{Z}} f(r, \tilde{\rho}_j) d\tilde{\rho}_j dr, \tag{5.12}$$

where $\mathbb{Z} = \int_{-1}^{1} \int_{-u}^{u} \frac{1}{\mathbb{Z}} f(r, \tilde{\rho}_j) d\tilde{\rho}_j dr$ and $\tilde{\gamma}\,(> 0)$ denotes a minimum correlation threshold.

In practice we set $\tilde{\gamma} = \gamma$, to make experiments as comparable as possible. To control Type I error (False Discovery Rate) across features we use the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) and set it at 10%

5.2.6 Comparing between alternative models

SCRaPL is a noise-aware approach fitting error models crafted to address challenges related to various multi-omics data across features. Our likelihood of choice for count data such as gene expression and accessibility in 10X data has been Poisson distribution. Since there is a debate surrounding the extend to which zero-inflation is required (Svensson, 2020), we take an unbiased stance and use posterior model samples to perform model selection using Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and Watanabe Information Criterion (WAIC; Watanabe, 2010). DIC is a method for assessing goodness of fit while penalizing large effective numbers of parameters between alternative models, with lower DIC values indicating more preferable models. WAIC has been shown to be asymptotically equal to using Bayesian leave one-out cross-validation (Watanabe, 2010).

To assess the need of zero-inflation in SCRaPL, we fit the zero-inflated and the standard Poisson in the methylation/expression layers of the mESC data and accessibility/expression of mEBC data across features. For the large majority of features in the mESC and mEBC, WAIC and DIC favor zero inflation as it is indicated from Figure 5.4 [6]. Furthermore, in Section A.10 we perform posterior checks where synthetic data are fitted from SCRaPL

[5] When $u = 0$ we get the standard Pearson/Spearman hypothesis test.

[6] Despite both WAIC and DIC strongly indicating that zero inflation is more appropriate for gene expression data, these results are tied to mESC/mEBC data. Users must perform a similar analysis to decide whether ZI is required for their data.

with and without inflation. For the majority of features the model with inflation seems to fit better the data, providing the necessary evidence in favor of zero inflation for both mESC and mEBC data. In mEBC data, SCRaPL's limitations due to underlying linearity assumptions become obvious when comparing simulated to raw data. However, when comparing SCRaPL with and without inflation no significant differences are apparent.

Apart from the zero inflation component the Poisson likelihood is chosen in single cell multi-omics for their ability to explain data over-dispersion. In SCRaPL the use of ZIFP model with exponential link introduces intractability which makes impossible to algebraically marginalize posteriors. Hence, to perform inference in each feature computationally intensive algorithms such as MCMC are required to explore the high dimensional posteriors. To demonstrate the importance of over-dispersion we follow a similar approach using WAIC. For this experiment, we introduce a simplified version of SCRaPL with Normal likelihood and fit it to mESC and mEBC data. More information about the model can be found in Section A.11. For the majority of features in mESC and mEBC, over-dispersion seems to improve the fit irrespective of inflation, based on WAIC, as it is evident from Figure 5.5.

5.3 Experimental setup

5.3.1 Synthetic data experimental design

To assess the estimation performance of SCRaPL, we experimented on synthetic datasets consisting of 300 simulated features (pairs of gene expression and promoter methylation values). The experiments were varied to cover a number of different scenarios [7]: numbers of cells; coverage levels; fraction of zeros in expression data (zero inflation, ZI); as well as different latent mean and covariance structures. In all cases, latent means and standard deviations were set as $\mu_{j1} = 4$, $\mu_{j2} = 1$, $\sigma_{j1} = 3$ and $\sigma_{j2} = 2$. Unless otherwise stated, our simulations were based on: $I = 60$ cells, $J = 300$ features, 20% ZI rate on average for the expression data ($\pi_j = 0.20$) and an average methylation coverage (n_{ij}) equal to 275 (sampled from a Uniform distribution with range $[50, 500]$) across cells and genes. When varying the number of cells, we use $I \in \{5, 10, 25, 50, 100, 200, 400, 800, 1600\}$. When varying expression ZI, we use $\pi_j \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8\}$. When varying methylation coverage, we sample n_{ij} from Uniform distributions with ranges given by $[5, 10]$, $[10, 20]$, $[20, 50]$, $[50, 250]$ and $[500, 1000]$. The simulation study illustrates the typical behavior of the model. As sensitivity analysis is performed on real data, we prioritized evaluating the difference due to sample size, coverage and fraction of zeros.

[7] Formal justification for experimenting with these parameters is provided in Section A.5.

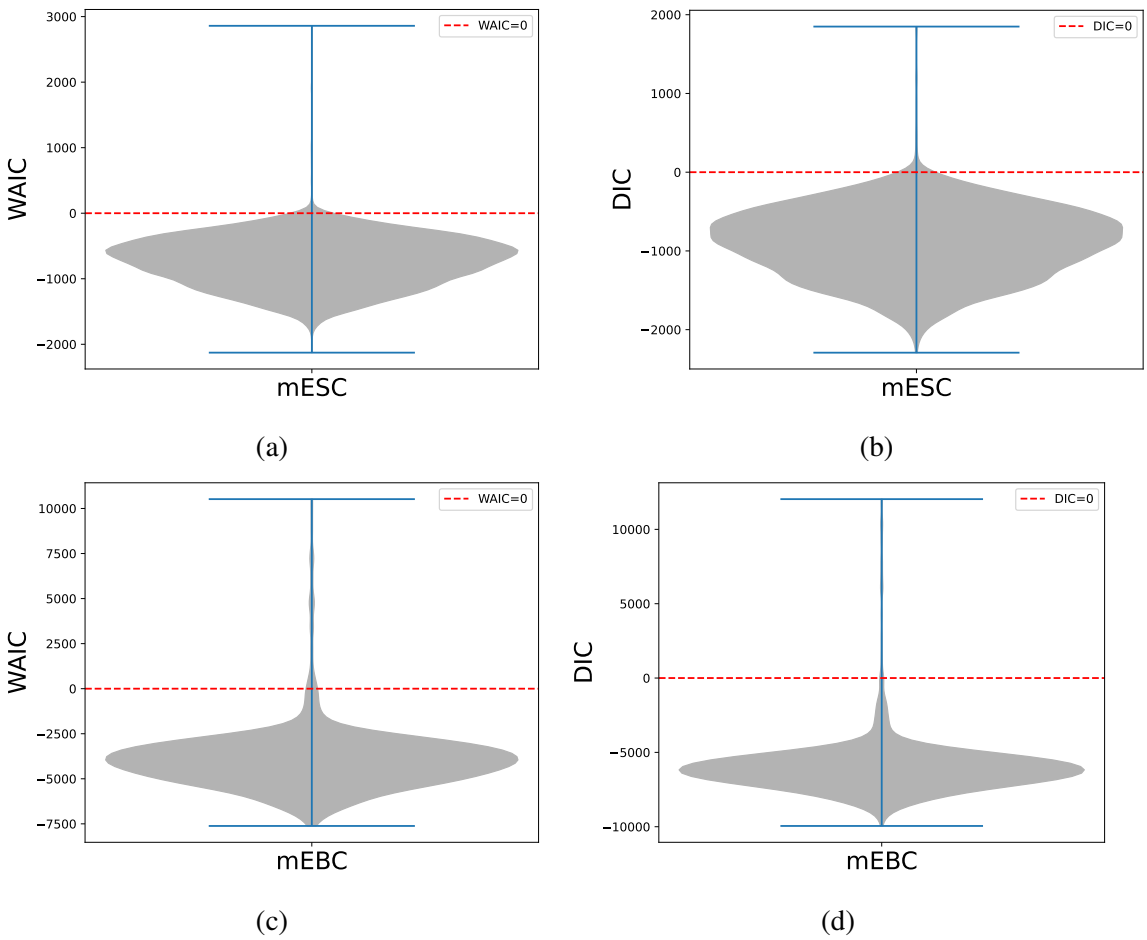


Fig. 5.4 **Distribution of WAIC and DIC difference between model with and without inflation for mESC and mEBC data across features. Evidence demonstrating the importance of zero inflation.** The more negative the difference, the stronger the evidence in favor of the model with zero inflation on the gene expression component for a particular feature. As a visual reference, zero is marked with dashed red line.

The first set of synthetic data experiments, described below, were generated using SCRaPL's generative model to asses its identifiability.

- **Experiment 1**: varying numbers of cells ($I \in \{5, 10, 25, 50, 100, 200, 400, 800, 1600\}$) and correlation values sampled from TB distribution ($\rho_j \sim \text{TB}(15, 15)$).
- **Experiment 2**: varying ZI rate ($\pi_j \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8\}$) and correlation values sampled from TB distribution ($\rho_j \sim \text{TB}(15, 15)$).
- **Experiment 3**: varying methylation coverage (n_{ij} sampled from Uniform distributions with ranges given by $[5, 10]$, $[10, 20]$, $[20, 50]$, $[50, 250]$, $[500, 1000]$) and correlation values sampled from TB distribution ($\rho_j \sim \text{TB}(15, 15)$).

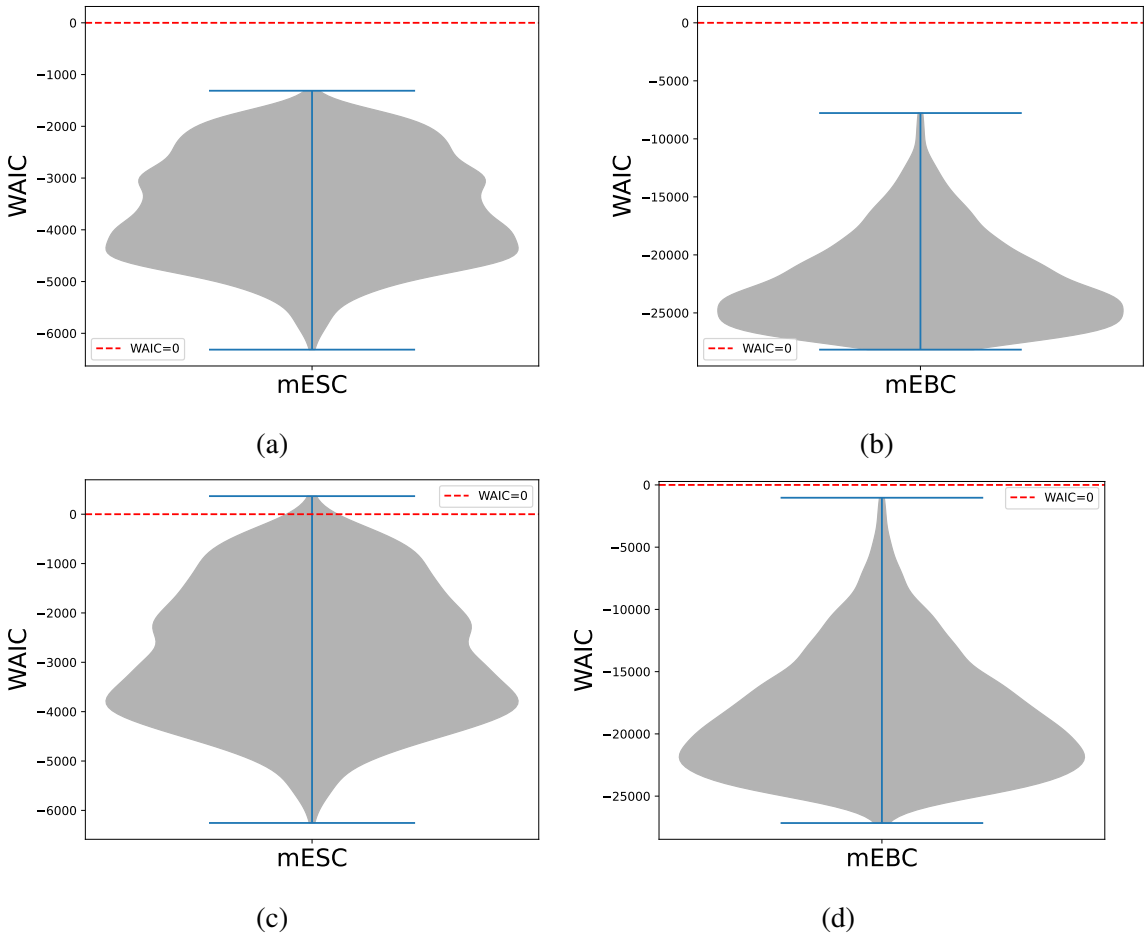


Fig. 5.5 **Distribution of WAIC difference between models with and without dispersion for mESC and mEBC data across features. FIgures demonstrating the importance of dispersion.** The more negative the difference, the stronger the evidence in favor of the model with dispersion on the gene expression component for a particular feature. As a visual reference, zero is marked with dashed red line. (5.5a) WAIC distribution of difference between ZIFP and Normal likelihood in mESC data. 5.5b) WAIC distribution of difference between ZIFP and Normal likelihood in mEBC data.(5.5c) WAIC distribution of difference between Poisson and Normal likelihood in mESC data. 5.5d) WAIC distribution of difference between Poisson and Normal likelihood in mEBC data.

The second set of experiments, explained below, were generated using SCRaPL's generative model where correlation ρ_j prior $\mathrm{TB}(15,15)$ was replaced with a $U_{[-0.8,-0.6]}$. Understanding the impact of the investigated parameters for extreme data generating samples helps develop useful insights for real data experiments.

- **Experiment 4**: varying numbers of cells ($I \in \{5,10,25,50,100,200,400,800,1600\}$) and correlation values sampled from a Uniform distribution ($\rho_j \sim U_{[-0.8,-0.6]}$).

- **Experiment 5**: varying ZI rate ($\pi_j \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8\}$) and correlation values sampled from a Uniform distribution ($\rho_j \sim U_{[-0.8,-0.6]}$).
- **Experiment 6**: varying methylation coverage (n_{ij} sampled from Uniform distributions with ranges given by $[5, 10]$, $[10, 20]$, $[20, 50]$, $[50, 250]$, $[500, 1000]$) and correlation values sampled from a Uniform distribution ($\rho_j \sim U_{[-0.8,-0.6]}$).

The final set of experiments data were partly sampled from a deep generative model described in (Lopez et al., 2018) and partly from the model. The deep generative model was used to generate latent expression and cell specific normalization constants. The rest of the parameters were sampled from the SCRaPL model conditioned on latent expression. **Experiments 7-9** help understand cases where our proposed linear generative model does not match the underlying data distribution.

- **Experiment 7**: varying numbers of cells ($I \in \{5, 10, 25, 50, 100, 200, 400, 800\}$) and correlation values sampled from TB distribution ($\rho_j \sim \text{TB}(15, 15)$).
- **Experiment 8**: varying ZI rate ($\pi_j \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8\}$) and correlation values sampled from TB distribution ($\rho_j \sim \text{TB}(15, 15)$).
- **Experiment 9**: varying methylation coverage (n_{ij} sampled from Uniform distributions with ranges given by $[5, 10]$, $[10, 20]$, $[20, 50]$, $[50, 250]$, $[500, 1000]$) and correlation values sampled from TB distribution ($\rho_j \sim \text{TB}(15, 15)$).

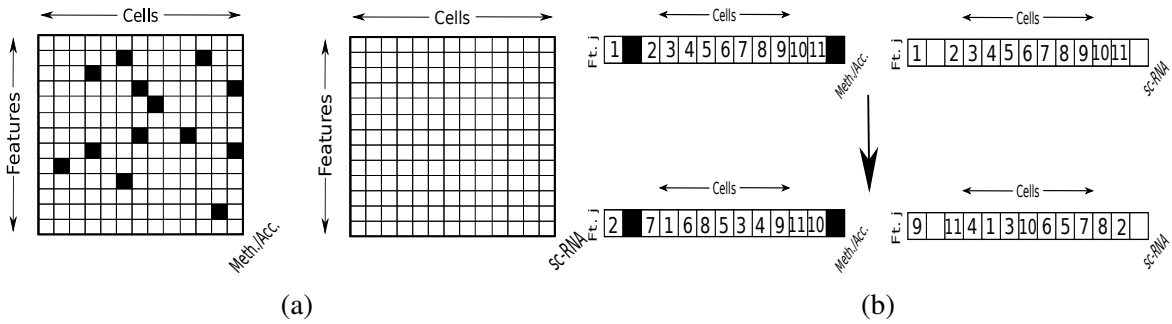
5.3.2 Comparisons on real data

To compare the power of different methods to detect associations between molecular layers, we considered, alongside SCRaPL, the classical Spearman and Pearson correlation tests (Section 5.2.6). The latter in particular has been widely used for single cell multi-omics data (e.g. Angermueller et al. (2016); Clark et al. (2018)); neither method takes into account noise in producing estimates of correlation. Molecular layer associations were retrieved as significant by controlling EFDR and FDR to 10%, respectively.

To characterize if SCRaPL's larger detection rate compared to alternatives was due to greater power, or simply to greater vulnerability to false positives, we proceed pragmatically by constructing 5 negative control data sets in which available observations of DNAm/chromatin accessibility and gene expression values for a particular feature in different cells are randomly permuted. Epigenomic and transcriptomic layers are stored in separate tensors with rows and columns representing features and cells respectively (as seen in Figure 5.6a). In plate based data missing readings in epigetic modalities is schematically denoted with

black boxes. To create negative control data we iterate through features, generating distinct permutation for each molecular layer while respecting position of missing epigenomic readings and making sure that normalization constants end up in the same position with permuted transcriptomic readings. This process is schematically and algorithmically presented in Figure 5.6b and Algorithm 2. By destroying any correlation structure between the two quantities, features detected as significant in negative control data could be considered as false positives.

To explain the result difference between SCRaPL and Pearson we carry out a series of experiments on a gene and cell population level. Initially, we consider the set of associations which are called as significant by at least one method, and split it into 3 categories: agreement between predictions, association labeling as significant by SCRaPL, but not by Pearson-/Spearman testing, and vice-versa. We then analyze these three sets attempting to detect common patterns, discussing some examples to substantiate our findings. Then we proceed to regress the absolute difference of Pearson from posterior correlation against coverage and zero inflation. Finally, the exploratory analysis is completed by linking genomic regions with strong regulatory action, identified by SCRaPL with biological phenomena observed in early embyogenesis and compare them with the Pearson predictions.



(a) (b)

Fig. 5.6 **Sketch of tensors storing methylation/accessibility(left) and expression(right) respectively. Visual explanation of the negative control experiments.**(5.6a) Black boxes represent missing methylation/accessibility data. (5.6b) Sketch of cell scrambling for a random feature j. It is important to note that despite cells are permuted at the end of scrambling, cell labels remain fixed.

Algorithm 2 Negative control dataset pipeline

Input: Dataset of interest.

Output: Negative control dataset

1: Load dataset of interest and compute the number of genomic features J.

2: **for** $ii = 1, 2, \ldots, J$ **do**

3: Compute number of cells I for that genomic region of interest.
4: Define 2 vectors with all integers ranging from 1 to I with dimension $1 \times I$.
5: Compute two distinct random permutations for these vectors.
6: Apply permutation in each epigentic mark.
7: Save updated data into a new file.

5.4 Results

5.4.1 Benchmarking SCRaPL on synthetic data

We start by considering a situation of perfect model specification (**Experiments 1-3** in Section 5.3.1), in order to assess the identifiability of our model and to document the degradation of correlation estimates obtained with classical methods. In this case, we observe that all methods provide estimates of correlation with zero-mean expected error, with an accuracy which increases with the number of cells in the data set. However, particularly for relatively low numbers of cells, the accuracy of the estimates was considerably higher for SCRaPL than for the classical Pearson and Spearman methods. Figure 5.7a shows a comparison of the three methods as we vary the number of cases, with a ZI level fixed to 20% (a benign setting similar to what encountered in high-depth plate-based technologies). SCRaPL outperforms both Spearman and Pearson by a large margin for all numbers of cells considered. Even in the most favourable case of 1600 cells (an unreasonably large number of cells for plate-based technologies), Pearson and Spearman systematically underestimate the (absolute value) of the correlation, while SCRaPL returns an accurate estimation for all true correlation values, as shown in Figure 5.7b. So, while overall all methods are unbiased in their estimates, Spearman and Pearson systematically underestimate the absolute value of the correlation, potentially leading to lesser power (see Section 5.4.2). As expected, the performance for all methods degrades with increasing levels of ZI (see Figure A.2 in Section A.1). However, we did not observe significant differences for SCRaPL correlation estimates across different levels of coverage (see Figure A.3 Section A.1). This is explained by the indirect impact of coverage in correlation, hinted by the analysis presented in Section A.5.

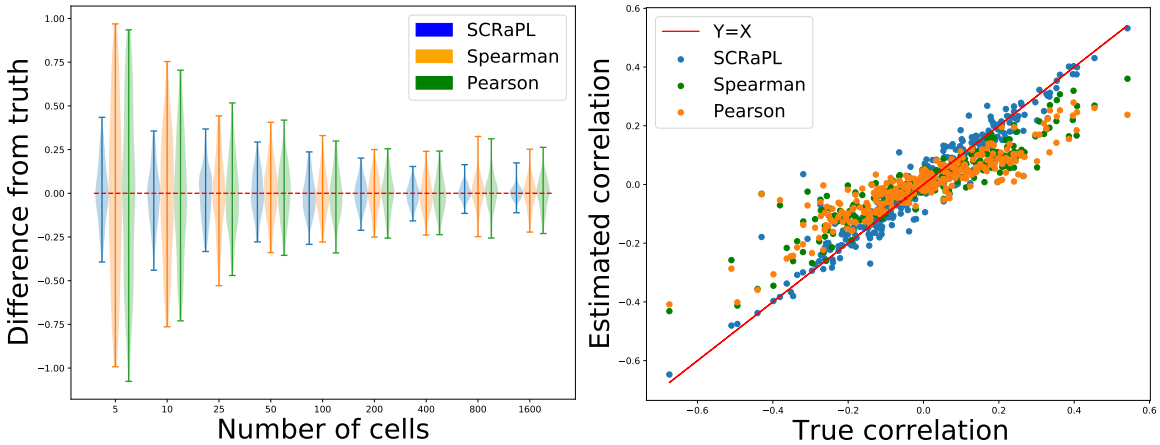
To probe the importance of prior specification, we generated data where the underlying correlation values ρ_j were in an area with low prior mass (**Experiments 4-6** in Section 5.3.1). In this case, we did observe some bias in our estimates (see Figures A.4-A.6 in Section A.1), particularly when the number of cells is low. Similarly, performance diminishes with increasing ZI levels and stays relatively intact across different coverage levels.

As a final test of more severe model mismatch, we evaluated predictive performance in a scenario where we retained the same noise model, but replaced the latent multivariate Gaussian distribution by expression rates inferred using a variational auto-encoder (scVI; Lopez et al., 2018) that was trained on the scRNAseq data from Argelaguet et al. (2019) (see Section 5.3.1, **Experiments 7-9**). Despite the model mismatch, we observed good estimation performance for ρ_j across a range of simulation parameters (see Section A.1 Figures A.7-A.9).

For the remaining of SCRaPL's latent variables, the impact of investigated parameters varied. Accuracy increasing behaviour with larger cell numbers is also encountered in the remaining latent parameters of **Experiments 1,4,7**, as seen from Figures A.1a-A.1d, A.4a-A.4e and A.7c-A.7b. With increased number of cells, likelihood overwhelms prior pushing inferred parameters close to the data generating ones. Then ZI mainly impact gene expression latent mean and standard deviation degrading accuracy. Hence in **Experiments 2,5,8** DNAm parameters accuracy stays intact while gene expression violin plots widen as seen from Figures A.2a - A.2d, A.5a - A.5d and A.8a-A.8c. As ZI levels increase the error model is pushed to rely on decreasing number of non-zero readings, increasing the uncertainty around prediction accuracy. Similarly to ZI, DNAm coverage affects DNAm latent mean and standard deviation. This is observed in Figures A.3a-A.3e, A.6a-A.6e and A.9a-A.9c of **Experiments 3,6,9**. Increased DNA methylation coverage improves prediction accuracy, as it shrinks the dataset of possible posterior parameters closer to data generating ones.

5.4.2 SCRaPL improves the power to identify associations between molecular layers in mouse embryonic stem and brain cells

In mESC data we are investigating correlations between methylation for protein coding promoters within $\pm$2.5kbps from Transcription Start Site (TSS) and expression. The mEBC data set consists instead of accessibility (measured by ATAC seq) and expression data. In that case we quantify the associations between expression and accessibility of enhancers lying in a region of 12.5kbps form the gene (associations between expression and promoter accessibility for synthetic mESC data is shown in Section A.1). Importantly, the two data sets were obtained using technologies with widely differing technical characteristics: the plate-based scNMT platform returns good coverage levels (and hence low ZI) for a limited number of cells, while the 10x platform assays many more cells but with lower coverage and higher dropout rates. After quality control, the resulting data sets contained 9480 features (gene



(a) (b)

Fig. 5.7 **Plots summarizing differences in correlation estimation between SCRaPL, Spearman and Pearson in Experiment 1 with synthetic data across features**. (5.7a) Estimated correlation difference from true correlation as a function of cells for SCRaPL, Spearman and Pearson. (5.7b) Estimated correlation as a function of true correlation for SCRaPL, Spearman and Pearson in synthetic datasets with 300 genes and 1600 cells. Each dot represents a gene and is color-coded based inference approach.

promoters) and 679 mESCs and 4249 features (enhancers) and 4052 mEBCs respectively (Section 5.1).

Figure 5.8 shows the summary results of SCRaPL, Spearman and Pearson correlations. On the mESC data, SCRaPL retrieves approximately 2.5 times more associations compared to both Pearson and Spearman testing, retrieving 217 (SCRaPL) versus 68(Pearson)/85(Spearman) (Figure 5.8c and Table 5.4) associations. Figure 5.8a shows Bayesian Volcano plots, demonstrating how SCRaPL captures many more associations than Pearson and Spearman correlation. The overwhelming majorities of the associations recovered by frequentist methods are also captured by SCRaPL (Pearson and Spearman tend to be in very good agreement on this data set), which captures many more associations. We also looked at accessibility-expression pairs but due to weak signal no significant features were found by any of the methods. Later, we will investigate the biological significance of these results, showing how the greater statistical power of SCRaPL improves the ability to uncover new biological insights.

Figures 5.8b and 5.8d show the analogous results for the analysis of mEBC data. Here the picture is completely different: while SCRaPL still detects many more associations than Pearson, Spearman testing collapses and can only detect one significant associations. This points to a statistical vulnerability of Spearman testing when applied to data with high zero inflation in both molecular layers. More precisely, the large number of zeros present in both expression and accessibility mEBC data creates a large set of ranking ties, creating an

intrinsic mathematical problem for Spearman correlation. That is reflected in the 4180, 816 and 1 detected associations for SCRaPL, Pearson and Spearman (Figure 5.8d and Table 5.5). Unsurprisingly, both SCRaPL and, to a lesser extent, Pearson testing identify as significant a greater fraction of association pairs between accessibility and expression, as expected from the basic biology of gene expression. In particular, the overwhelming majority of correlations between proximal enhancers accessibility and gene expression were deemed to be significant by SCRaPL, reflecting the importance of proximal enhancers in the regulation of gene expression.

In summary, these results demonstrate that SCRaPL displays significantly increased statistical power in detecting associations between different molecular layers for both main types of multi-omics technological platforms. Intriguingly, both SCRaPL and Pearson testing appear to be largely insensitive to the type of technology, with SCRaPL identifying between 2.5 and 5 times more associations. Instead, Spearman testing reveals an intrinsic weakness in dealing with high sparsity data, making it potentially unsuitable as a tool for 10x multi-omics data analysis.

	Orig.	Neg.1	Neg.2	Neg.3	Neg.4	Neg.5
SCRaPL	217	2	2	8	3	0
Pearson	68	0	0	0	0	0
Spearman	85	0	0	0	0	0

Table 5.4 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in methylation-expression modalities of mESC data when** $\gamma = 0.205$

	Orig.	Neg.1	Neg.2	Neg.3	Neg.4	Neg.5
SCRaPL	4180	31	22	27	10	21
Pearson	816	0	0	0	0	0
Spearman	1	0	0	0	0	0

Table 5.5 **Number of detected features for original and negative control data with SCRaPL,Pearson and Spearman correlation in accessibility-expression modalities of mEBC data when** $\gamma = 0.145$

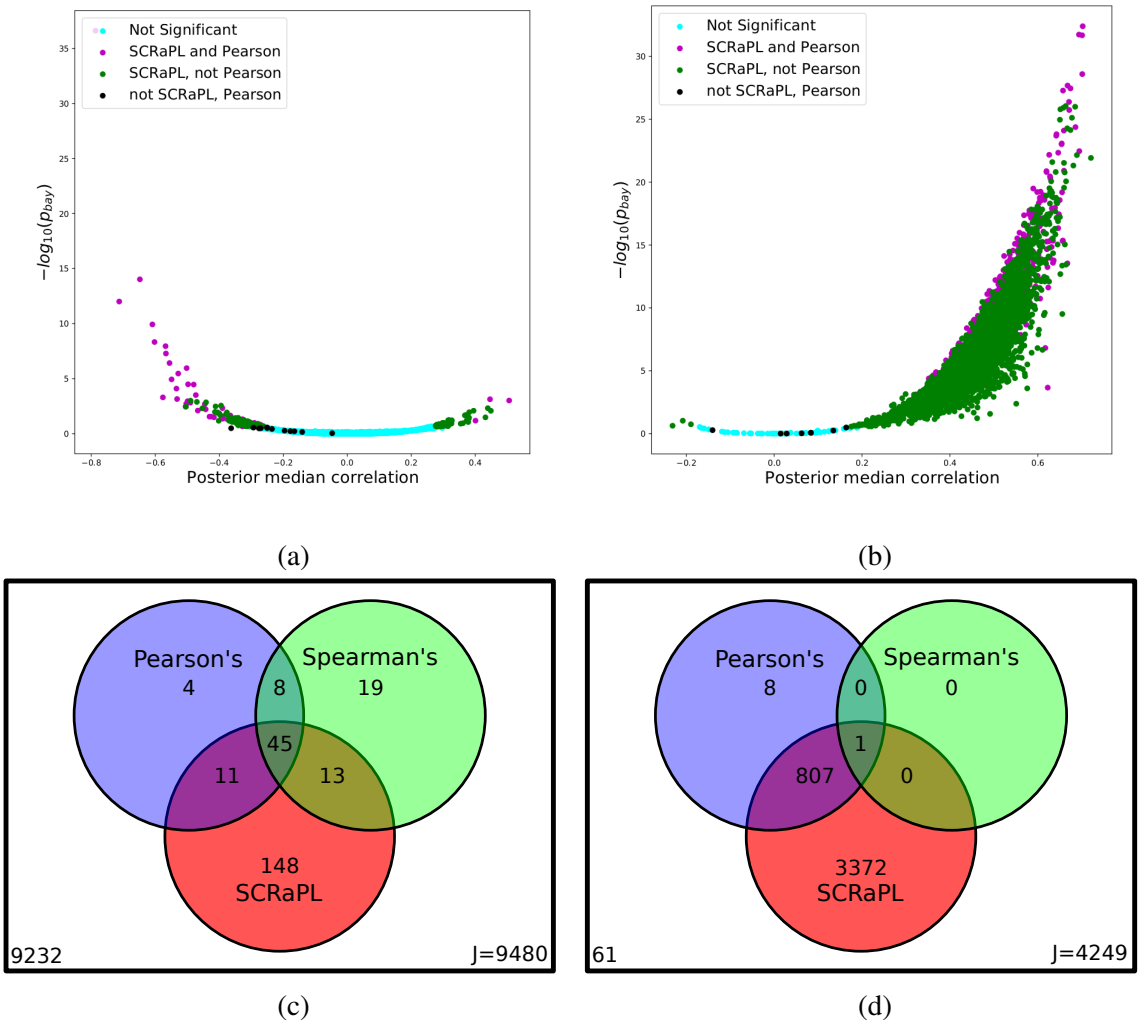


Fig. 5.8 **Summary of experiments on real data. Plots demonstrating that SCRaPL can significantly improve detection rates, by choosing features picked up by Pearson and Spearman correlation while avoiding spurious correlations.** (5.8a,5.8b) Bayesian volcano plots for mESC and mEBC data respectively. Scatter plot of posterior probability under the null hypothesis (in log scale) as a function of posterior median correlation. Each dot represents a feature and is marked with different color depending the method that labels it as a significant association. (5.8c,5.8d) Venn diagrams summarizing detection rates for SCRaPL, Pearson and Spearman in mESC and mEBC data. By accounting for different sources of noise it detects a large set of features identified by Pearson and Spearman correlation. SCRaPL also uncovers a additional large set that would be impossible for frequentist methods to identify in a robust way.

5.4.3 SCRaPL associations are influenced by data sparsity and are robust to outliers.

Figures 5.8a - 5.8d (and Figures A.11-A.13 in Section A.3) show clearly that, while most Pearson/ Spearman associations are also detected by SCRaPL, there are still some discrepancies. It is therefore natural to wonder to what extent the signals detected by the alternative methods are different, and what factors influence the different outcomes, in particular the much greater detection power obtained by SCRaPL.

From the modeling perspective, there are two major differences: first, SCRaPL considers noise models which capture overdispersion and take into account coverage in the epigenomic data. This should make SCRaPL associations less vulnerable to outlier values (eg. genes with low average expression with one or two high readings) or to epigenomic measurements with low coverage. Secondly, SCRaPL includes zero inflation in its accessibility/expression model, and can therefore attribute to that component some measurements of zero expression should the evidence dictate so. In the rest of this section, we present some empirical evidence that indeed supports the presence of these benefits in our real data analysis.

Features for which Pearson/Spearman testing and SCRaPL agree tend to have high coverage and small number of zeros in case of expression (or accessibility in 10X). An example feature called as significant by SCRaPL and Pearson is in Figure 5.9a.

To gain more insight on the factors driving SCRaPL inferences it is interesting to focus on associations, whose significance differs between the two methods. An example of an association detected by Pearson/Spearman testing but not SCRaPL is shown in Figure 5.9b. As we can see, we have a large fraction of zero expression values with very low methylation coverage. As a result, SCRaPL, while placing most of the posterior mass over negative correlation values, cannot confidently exclude the possibility of no correlation. This example perfectly illustrates that divergences between SCRaPL and Pearson/Spearman testing are often driven not by expected values, but by the fact that SCRaPL additionally performs uncertainty quantification on its results.

An example of an association deemed significant by SCRaPL, but not by Pearson/Spearman testing, is shown in Figure 5.9c. In this case, we tend to have medium to high expression and good coverage. However, Pearson/Spearman correlation remain below detection levels due to a number of observations with zero expression. This is an example where SCRaPL can be particularly beneficial, since the noise model can better capture potential effect of zero inflation.

To provide a more quantitative, global explanation of the differences between SCRaPL and Pearson, we regress the absolute difference in inferred correlation against methylation coverage and percentage of zero counts for each feature across all cells. The resulting

regressions, shown in Figure 5.9d, demonstrate a weak but consistent effect of both forms on noise, confirming that differences between the two methods are more prominent in noisier situations where methylation coverage is low or sparsity is high. This analysis is also confirmed for Pearson and Spearman in both mESC and mEBC data in Figure A.20 of Section A.6.

5.4.4 SCRaPL identifies biologically meaningful epigenetic regulation in early mouse gastrulation

To provide further biological support to SCRaPL associations, we explored the features identified to have statistically significant associations between DNAm and gene expression during the early stages of gastrulation.

We start by choosing early pluripotency and germ cell markers where methylation's strong repressive role is widely investigated (e.g. Argelaguet et al. (2019); Eckersley-Maslin et al. (2019); Hu et al. (2008)). Developmental pluripotency markers (i.e. Dppa2, Dppa4, Dppa5a) exhibit strong regulatory patterns with the generally high expression levels in days 4.5 and 5.5 being gradually suppressed as cells diversity to progenitors of major organs (Figures A.14a, A.14c, A.14d in Section A.3.1). Methylation's strong silencing role was also found in Dnmt3l (Figure A.15a in Section A.3.1), a catalytically inactive DNA methyltransferase that cooperates with Dnmt3a and Dnmt3b to methylate DNA (Neri et al., 2013). In addition, our analysis identified a series of genes with strong regulatory action vital to embryonic development Atp6v0d1 (Miura et al., 2003), to spermatogenesis/placenta-supported development Tex19.1 (Tarabay et al., 2013) and others with unknown roles like Zfp981 and Trap1a (Figures A.16a, A.16b, A.16c in Section A.3.1).

To complete the exploratory analysis, we look at Gene Set Enrichment Analysis (GSEA) using DAVID (Huang, 2009) to establish links with biological phenomena observed in early embryogenesis and gene promoter methylation. To identify the processes we allow a minimum of 7 genes, p-values up to 0.3 and sort them based on their enrichment score. As a result we have identified a total of fifteen developmental and house-keeping processes (see Figure A.17 in Section A.4). The highest enrichment scores are encountered for angiogenesis and in utero embryonic development with 2.6 and 2.1 respectively. For house-keeping processes we get proteolysis, ion transport and negative regulation of transcription with enrichment scores 2.2,1.9 and 1.7 respectively. Using the same filtering parameters in DAVID with the set of genes detected by Pearson we would recover a single process, regulation of transcription with enrichment score 1.5. Spearman testing detects a larger number of associations than Pearson on the ESC data set, and consequently has increased

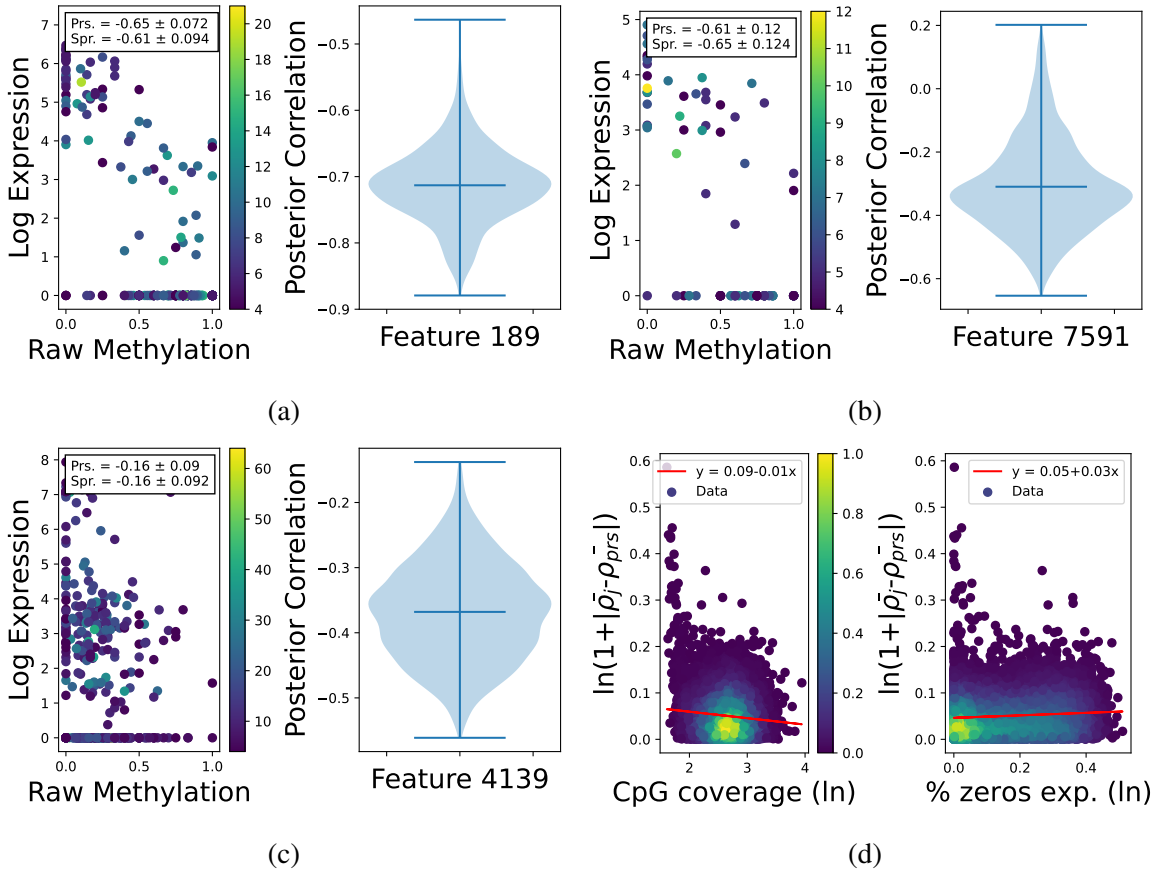


Fig. 5.9 **SCRaPL's behavior compared to Pearson/Spearman correlation in micro and macro scale. Specific features demonstrate examples where frequentist approaches agree and disagree with SCRaPL's estimates. The scatter plots show the link between coverage and zero inflation and distance between frequentist and SCRaPL correlation.**. In all figures apart from 5.9d the scatter plot depicts raw data for chosen features color-coded by CpG coverage, and normalized expression plotted in the $log(1+x)$ scale. The violin plots depict the posterior correlation densities estimated by SCRaPL for the raw data in their left hand side. (5.9a) Example where both SCRaPL and Pearson/Spearman identify the feature's association as significant. (5.9b) Example were only Pearson/Spearman identifies the feature's association significant. (5.9c) Example were only SCRaPL identifies the feature's association significant. (5.9d) Scatter plots to demonstrate the negative/positive relationship between alternative correlation estimates and CpG coverage/% zeros in expression respectively. ($\bar{\rho}_j$ and ρ_{prs} in figure 5.9d are posterior mean and Pearson correlation for feature j.)

power in detecting enriched processes. In this case the number of recovered processes increases to 7, which consist however of primarily house-keeping processes (see Figure A.18 in Section A.4).

This analysis confirms the biological plausibility of the identified SCRaPL associations. It should be emphasised that the enrichment analysis has only been possible due to the larger number of associations identified by SCRaPL: GSEA analyses require considerable numbers of genes to identify any significant enrichment. This underlines the fact that technical variability not only erodes correlation but significantly under-powers downstream exploratory analysis in multi-omics data. Hence by modelling data generative processes we can increase substantially the scope of downstream interpretative analyses of single-cell multi-omics data.

5.5 Discussion

Single cell multi-omics sequencing technologies are rapidly becoming an important tool to understand epigenetic regulation for individual cells in complex biological processes, such as early embryo development. However, analysis of such data still presents a major bottleneck, due to the high-dimensionality, sparsity and heterogeneous noise affecting them.

In this chapter, we argued that the introduction of noise-aware approaches is fundamental in developing the field of single-cell multi-omics. We introduced SCRaPL, a Bayesian approach to perform perhaps the most basic and common multi-omics analysis, the discovery of correlative associations between different data modalities. By employing dedicated noise models in a latent-Gaussian framework, SCRaPL achieves more powerful and more robust results than simple analyses based on Pearson correlation, which is by far the most widespread tool currently used. As with most Bayesian methods, SCRaPL does suffer from a higher computational burden, particularly when compared with extremely simple analyses, such as Pearson correlation. Extension of noise-aware Bayesian methods to different single-cell multi-omics analyses may therefore require the adoption and evaluation of more efficient computational inference techniques, such as variational inference (Kingma and Welling, 2014).

Our analyses were based on existing annotation, where the expression of a given gene was correlated with epigenetic data from a nearby genomic region (promoters or nearby enhancers). This appears to be a reasonable demonstration of the tool, although it clearly limits the scope for discovery of interesting biological processes such as distal regulation. It should be pointed out that SCRaPL could also be used to test associations between unannotated regions along the lines explored in e.g. Kapourani et al. (2021), Cusanovich et al. (2018).

Understanding the impact of noise in correlation is challenging. There are methods addressing the problem both at a population and at a feature specific level. For example, Seurat's (Hao et al., 2021) CCA approach looks at noise from a population perspective. It

works by averaging epigenomic and trascriptomic noise during dimensionality reduction creating correlated embeddings. To improve our understanding towards that direction, we employed SCRaPL's Bayesian framework for quantifying uncertainty around correlative analyses. Our experiments comparing the two methods indicate similar noise filtering results. However, future analysis could thoroughly investigate the precise connections between these approaches and identify which one is more suitable for specific problems.

Chapter 6

scMVI: A deep generative model for integrated analysis of single-cell multiomics data

In Chapter 5 we introduced SCRaPL, a Bayesian Hierarchical model designed to handle noise in epigenomic versus transcriptomic correlative analyses. The robust noise treatment improved detection rates of statistically significant associations, allowing for better quantification of the regulatory landscape compared to traditional alternatives like Pearson or Spearman correlations. As demonstrated by SCRaPL, underlying assumptions such as the independence across genomic regions yield scalable statistical models which work well for correlation inference. In reality, such assumptions are at odds with complex regulatory mechanisms, employing large sets of genes. Hence, to create models that are more appropriate to describe such phenomena, simplistic independence postulates need to be abandoned.

As we saw in Section 2.3.1, single cell multi-omics platforms have revolutionised the way we interrogate interactions between epigenome and transcriptome. Single cell multi-omics protocols rely on a series of approaches, each with its advantages compared to the rest. Despite their immense potential, common limitations like the small amount of starting DNA, low protocol sensitivity and technical artifacts introduced during sequencing (Muskovic and Powell, 2021; Schwartzman and Tanay, 2015), create datasets with large number of zeros with many cells failing to pass quality control steps. Often this forces practitioners to discard significant data subsets, making subsequent analysis more challenging.

In this chapter, we introduce single cell Multi View Inference (scMVI), a deep generative probabilistic model designed to accommodate analyses on partially and fully observed single cell multi-omics data. Using single cell multi-omics data, scMVI offers an alternative

approach for aligning latent omics representations based on Jeffrey's divergence [1]. In the case of partially observed data, this allows to fill in missing modalities either by imputing or by matching values across similar cells, effectively providing transcriptomic and epigenomic layer integration. Finally, scMVI's robust noise treatment makes it an effective tool for downstream analysis.

6.1 Datasets

We consider two single cell multi-omics datasets. The first dataset was generated using the 10X Genomics Single Cell Multiome ATAC plus Gene Expression platform and consists of Peripheral Blood Mononuclear Cells (PBMC) of a healthy female donor aged 25 [2]. The second dataset consists of mouse Embryonic Stem Cells (mESC_Ag) at various developmental stages (embryonic days 4.5, 5.5, 6.5, 7.5) comprising the exit from pluripotency to lineage specification (Argelaguet et al., 2019) and was generated using scNMT-seq (Clark et al., 2018). As part of some experiments, we also consider a single-omics example with scRNA-seq data generated by the 10x Genomics Chromium Single Cell Gene Expression platform. Samples correspond to mouse Embryonic Stem Cells (mESC_Pj) at embryonic days 6.5, 6.75, 7.0, 7.5, 7.75, 8.0 and 8.5 (Pijuan-Sala et al., 2019). For the mESC_Pj and PBMC datasets, our analysis focuses on comparing scMVI with related methods, while the mESC_Ag dataset is used to demonstrate that scMVI can achieve state of the art performance even in cases where few observations are available.

Effective data preprocessing and quality control techniques are important to subsequent downstream analysis. To preprocess chromatin accessibility and gene expression in PBMC, a mixture of Seurat (Hao et al., 2021) and scran were used (Lun et al., 2016). Low quality cells with library size less than 3 absolute deviations below the median were subsequently removed. Then, using Seurat's built-in library and a 2kbps window size, we aggregated accessible peaks around Transcription Start Site (TSS) [3]. For the mESC_Ag dataset, fragments were mapped to a reference genome using BPRMeth (Kapourani and Sanguinetti, 2018) with window size of $\pm$2.5kbps and $\pm$0.25kpbs for DNA methylation (DNAm) and chromatin accessibility, respectively. As above, the expression data were filtered using library size and offsets for the gene expression model were inferred with scran (Lun et al., 2016). Detailed analysis of the

[1]During the development of scMVI, a preprint of MultiVI (Ashuach et al., 2021) was uploaded onto the bioarxiv repository which also uses Jeffrey's divergences.

[2]More information can be found in https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k.

[3]This preprocessing steps are thoroughly explained by Seurat's authors in https://satijalab.org/seurat/articles/atacseq_integration_vignette.html.

preprocessing steps for this dataset can be found in Section 5.1. The preprocessing of the mESC_Pj data had largely been done by the authors (Pijuan-Sala et al., 2019).

Subsequently, all three datasets passed through Quality Control (QC). In mESC_Ag genes with more than 65% zeros in DNAm, gene expression and chromatin accessibility were discarded, creating a dataset with $\sim$ 6000 genes and $\sim$ 800 cells. For DNAm data, we have the number of captured CpGs in the 5 kbp window and how many of them were methylated in gene promoter region. Similarly for chromatin accessibility we have the number of captured GpCs in the 0.5 kbp window and how many of them were methylated in gene promoter region. For gene expression we have the count of captured mRNAs for each genomic region. To make mESC_Pj and mESC_Ag comparable, genes not appearing in the latter were removed from the former, creating a dataset with $\sim$ 6000 genes and $\sim$ 52500 cells. For mESC_Pj which has only gene expression data we have the count of captured mRNAs for each genomic region. For PBMC cells, similar QC steps to Minoura et al. (2021) were applied, keeping 5000 and 10000 most variable genes and peaks. Here, for chromatin accessibility we have the number of accessible peaks in a 2 kbp window around TSS. For gene expression we have the count of captured mRNAs for each genomic region.

The resulting dataset contained $\sim$ 10000 cells. A summary of datasets is presented in Table 6.1.

Dataset	Protocol	# Cells	#Features	DNAm	Acc.	Expr.	Expr. Covg
PBMC	Droplet	$\sim$10000	$\sim$10000 (Acc.) / $\sim$5000 (Expr.)	-	+	+	0.22
mESC_Ag	Plate	$\sim$800	$\sim$6000 per layer	+	+	+	89.64
mESC_Pij	Droplet	$\sim$52000	$\sim$6000	-	-	+	1.91

Table 6.1 **Table summarizing data used in Chapter 6.** Table giving a synopsis of the data used for this study. Columns 2-4 describe the sequencing protocol, the number of cells and the number of features per modality for each of the datasets used. Then columns 5-7 present the available data modalities and column 8 shows the coverage using average expression across genes and cells in each dataset.

6.2 Methods

scMVI is a deep probabilistic model designed to analyze partially and fully observed single cell multi-omics data, gathered with modern plate and droplet based protocols. To motivate scMVI, we focus on epigenomic and transcriptomic datasets with two or three molecular layers. However, scMVI's flexible design allows for seamless handling of protocols targeting different modalities with the appropriate likelihood modifications or the adaption to future protocols scaling to more molecular marks.

Low amount of input DNA combined with limited sensitivity of multi-omics sequencing protocols often results to the partial or full data dismissal, under-powering downstream analysis. A solution could be to post-experimentally address protocol limitations through computational strategies. scMVI introduces a deep hierarchical framework (Figure 6.1), designed to impute (Figure 6.2) and integrate (by identifying matches between) partially observed readings, minimizing the impact of the outlined problems.

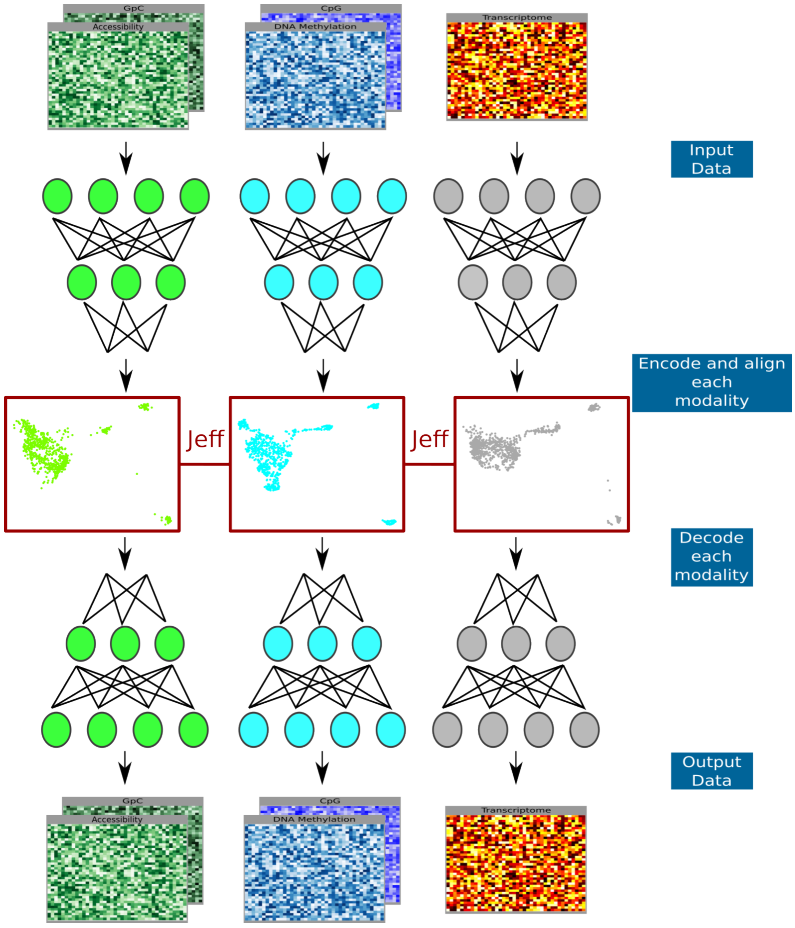


Fig. 6.1 **scMVI schematic representation. This is to demonstrate the how the model looks like.** scMVI is a deep generative model designed to handle partially and fully observed single cell multi-omics data. DNAm, gene expression and chromatin accessibility are non-linearly encoded to modality dependent low dimensional manifolds. During training, the manifolds are aligned using Jeffrey's divergence. In the generative step, a modality specific latent representation and decoder is used to reconstruct observations. This graph is given, as the way scMVI is defined and trained has some differences to traditional VAEs.

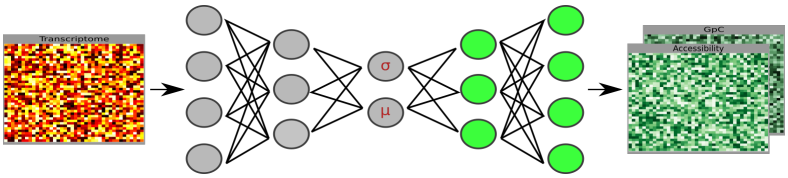


Fig. 6.2 **scMVI cross modal schematic representation. This is to demonstrate how scMVI imputes missing profiles.** Schematic depiction of cross modal imputation. Using relevant encoder(s), we project data to create a latent space representation, which is then used to impute missing data with the corresponding decoder.

6.2.1 scMVI: Single Cell Multi View Inference

In a dataset with J_r distinct mRNAs from DNA coding regions, DNAm recordings from J_m cis-regulatory regions, chromatin accessibility from J_a regions and I cells, scMVI models the generative process of the epigenetic and transcriptomic molecular layers using separate Variational Autoencoders (VAEs) (Kingma and Welling, 2014) [4]. Since our primary example is the mESC_Ag dataset where J_r, J_m, J_a are equal, we replace them with J to simplify the notation, but in general they can attain different values. Apart from epigenomic and transcriptomic layers in plate and droplet based protocols, scMVI can be extended to other multi-omics protocols (eg. gene expression-protein) with the appropriate likelihood modifications.

For a cell with index $i \in I$, we denote gene expression, DNAm and chromatin accessibility with vector variables $\mathbf{Y}_{ri}$, $\mathbf{Y}_{mi}$ and $\mathbf{Y}_{ai}$ respectively where $\mathbf{Y}_{ri} = \left[Y_{ri1}, \ldots, Y_{rij}, \ldots Y_{riJ}\right]$, $\mathbf{Y}_{mi} = \left[Y_{mi1}, \ldots, Y_{mij}, \ldots Y_{miJ}\right]$ and $\mathbf{Y}_{ai} = \left[Y_{ai1}, \ldots, Y_{aij}, \ldots Y_{aiJ}\right]$. Each encoder estimates a mapping between molecular layers ($\mathbf{Y}_{ri}$, $\mathbf{Y}_{mi}$,$\mathbf{Y}_{ai}$) and their corresponding latent cell states ($\mathbf{Z}_{ri}$, $\mathbf{Z}_{mi}$,$\mathbf{Z}_{ai}$). By design, latent representations are modeled to follow multivariate isotropic Gaussian distributions $\mathbb{P}(\mathbf{Z}_{ri}|\mathbf{Y}_{ri})$, $\mathbb{P}(\mathbf{Z}_{mi}|\mathbf{Y}_{mi})$ and $\mathbb{P}(\mathbf{Z}_{ai}|\mathbf{Y}_{ai})$, respectively. To ensure consistent latent representations from different molecular modalities, distances between $\mathbb{P}(\mathbf{Z}_{ri}|\mathbf{Y}_{ri})$, $\mathbb{P}(\mathbf{Z}_{mi}|\mathbf{Y}_{mi})$ and $\mathbb{P}(\mathbf{Z}_{ai}|\mathbf{Y}_{ai})$ are penalized in pairs using Jeffrey's divergence. Jeffre's (or symmetrized KL) divergence is used as it is symmetric and can ensure that different latent molecular representations end up as close as possible. Moreover, each encoder is regularized by minimizing the KL divergence with respect to a standard multivariate normal distribution.

Gene expression, DNAm and chromatin accessibility are generated by decoding the corresponding latent space samples. More precisely, samples $\mathbf{Z}_{ri}$, $\mathbf{Z}_{mi}$ and $\mathbf{Z}_{ai}$ are fed to modality specific decoders to generate likelihood parameters. Similarly to SCRaPL, the noise models are tailored to the analysed assays. More specifically, chromatin accessibility

[4]For more information regarding VAEs please refer to Section 3.5.1.

data are handled using Binomial or Zero Inflated Negative Binomial (ZINB) likelihoods for plate or droplet based sourcing protocol respectively. DNAm observations are analyzed using Binomial likelihood and gene expression with ZINB. scMVI's Graphical representation is summarized in Figure 6.3.

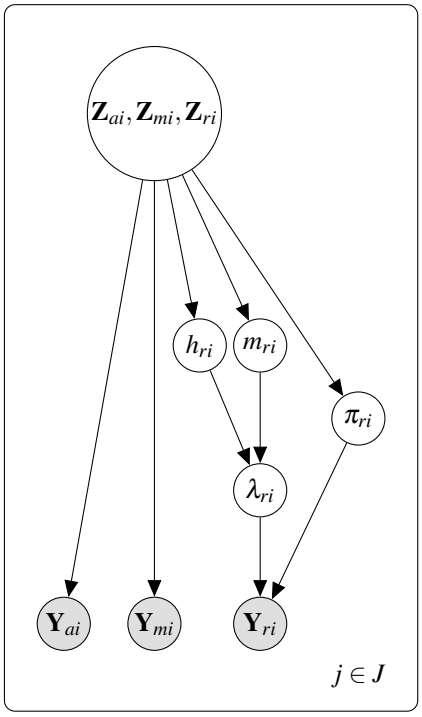


Fig. 6.3 **scMVI's graphical model for mESC data.** For each cell $i \in I$, scMVI learns a forward and reverse mapping between molecular layers ($\mathbf{Y}_{ri}$ is Gene expression; $\mathbf{Y}_{mi}$ is DNAm; $\mathbf{Y}_{ai}$ is Chromatin accessibility) and their associated latent parameters ($\mathbf{Z}_{ri}, \mathbf{Z}_{mi}, \mathbf{Z}_{ai}$). The flexible design assuming one VAE per molecular layer allows for accurate imputation and integration even in cases with partially observed data. Depending the sourcing protocol VAEs could be added or removed, from scMVI, to accommodate imputation and integration.

RNA expression noise model. Let $\mathbf{Y}_{ri}$ be the vector of expression read counts observed for each cell $i \in I$ in the gene expression modality. Conditioned on the latent expression vector $\mathbf{Z}_{ri}$ for gene $j \in J$ and cell i, we assume that

$$\mathbb{P}\left(\mathbf{Y}_{rij} = y_{rij} \,|\, \mathbf{Z}_{ri} = z_{ri}, s_{ri}\right) = \begin{cases} \left(1-\pi_{rij}\right) \text{Poisson}\left(s_{ri}\lambda_{rij}\right) & \textbf{if } y_{rij} > 0, \\ \pi_{rij} + \left(1-\pi_{rij}\right) \text{Poisson}\left(s_{ri}\lambda_{rij}\right) & \textbf{if } y_{rij} = 0. \end{cases} \tag{6.1}$$

where

$$\lambda_{rij} \sim \text{Gamma}\left(h_{rij}, \frac{h_{rij}}{m_{rij}}\right) \tag{6.2}$$

and

$$\pi_{rij} = f_{rj}^{\pi}(z_{ri}), h_{rij} = f_{rj}^{h}(z_{ri}), m_{rij} = f_{rj}^{m}(z_{ri}), \tag{6.3}$$

where f_r^{π}, f_r^{h} and f_r^{m} are model decoders sharing all inner layers with:

$$f_r^{\pi}(\cdot) = \left[f_{r1}^{\pi}(\cdot), \ldots, f_{r|J|}^{\pi}(\cdot)\right], f_{rj}^{h}(\cdot) = \left[f_{r1}^{h}(\cdot), \ldots, f_{r|J|}^{h}(\cdot)\right], f_{rj}^{m}(\cdot) = \left[f_{r1}^{m}(\cdot), \ldots, f_{r|J|}^{m}(\cdot)\right], \tag{6.4}$$

This error model formulation corresponds to a ZINB where s_{ri} (> 0) is a cell specific scaling factor accounting for global differences across cells (e.g. due to sequencing depth), λ_{ri} (> 0) represents the normalized feature expression vector , h_{rij} (> 0) captures cell-cell unexplained technical variability for gene j, m_{rij} (> 0) is the expected normalized feature expression and π_{rij} ($\in [0,1]$) represents zero-inflation parameter for gene j allowing the likelihood to account for excess zeros (if $\pi_{rij} = 0$, (6.1) reduces to a Negative Binomial model). In practice, we infer scaling factors s_{ri} using scran (Lun et al., 2016) and use them as known model offsets and the remaining parameters are estimated from functions of equation 6.3 by learning neural network weights during training. Similarly to previous section we chose a model that can handle count data. However, this time instead of going with a ZINF Poisson log-normal we chose the ZINF Poisson gamma distribution to make our results as comparable as possible to related methods.

DNAm noise model. Let the DNAm vector $\mathbf{Y}_{mi}$ in cell $i \in I$ correspond to the number of methylated CpG cites and vector N_{mi} captured CpG sites within a pre-specified genomic region (e.g. gene promoter)[5]. Then, conditioned on latent methylation vectors z_{mi}, N_{mi} and CpG methylation levels are modeled as follows:

$$\mathbb{P}(\mathbf{Y}_{mi} = y_{mi} \,|\, \mathbf{Z}_{mi} = z_{mi}, N_{mi}) = \text{Binomial}(N_{mi}, f_m(z_{mi})), \tag{6.5}$$

where $f_m(\cdot)$ and z_{mi} denote the methylation decoder and latent space sample for cell i respectively [6].

Chromatin accessibility noise model. For chromatin accessibility, we use data generated by scATAC-seq and scNONe-seq technologies, so our modeling practices closely resemble what was done in SCRaPL. If cells are sequenced with scATAC-seq, then vector $\mathbf{Y}_{ai}$ in cell $i \in I$ denotes open peaks across genomic regions and follows a ZINB as described for gene expression noise model. In case of scNOMe-seq, $\mathbf{Y}_{ai}$ describes the number of methylated

[5] N_{mi} representes coverage level for DNAm data.

[6] In Equation 6.5 we use vector notation to denote the model that applies to each element of the associated vector. We abuse notation for simplicity.

GpC sites across the pre-specified genomic regions, where N_{ai} CpG[7] sites were captured and we resort to the same Binomial likelihood as described in DNAm noise model.

Objective function. For a batch of cells $B \in I$ in the training set, scMVI's objective function is computed as follows:

$$\sum_{i \in B} -\ln\left[\mathbb{P}\left(\mathbf{Y}_{mi} = y_{mi} \,|\, \mathbf{Z}_{mi} = z_{mi}, N_{mi}\right)\right] - \ln\left[\mathbb{P}\left(\mathbf{Y}_{ai} = y_{ai} \,|\, \mathbf{Z}_{ai} = z_{ai}, N_{ai}\right)\right] - \ln\left[\mathbb{P}\left(\mathbf{Y}_{ri} = y_{ri} \,|\, \mathbf{Z}_{ri} = z_{ri}, s_{ri}\right)\right] + \mathrm{Jeff}\left[\mathbb{P}\left(\mathbf{Z}_{ri} = z_{ri} | \mathbf{Y}_{ri} = y_{ri}\right) \| \mathbb{P}\left(\mathbf{Z}_{ai} = z_{ai} | \mathbf{Y}_{ai} = y_{ai}\right)\right] + \mathrm{Jeff}\left[\mathbb{P}\left(\mathbf{Z}_{ri} = z_{ri} | \mathbf{Y}_{ri} = y_{ri}\right) \| \mathbb{P}\left(\mathbf{Z}_{mi} = z_{mi} | \mathbf{Y}_{mi} = y_{mi}\right)\right] + \mathrm{KL}\left[\mathbb{P}\left(\mathbf{Z}_{ri} = z_{ri} | \mathbf{Y}_{ri} = y_{ri}\right) \| \mathbb{N}\left(0, I\right)\right] + + \mathrm{KL}\left[\mathbb{P}\left(\mathbf{Z}_{ai} = z_{ai} | \mathbf{Y}_{ai} = y_{ai}\right) \| \mathbb{N}\left(0, I\right)\right] + \mathrm{KL}\left[\mathbb{P}\left(\mathbf{Z}_{mi} = z_{mi} | \mathbf{Y}_{mi} = y_{mi}\right) \| \mathbb{N}\left(0, I\right)\right] \tag{6.6}$$

where $\mathrm{Jeff}(\cdot)$, $\mathrm{KL}(\cdot)$ denote Jeffrey's and KL divergence respectively, and $\mathbb{N}(0, I)$ denotes the standard isotropic normal. Depending on the training data, unrelated terms are omitted (eg. for PBMC data, all terms involving an index m are dropped) from equation 6.6.

The objective function in equation 6.6 is used to jointly train three distinct VAEs with an additional Jeffrey's divergence term use to penalize distant latent representation from the same cell. The part of the objective function training the gene expression VAE is presented in equation 6.7.

$$\sum_{i \in B} -\ln\left[\mathbb{P}\left(\mathbf{Y}_{ri} = y_{ri} \,|\, \mathbf{Z}_{ri} = z_{ri}, s_{ri}\right)\right] + \mathrm{KL}\left[\mathbb{P}\left(\mathbf{Z}_{ri} = z_{ri} | \mathbf{Y}_{ri} = y_{ri}\right) \| \mathbb{N}\left(0, I\right)\right] \tag{6.7}$$

The part of the objective function training the chromatin accessibility VAE is presented in equation 6.8.

$$\sum_{i \in B} -\ln\left[\mathbb{P}\left(\mathbf{Y}_{ai} = y_{ai} \,|\, \mathbf{Z}_{ai} = z_{ai}, N_{ai}\right)\right] + \mathrm{KL}\left[\mathbb{P}\left(\mathbf{Z}_{ai} = z_{ai} | \mathbf{Y}_{ai} = y_{ai}\right) \| \mathbb{N}\left(0, I\right)\right] \tag{6.8}$$

The part of the objective function training the DNAm VAE is presented in equation 6.9.

$$\sum_{i \in B} -\ln\left[\mathbb{P}\left(\mathbf{Y}_{mi} = y_{mi}, \,|\, \mathbf{Z}_{mi} = z_{mi}, N_{mi}\right)\right] + \mathrm{KL}\left[\mathbb{P}\left(\mathbf{Z}_{mi} = z_{mi} | \mathbf{Y}_{mi} = y_{mi}\right) \| \mathbb{N}\left(0, I\right)\right] \tag{6.9}$$

[7]Similarly to DNAm data N_{ai} refers to coverage levels for chromatin accessibility in plate based protocols.

To ensure consistent representations between alternative low dimensional cell representations estimated from each molecular layer we also used the terms described in equation 6.10.

$$\begin{aligned}\sum_{i\in B} \text{Jeff}\left[\mathbb{P}\left(\mathbf{Z}_{ri}=z_{ri}|\mathbf{Y}_{ri}=y_{ri}\right)\|\mathbb{P}\left(\mathbf{Z}_{ai}=z_{ai}|\mathbf{Y}_{ai}=y_{ai}\right)\right]+\\ \text{Jeff}\left[\mathbb{P}\left(\mathbf{Z}_{ri}=z_{ri}|\mathbf{Y}_{ri}=y_{ri}\right)\|\mathbb{P}\left(\mathbf{Z}_{mi}=z_{mi}|\mathbf{Y}_{mi}=y_{mi}\right)\right]\end{aligned} \tag{6.10}$$

The two Jeffrey's divergences ensure that latent representations $\mathbb{P}\left(\mathbf{Z}_{ri}=z_{ri}|\mathbf{Y}_{ri}=y_{ri}\right)$ is as close as possible to manifolds defined by $\mathbb{P}\left(\mathbf{Z}_{mi}=z_{mi}|\mathbf{Y}_{mi}=y_{mi}\right)$ and $\mathbb{P}\left(\mathbf{Z}_{ai}=z_{ai}|\mathbf{Y}_{ai}=y_{ai}\right)$ and visa versa.

Upon successful convergence, the objective function in equation 6.6 achieves three goals: (i) it optimizes parameters such that DNAm, gene expression and chromatin accessibility conditional likelihoods are maximized; (ii) the Jeffrey's divergence terms ensure that all three modalities construct a consistent and meaningful joint manifold [8], even when some molecular layers are missing and (iii) the KL divergence terms further regularize the network to prevent over-fitting.

Model architecture. scMVI consists of three VAEs. The encoders map high dimensional epigenetic and transcriptomic samples, to a low latent space (in $\mathbb{R}^{10}$). Latent space dimensionality was chosen to facilitate comparisons with similar methods like MultiVI (Ashuach et al., 2021), scMM (Minoura et al., 2021) and scMVAE (Zuo and Chen, 2021). Similarly, the decoders follow the reverse order reconstructing likelihood parameters. Each encoder/decoder contains building blocks made of a dense layer with a bias term, a LayerNormalization [9], a ReLU activation function and a dropout with rate 0.1 in the described order. The only exception is in last hidden layer of the encoders and decoders (i.e. last layer of encoder and decoder, to latent Gaussians and unconstrained likelihood parameters), where only a dense layer is used. The encoder and decoder hidden dimensions are 200, 100, 100 and 10 totaling approximately 12 million parameters for the entire model[10]. Since our largest training set contains $\sim$ 36000 cells while the network several million parameters, strong layer and objective function regularization described above is vital to avoid over-fitting.

Implementation. Single cell multi-omics datasets were randomly split into training (77% of cells) and testing data (23% of cells). We omit the use of validation data as we used a fixed architecture and did not perform any hyper-parameter optimization. To train scMVI we use Adam optimizer (Kingma and Ba, 2015) with step and batch size equal to 10^{-3} and 10 respectively. The total number of allowed epochs was 250. Due to the 10 epoch patience

[8] In case of three molecular layers, Jeffrey's divergence is estimated in pairs.

[9] This layer normalizes data across sample dimension. More information can be found in (Ba et al., 2016)

[10] Related deep learning methods like scMM and scMVAE use up to 60 million parameters.

early stopping strategy [11] on the test set loss, scMVI did not train for more than 90 epochs (maximum 10 minutes) in both plate and droplet based data. A Tensorflow implementation of scMVI is available at https://github.com/chrmaniatis/scMVI.

6.2.2 Related Methods

Since the introductions of the first single cell sequencing protocol by Tang et al. (2009), various sequencing techniques have emerged aiming to quantify single and multiple omics layers at a single cell resolution. Together with sequencing technologies, multiple methods addressing related limitations have appeared with attention recently shifting to VAEs (i.e. MultiVI, scMM, scMVAE and Cobolt (Gong et al., 2021)) capable of handling partially observed data. The ways in which such methods differ are in terms of their noise models, the problem they target and manifold building strategy.

Noise models. The rapid expansion of droplet based sequencing protocols have generated large scale multi-omics datasets, allowing to train reasonably complex deep generative models. The most popular datasets combine proteomics with transcriptomics or chromatin accessibility with transcriptomics (i.e. (Stoeckius et al., 2017) or (Cao et al., 2018) respectively). As these multi-omics datasets consist of counts, the most popular error model for analyzing both pairs is ZINB. MultiVI differs from related methods as it models chromatin accessibility data using Bernoulli likelihood. Furthermore, MultiVI offers more flexibility as it allows the use of Negative Binomial and Poisson likelihoods to analyze gene expression data.

Addressed problem. Multi-omics protocols try to extract information for multiple molecular layers, amplifying poor coverage problems related to low starting DNA amounts. In contrast, post-sequencing integration of individually measured molecular layers keeps sparsity issues to a minimum as it works with single omics layers. The methods reviewed here place much emphasis on addressing protocol limitations post-sequentially, either by integrating accessibility with gene expression layers of partially observed data or by imputing missing values. The method that stands out is MultiVI which also corrects for batch effects. The difference of MultiVi compared to the rest is that it fits conditional variational autoencoders with batch labels being auxiliary covariates. The difference to traditional VAEs is that the input label is augmented using batch label infomration. This has the advantage of considering

[11] Early stopping strategy is a common form of regularization used to avoid overfitting. During training validation set loss is monitored to detect overfitting signs. When validation loss starts to increase training is stopped and model parameters are reverted to the state with the smallest generalization error. For more information please look Yao et al. (2007).

more diverse data, but there is an added complexity of accurately choosing batch label when integrating or imputing values in previously unseen data.

Manifold building strategy. As we are dealing with models that separately encode each modality into a low dimensional representation, the decoding task can be interpreted as experts' prediction task. Therefore, there is a need of combining and decoding multiple predictors. Here we outline a series of expert combining strategies in order to highlight that what we do is is slightly different. Handling partially missing multi-modal data requires an elaborate strategy for building the latent space manifold. In this aspect, almost all methods offer a unique perspective. We begin with Cobolt, where they combine encoder predictions into a unique point in the manifold using a Product of Experts (PoE)(Hinton, 2002) strategy. Essentially, this combines the alternative realizations of the same point using the product distribution. Then we have the Mixture of Experts (MoE) (Jacobs et al., 1991) approaches described in scMM and scMVAE, which partitions the latent space. MoE uses a divide and conquer training strategy, forcing each encoder to capture unique features. Predictions are then merged creating a richer representation. The differences between scMM and scMVAE are in the objective function and the trust they place in latent representations coming from different omics layers. In scMM they set weights to be equal while scMVAE infers the weights from the data.

scMVI compared to related methods. Comparing scMVI with the methods summarised above, we identify both novelties and similarities. Related methods chose to examine larger and more homogeneous datasets depicting static phenomena such as chromatin accessibility and gene expression in peripheral blood cells. In contrast, the types of data we analyzed notably vary in size, capturing technology and biology. Apart from peripheral blood cells, scMVI considered droplet and plate based sequenced mouse embryonic stem cells in various stages between pluripotency and gastrulation. The manifold building strategy we chose closely resembles the one presented in MultiVI used to penalize the model if individual omics latent representations are far apart in the latent space. However, they average posterior epigenomic and transcriptomic latent predictions while we decode them separately. By construction we force latent epigenomic and transcriptomic representations to match, implying that respective dynamics match. This is an assumption that might not always hold. An example might be in transcriptome and proteome multi-omics pairs. Furthermore, MultiVI rely on their network structure to account for batch effects and normalization, while we only normalize gene expression data using scran (Lun et al., 2016) to estimate cell specific offset.

Method	Epigen. Noise Model	Trancr. Noise Model	Manifold strategy	Batch. Cor
Cobolt	ZINB	ZINB	PoE	-
scMVAE	ZINB	ZINB	MoE	-
scMM	ZINB	ZINB	MoE	-
MultiVI	Bernoulli	ZINB,NB,Pois	Jeff. Diver.	+
scMVI	ZINB/Binomial	ZINB	Jeff. Diver.	-

Table 6.2 **Related methods using VAEs to model partially or fully observed single cell multi-omics data**. Table summarizes key differences for each of the reviewed methods.

6.3 Experimental setup

scMVI is a method designed to perform downstream analysis in droplet and plate based data. By leveraging information from multiple sources in a noise aware framework, it can provide robust noise treatment and important biological insights even under partially observed data. To establish that, we run a series of demonstrations on datasets with different properties.

Effectively handling missing data in downstream analysis requires accurate imputation along with an effective manifold building strategy. To asses imputation performance, one of the available layers was used to impute the rest in test set cells which were not used to train the model. Imputed values were plotted against ground truth for reference in test set cells. Furthermore, imputation accuracy was judged by tracking the number of observed genes lying within a 95% confidence interval around imputed values for each cell. In case of mESC_Pj, where only one modality existed, performance was similarly assessed for reconstruction instead of imputation. The lack of tools for performing single cell integration across platforms, makes integration assessment challenging. Our first step is to inspect the visual outcome of artificially unpaired test data. To quantify how effective scMVI would be in integrating unseen data, we compare the Euclidean distances between training and test set latent representation and examine proximal cell types in the trained latent manifold.

Assessing manifold building strategy is difficult on multiple levels. The lack of universally accepted metrics along with few labeled single cell multi-omics datasets means that results will vary. Here we consider a set of qualitative and quantitative approaches. We begin by looking at the biological insights the latent space offers. In turn, latent manifolds' consistency across data modalities is evaluated by plotting proximal cell types heatmaps. These experiments are repeated for multiple related methods (MultiVI, scMM, scMVAE and Cobolt) allowing to compare and contrast. To the best of our knowledge, there are not any methods addressing epigenomic and transcriptomic layers with more than two modalities. Hence, for mESC_Ag data we focus on interpreting latent structure.

6.4 Results

6.4.1 scMVI achieves effective integration accurate and probabilistic imputation.

We start our analysis by looking at imputation and integration performance on data generated from mESC and PBMC. For mESC, we have two datasets (mESC_Pj and mESC_Ag), obtained at roughly the same developmental stages but with substantially different protocols. mESC_Pj contains gene expression counts, sequenced with commercial droplet based protocol. In mESC_Ag, DNAm, gene expression and chromatin accessibility observations were gathered using scNMT plate-based platform. The PBMC dataset is also harvested with a droplet-based protocol and consists of gene expression and chromatin accessibility readings. Qualitative differences between sequencing methods are reflected from the number of cells in each dataset. On the one hand, mESC_AG generated with scNMT has $\sim$ 800 cells with good coverage. On the other hand, commercial 10X platforms stand out for their scale, yielding $\sim$ 52k and $\sim$ 10k for mESC_Pj and PBMC respectively, at the cost of high drop-out. This coverage is neatly demonstrated in Figures 6.4a- 6.4c and Table 6.1.

Figure 6.4 presents three examples of imputed cells, one for each dataset. Despite the challenging nature of imputation in single cell omics, cross modal imputation works well irrespective of sequencing assay (Figures B.5 - B.12 in B.1). In all cross modal imputations, genomic region predictions fall almost symmetrically around the theoretical perfect match (line $y = x$). Furthermore, error bars around predictions extrapolated from randomly chosen cells from each dataset demonstrate that $\sim 96.9\%$[12] of imputed data, across all data, lie at most 2 standard deviations away from the observed ones. Breakdown of the statistic by sequencing technology, indicates that in droplet based data 97% of imputed vales are within 2 standard deviations while in plate based this drops to 93%. Plots summarizing average error as a function of confidence interval can be found in Figure B.1.

In the 3.1% of imputed values outside the confidence interval, we find cells that present high error and low uncertainty. The source of error can be traced back to the training set. For example, in mESC_Ag where only few observations existed, scMVI could not make accurate predictions in hypo-methylated regions when the majority of training observations were hemi-methylated (i.e. partially methylated). Despite the abundance of observations present in mESC_Pj and PBMC, there are regions which for most training set examples are

[12]This statistic was acquired with stratified sampling. Initially 100 cells are sampled from each dataset to get an estimate of predictions outside confidence interval. Then using number of observations in each dataset, we extrapolate to the whole population.

zero, pushing the zero inflation parameter of the likelihood to high values. This makes it challenging for the model to match large non-zero predictions in the test set.

An alternative way of assessing imputation performance is by using different modalities to impute missing values and then compare moments with the ground truth data. In Fig. 6.5 and 6.6 means and standard deviations of gene expression, chromatin accessibility and DNA methylation imputed profiles are compared to test set mESC_Ag ground truth data. Fig. 6.5 suggested that for chromatin accessibility and DNA methylation data the distribution of the first moment for imputed values give a relatively accurate representation. On the other hand, imputed gene expression cell means tend to be more dispersed compared to ground truth as indicated by Fig. 6.5a. By looking at Fig. 6.6 it also looks that in some cases the imputed standard deviation distributions do not exactly the ground truth. Most obvious example is for DNA methylation in Fig. 6.6c, where imputed profiles standard deviations are under-estimated. To a lesser extent this is also present in Fig. 6.5b. For gene expression, namely Fig. 6.6a, imputed profiles standard deviations are closer to ground truth. In mEBC data (Fig. B.2a and B.3a) we also see that both mean and standard deviation for both molecular layers (ie. epigenome and transcriptome) tend to be more dispersed compared to ground truth. Considering the Pearson correlation between imputed profiles and comparing it to raw data for mESC_Ag and mEBC (Fig. 6.7 and B.4 can justify the difference in means and standard deviations dispersion observed before. In almost every case the distribution of Pearson correlation between imputed profiles matches the one obtained from ground truth data. This suggests that the model has adjusted its moments to match data patterns.

Due to a lack of single cell integration tools capable of handling data sourced from different protocols, obtaining a reliable ground truth for partially observed cells is challenging. To asses scMVI's integrative performance we proceed by artificially unpairing jointly quantified molecular layers from PBMC and mESC_Ag data. To demonstrate scMVI's integration performance, we assume that test set cells were sourced from single omics protocols comprising the multi-omics pipeline. These single modality datasets are projected to scMVI's latent manifold where a series of experiments demonstrate how well scMVI would match partially quantified data to produce a multi-omics dataset.

Figures 6.8a and 6.8b give a qualitative summary of the PBMC and mESC_Ag unpaired modalities projected by scMVI. By examining the manifolds we see that irrespective of the dataset, scMVI manages to create consistent representations from partial observations. This macroscopic view is an encouraging first step, however for downstream analysis it is equally important to match similar lineages. In absence of ground truth, we run two series of experiments. The first one compares latent space euclidean distances between the paired training and test set cells with artificially unpaired modalities. If the distribution of distances

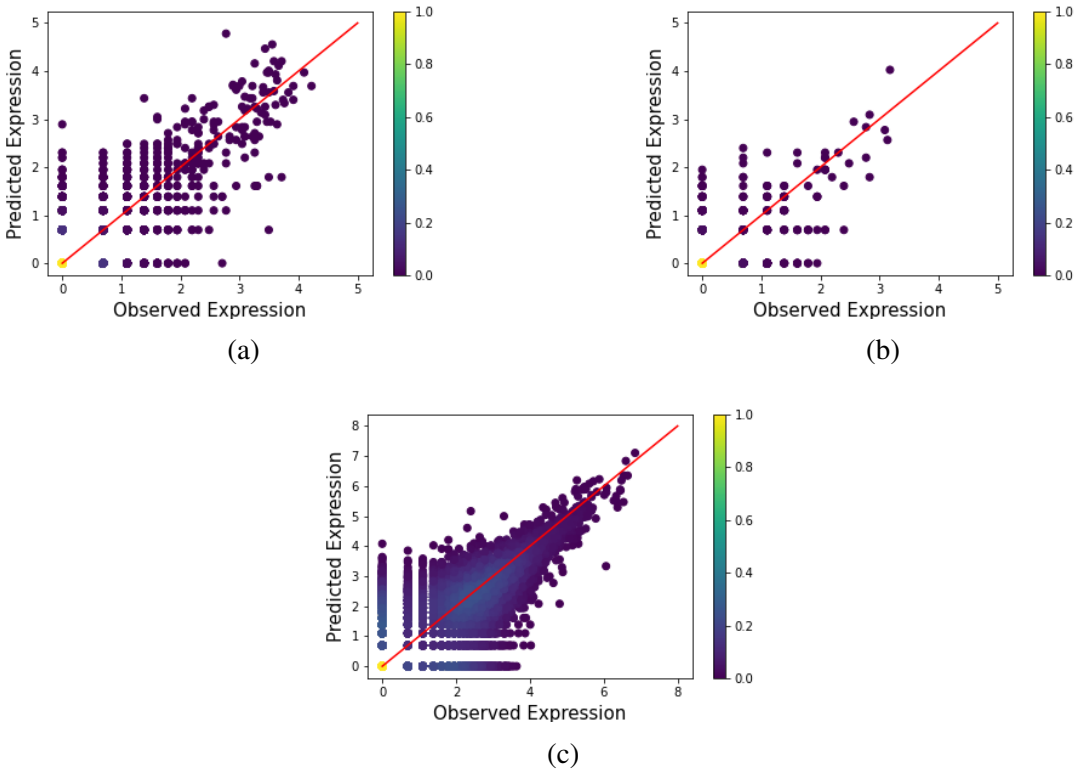
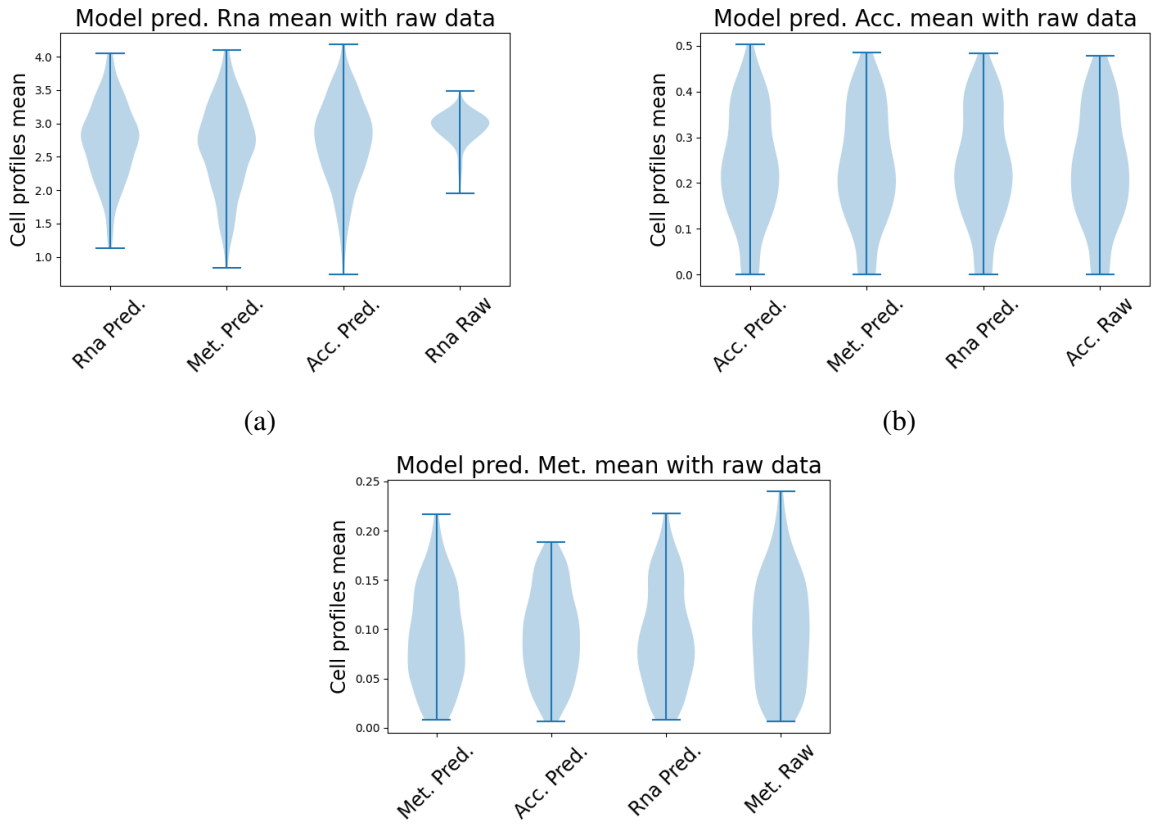


Fig. 6.4 **Summary of imputed expression for data generated with commercial 10X and scNMT (Clark et al., 2018). Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots of imputed gene expression for one cell of the test set used when analysing the mESC_Pj (6.4a), PBMC (6.4b) and mESC_Ag (6.4c) datasets. For the multi-omics datasets (PBMC and mESC_Ag), gene expression is imputed using chromatin accessibility. For mESC_Pj we present the reconstruction of gene expression data since there are no other modalities to impute from. Each dot represents a gene and is color-coded using frequency of point occurrence. The red line is $y = x$ corresponds to the perfect match. Error bars are omitted for clarity of plots.

is similar between paired and unpaired values, then similar cells sequenced with a single omics protocol could be accurately matched. The second set of experiments looks at lineage of neighbouring cell in paired data. If lineages in close proximity match, then integrating partial observations by choosing the closest cell likely gives desired results.

Violin plots in Figures 6.9a- 6.9d present the distribution of euclidean distances between alternative representations for selected matched and unmatched cells. Apart from endoderm and primitive streak cells where on average the eucldiean distance is higher for unmatched cells, differences are minuscule. Given similar cells and sequencing protocol, we have good indication that alternative latent realizations will be as close to each other as in the training data.



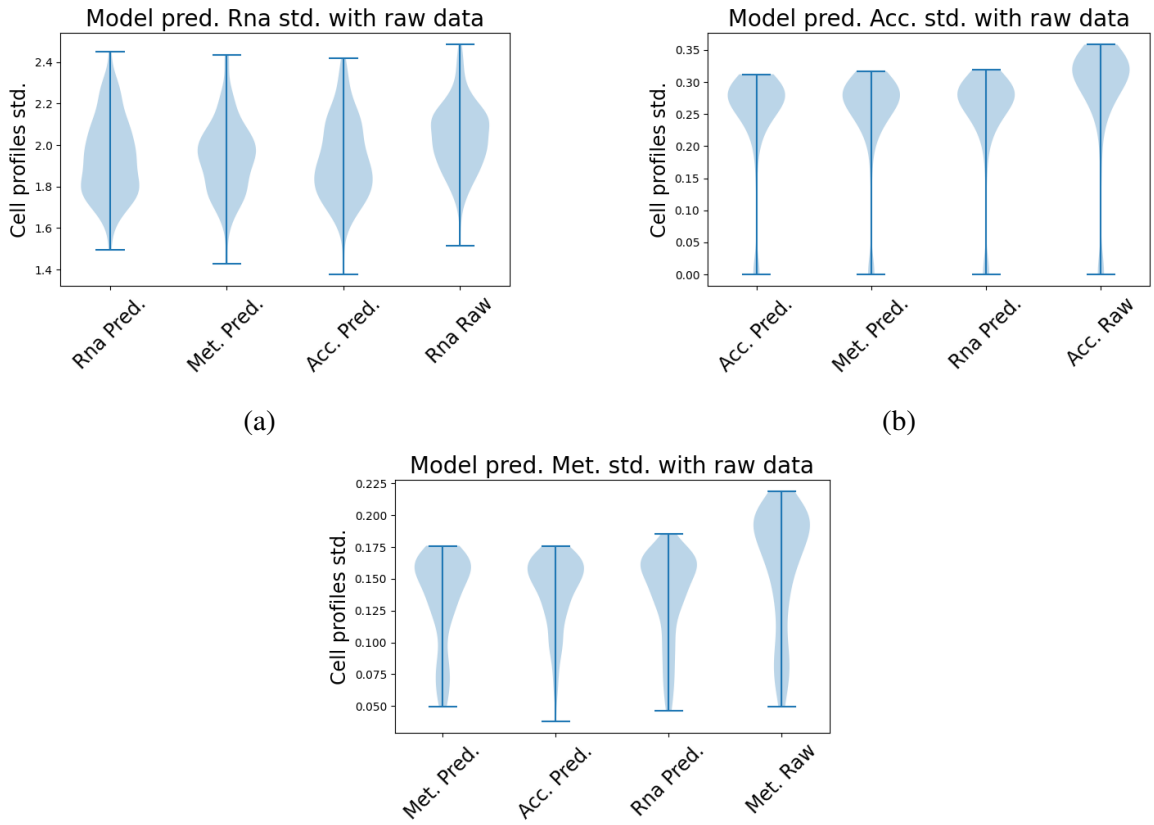
(a) (b)

(c)

Fig. 6.5 **Violin plots summarizing distribution of imputed and raw profile means for gene expression, chromatin accessibility and DNA methylation in mESC_Ag. Useful comparison between statistics of imputed and ground truth values.** (6.5a) Violin plots summarizing distribution of imputed and raw profile means for gene expression in mESC_Ag test cell data. (6.5b) Violin plots summarizing distribution of imputed and raw profile means for chromatin accessibility in mESC_Ag test cell data. (6.5c) Violin plots summarizing distribution of imputed and raw profile means for DNA methylation in mESC_Ag test cell data.

Figure B.18a indicate that for paired PBMC data, the majority of proximal cells are of the same cell type. In paired mESC_Ag (Figure B.17), there are some lineages where integration is going to be challenging. As most primitive streak cells are neighboured by exe ectoderm, a primitive streak latent representation in that region will probably be mismatched assuming a strategy of picking the nearest cell. On the other hand, the majority of lineages by and large are bordered by same lineage cells.

In summary, these results demonstrate that scMVI can effectively impute missing values while accurately tracking uncertainty. Using simple heuristics like confidence interval it is possible to prevent unreliable predictions from impacting subsequent analysis. In integration



(a) (b)

(c)

Fig. 6.6 **Violin plots summarizing distribution of imputed and raw profile standard deviations for gene expression, chromatin accessibility and DNA methylation in mESC_Ag. Useful comparison between statistics of imputed and ground truth values.** (6.6a) Violin plots summarizing distribution of imputed and raw profile standard deviations for gene expression in mESC_Ag test cell data. (6.6b) Violin plots summarizing distribution of imputed and raw profile standard deviations for chromatin accessibility in mESC_Ag test cell data. (6.6c) Violin plots summarizing distribution of imputed and raw profile standard deviations for DNA methylation in mESC_Ag test cell data.

we have strong indication that scMVI will accurately match most paired and unpaired data. These properties lay the foundation of a state of the art method, capable of performing reliable downstream analysis across various protocols.

6.4.2 scMVI leverages effectively multi-source information to build biologically meaningful latent spaces.

Using available datasets we seek to understand what biological patterns scMVI captures in its latent space. Our analysis will primarily focus on the multi-omics datasets (PBMC and

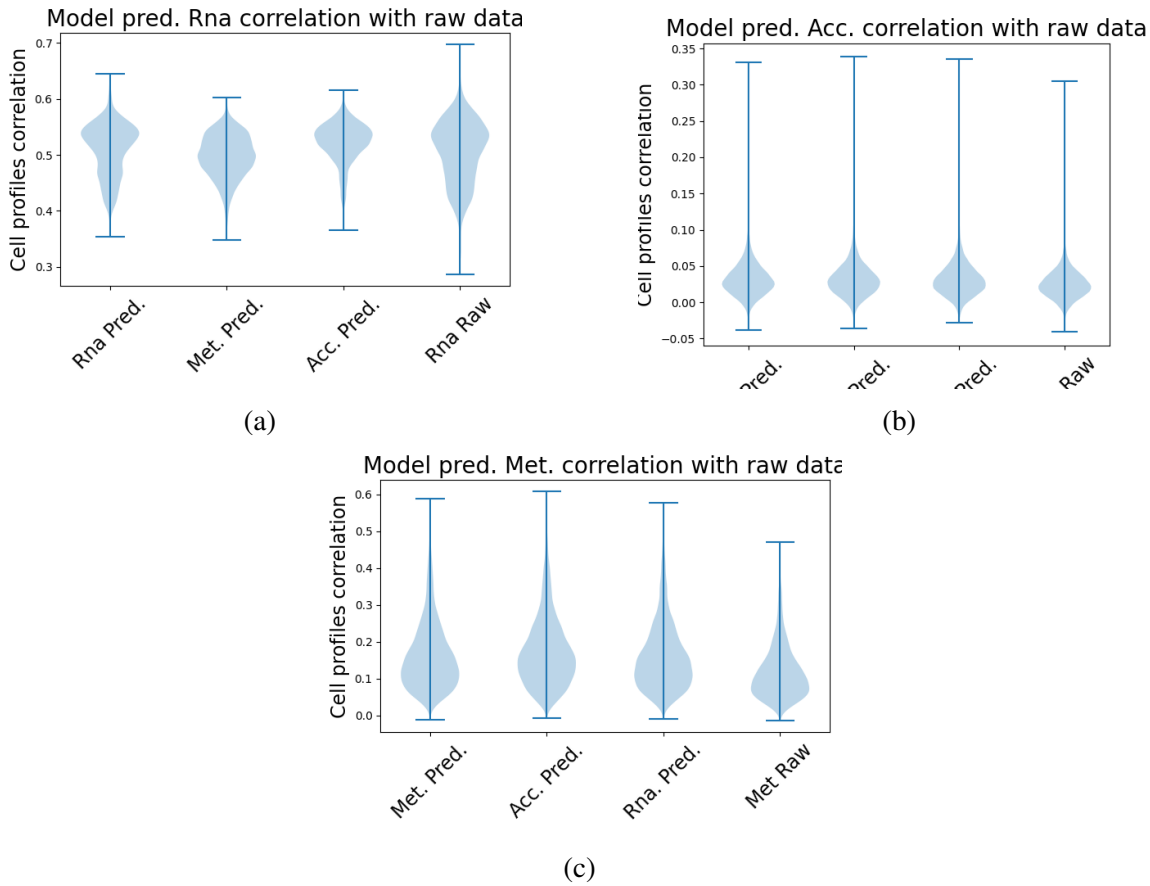


Fig. 6.7 **Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression, chromatin accessibility and DNA methylation in mESC_Ag. Useful comparison between statistics of imputed and ground truth values.** (6.7a) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression in mESC_Ag test cell data. (6.7b) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for chromatin accessibility in mESC_Ag test cell data. (6.7c) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for DNA methylation in mESC_Ag test cell data.

mESC_Ag), to demonstrate that the aligning strategy based on Jeffrey's divergence can yield important insights. We also look at mESC_Pj data to get a diverse understanding around scMVI's exploratory capabilities. In PBMC, we qualitatively and quantitatively compare scMVI's latent manifold with related methods (MultiVI, scMM, scMVAE and Cobolt). For mESC_Ag, we consider lineage and developmental stage metadata to perform exploratory analysis. A similar analysis is also applied to mESC_Pj. To support our effort, data from respective modalities were projected to the latent space and averaged across modalities creating a 10 dimensional latent cell representation. To ease the manifold examination a

(a) (b)

Fig. 6.8 **UMAP visualization of artificially unpaired data projected to the latent space for PBMC and mESC_Ag. Visualization qualitatively demonstrating that similar manifolds can be generated from partial representations.** UMAP plots summarizing latent space projections of artificially unpaired readings from jointly profiled PBMC (6.8a) and mESC_Ag (6.8b). Each dot represents a cell and is color-coded using its source modality.

two dimensional depiction was created with UMAP (McInnes et al., 2018), where latent representation were color-coded using cell metadata.

The manifold depicted in Figure 6.10a presents a unique cell structure. Cells with different function (i.e. T-cells from B-cells) along with different subtypes of the same cells can easily be distinguished, only by looking at their color-coding. This clustering consistency is also confirmed by the heatmap summarizing subtype of cell in closest proximity to each cell (Figure B.18a).

For example, the majority of naive helper T-cells (labeled CD4_Naive) are neighbored by same helper T-cells and central memory T-cells (labeled CD4_TCM). However, when naive helper T-cells are contrasted with any B-cell subtype, no two cells are to be found neighboring each other. Comparing these findings to results of alternative methods (MultiVI, scMM, scMVAE and Cobolt), we get consistent results (Figures B.18 - B.20 in Section B.1) despite model differences. All models distinguish between different cell types while keeping close similar subtypes. This is not surprising as an analysis [13] by Seurat's authors Stuart et al. (2019), indicated that a linear projection is equally effective. To some extent this is expected as PBMC consists of healthy cells that do not change much over time.

Figures 6.10b and 6.10c present a thorough summary of scMVI's latent manifold on mESC_Ag sequenced with scNMT. Lineage metadata allow for a comprehensive analysis of the developmental dynamics (Figure 6.10c). Combining developmental stage with

[13] https://satijalab.org/seurat/articles/atacseq_integration_vignette.html

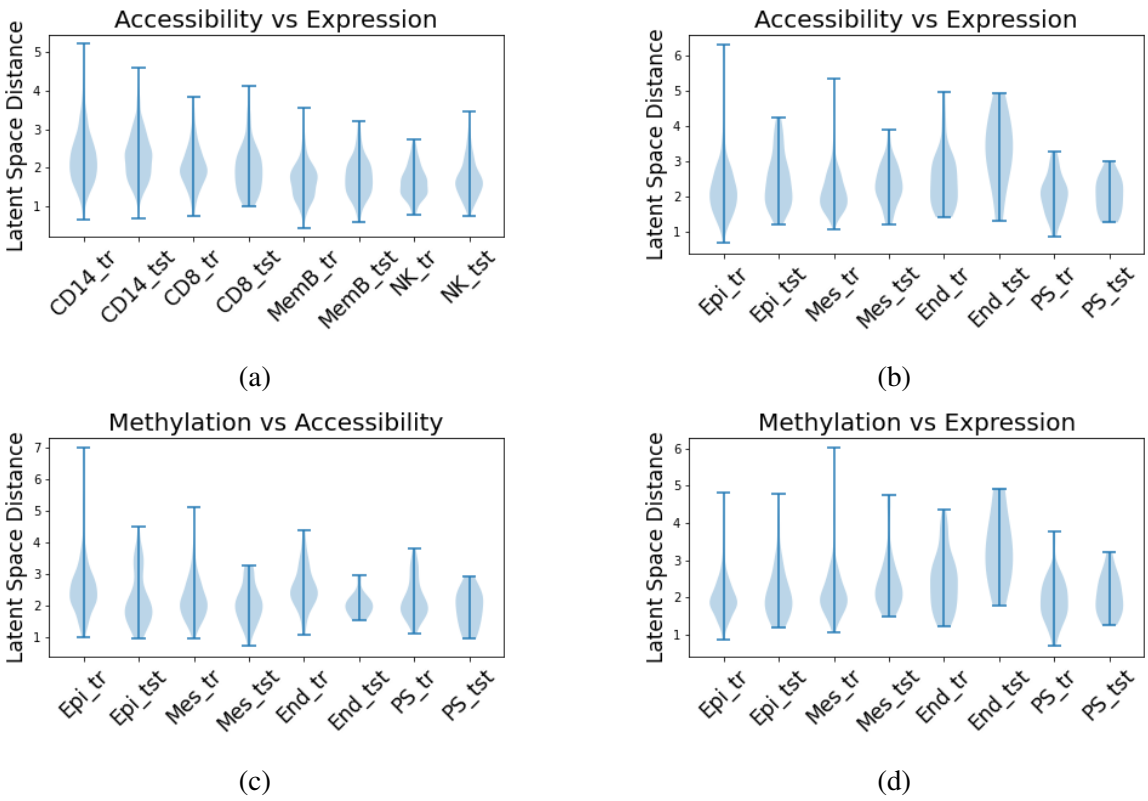


Fig. 6.9 **Violin plots summarizing latent representation euclidean distances in training set and artificially unpaired modalities in test sets cells, broken by technology and selected cell types. Figures used to asses how close in the latent manifold would previously unseen data would sit.** Violin plots summarizing latent representation euclidean distance in training set cells and artificially unpaired modalities in test set cells for droplet (6.9a) and plate based data (6.9b,6.9c,6.9d). In mESC_Ag the selected cells are Epiblast (Epi), Mesoderm (Mes), Endoderm (End) and Primitive Streak (PS). Similarly for PBMC the selected cells were CD14 Mono (CD14), CD8 T cells (CD8), Memory B cells (MemB) and Natural Killer cells (NK). Violin plots of euclidean distances on the training set are distinguished from the ones of the test set using labels "_tr" and "_tst" respectively. For example violin plot of euclidean distances between epiblast cells on the training set is denoted by Epi_tr.

cell type labels enables the development of fine grained distinctions within development. In embryonic day $E4.5$, clear boundaries can be drawn between primitive endoderm and epiblast cells. In embryonic days $E5.5 - E6.5$ visceral endoderm can easily by discriminated from epiblast and primitive streak cells, but differences between the latter two types are more subtle. In embryonic day $E7.5$, the endoderm and ectoderm cells branch off to two directions. Mesoderm cells can equally be found close to endoderm and ectoderm cells. The

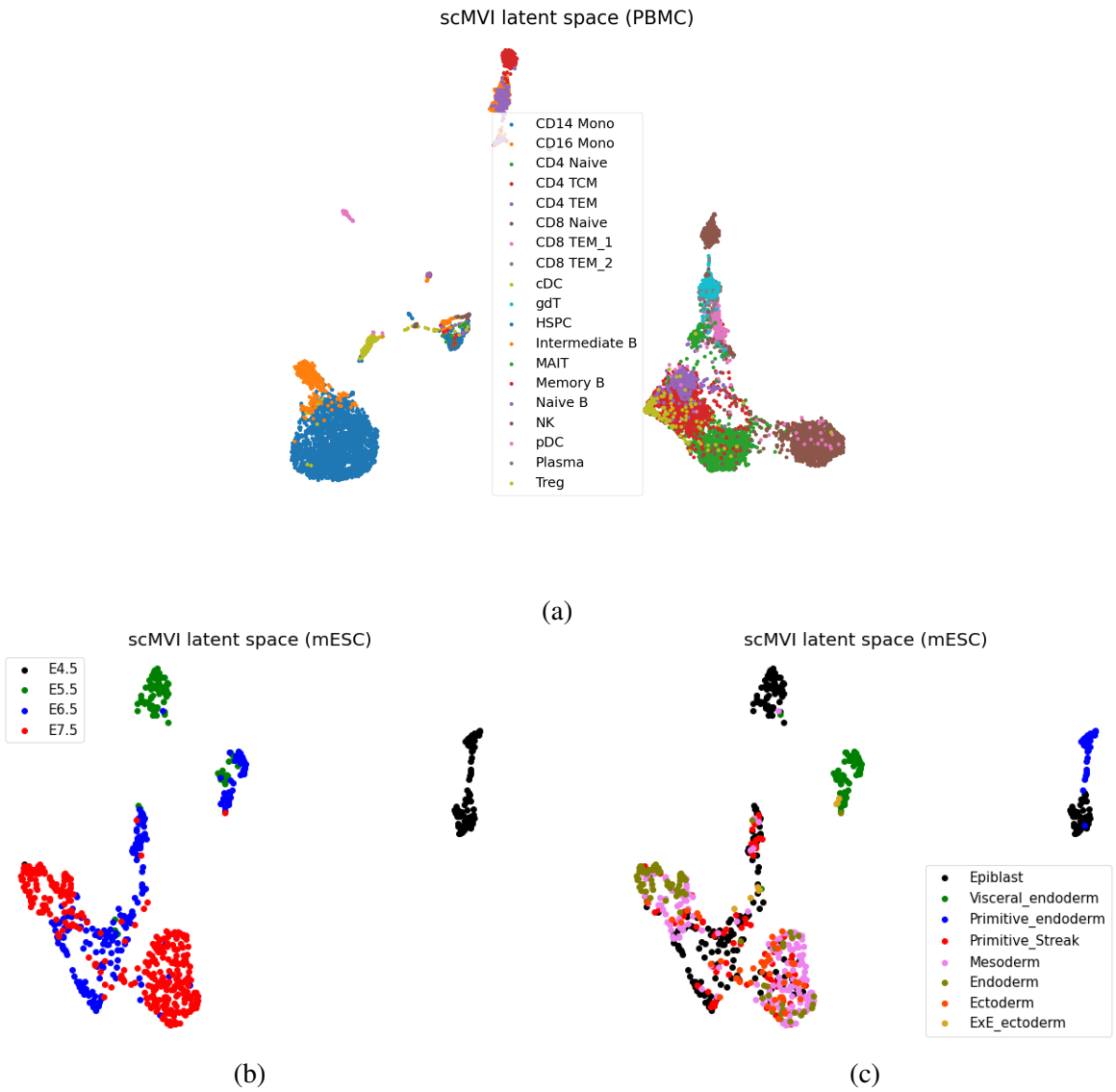


Fig. 6.10 **Summary of scMVI's latent space on mESC_Ag and PBMC. Figures demonstrating that scMVI learns latent representations correlated with cell phenotype or developmental state.** 2D UMAP projection of scMVI's, trained on mESC_Ag and PBMC with each dot representing a cell. Since scMVI estimates one latent representation per molecular layer these plot are generated by averaging across them. (6.10a) scMVI latent space on PBMC cells color-coded by cell type. (6.10b) scMVI latent space on mESC_Ag data color-coded by developmental stage. (6.10c) scMVI latent space on mESC_Ag color-coded by lineage.

differentiation of primitive streak Cells to mesoderm and ectoderm is also reflected in the manifold with the three cell types being closely mapped.

Using Figure 6.10b developmental patterns are easily spotted. Young cells (embryonic day $E4.5$) are concentrated on the right (labeled in black), intermediate cells (embryonic day $E5.5$) in the middle (labeled in green) and more mature (embryonic days $E6.5$-$E7.5$) in the bottom left (labeled in red and blue). Cells transitioning from embryonic day $E6.5$ to $E7.5$ form two branches, signaling strong differentiation, consistent with findings in Argelaguet et al. (2019); Pijuan-Sala et al. (2019).

Epiblast cells are spotted in all developmental steps. Using the schematic of the developing mouse embryo in Figure 1 of Argelaguet et al. (2019), we expect to encounter most Epiblast cells around embryonic days $E4.5$ and $E5.5$. Hence, this looks like an artifact. However, reviewing mouse development bibliography explains this phenomenon. According to Qiu and Shendure (2021), epiblast cells differentiate to ectoderm and srimitive streak, which is consistent with our findings. As these lineage labels are derived using marker genes and the theorized mouse development models present an approximation of the underlying process, a small mixing between mature epiblast and young ectoderm/primitive streak (found in $E6.5$ and $E7.5$) is expected.

Figures 6.11a and 6.11b breakdown the various properties of scMVI's latent manifold when training on mESC_Pj. By examining linage position in Figure 6.11a, it is evident that scMVI has learned to discriminate between cells. In the bottom and top left we identify exe ectoderm, exe endoderm and visceral endoderm cells in purple, brown and pink clusters. Then we have a small cluster in the top right consisting of blood progenitors and erythroid cells. On bottom cluster and the lower part of the main cluster, epiblast and primitive streak cells can be found.

Cells in the center of the large central cluster are a mixture of mesoderm, ectoderm and early developmental tissue. Looking at the upper area of the main cluster, we identify a mixture of mesoderm and ectoderm cells. Focusing on mesoderm we find clusters of praxial, somitic, nascent, intermediate and mixed mesoderm along with mesenchyme in close proximity. scMVI has a hard time distinguishing between mesoderm subtypes, but it can much better tell the difference between mesenchyme, presumably because they are sufficiently differentiated. In ectoderm, we identify two major subtypes: surface ectoderm and neural plate, consisting of neural crest and neural tube (Defined by Brain and Spinal Chord labels). These subtypes are spotted in the left, upper left part and bottom right of the main cluster.

In some cases, lineages in close proximity hint known developmental patterns. For example combining exe endoderm with visceral and definitive endoderm cells hints gut endoderm formation. Similarly, when praxial/somitic mesoderm cells are considered with

blood and haematoendothelial progenitors, endothelium and mesenchyme, it hints early hematopoietic and myoskeletal system development.

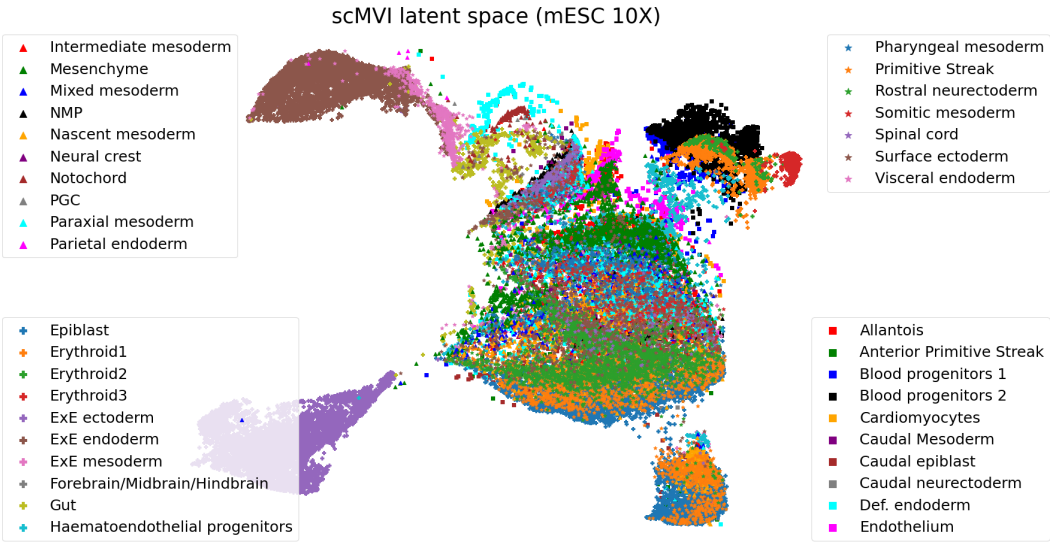
Identifying clear developmental dynamics from Figure 6.11b is challenging. The gastrulation landscape between embryonic days $E6.5$ and 8.5 is convoluted as many developmental phenomena are simultaneously taking place. Still, it is possible to identify localized patterns. For example in exe ectoderm and endodrem cells' clusters, qualitative differences between younger and more mature cells force scMVI to sort them accordingly. Furthermore, in the main cluster developmental ordering can be spotted. Starting from epiblast and visceral endoderm we can trace various differentiated cells until we reach the most specialized which are color-coded in gold (neural crest and neural tube cells).

Comparing results with scVI (Lopez et al., 2018) (Figure B.13) we identify both differences and similarities. Looking at exe endoderm and exe ectoderm in Figure B.13a we can trace similar structure to scMVI. However, blood progenitors, gut and definitive endoderm cells are only some examples of cells that are not easily distinguishable from the rest. Developmental patterns in Figure B.13b are in some cases easier to spot. For example, paths of cells differentiating from epiblast or visceral endoderm cells can be easily pinpointed in the main cluster of Figure B.13b.

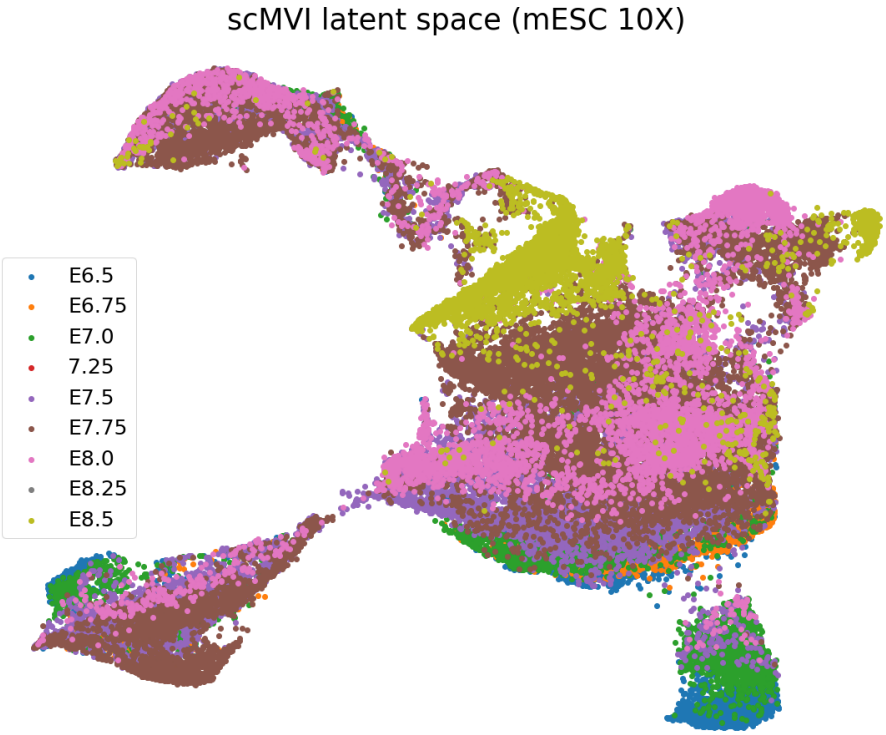
In all three datasets we looked, scMVI's manifold building strategy yields latent spaces where cell arrangements strongly correlate with phenotype. For PBMC it can accurately distinguish between different cells and distinct subtypes of the same cell family. In the rapidly differentiating mESC cells, scMVI extracts biological patterns whose complexity directly links their developmental phase. In mESC_Ag where mESC are sourced between embryonic days $E4.5 - E7.5$, scMVI tends to accurately distinguish between lineages. In mESC_Pj, where mESC are sourced between embryonic days $E6.5 - E8.5$, scMVI captures many convoluted processes such as Neural Plate development. In both mESC datasets, scMVI has learned to distinguish between various cell types. However, it is impossible to completely decipher the complex structure without a joint consideration of lineage and developmental stage metadata. Taking everything into account scMVI's latent manifold learns to discriminate between different cell types while accounting for developmental dynamics whenever necessary.

6.4.3 Exploratory analysis using Differential directions

As established before, scMVI's latent manifold arranges latent cell representations using phenotype properties. More precisely, scMVI learns to differentiate between lineages while keeping cells with similar phenotypes in close proximity. In this section we explore an



(a)



(b)

Fig. 6.11 **Summary of scMVI's latent space on mESC_Pj color-coded by lineage and developmental stage. Figures demonstrating that scMVI learns clear latent representations correlated with cell phenotype or developmental state.** 2D UMAP projection of scMVI's latent space, trained on mESC_Pj. Each dot represents a cell and is color-coded by lineage (6.11a) and developmental stage (6.11b).

approach that relates these meaningful arrangements to measured molecular marks, aiming to improve downstream analysis.

We introduce Differential Direction (DD), a novel estimator for tracking direction of differentiation between average group representations in the latent space. To relate DD with measured gene expression, we apply scMVI's expression decoder to the difference vector creating a heuristic similar to Log Fold Change (LFC) of a gene expression values in two different conditions used in transcriptomic differential testing. We call it Differential Direction Log Fold Change (DDLFC)[14] and use it to demonstrate it can lay the foundation for differential testing alternatives. In case of a small number of cells or partially observed data, gene expression differential testing might be more stable when performed using DDLFC compared to LFC. This is because DDLFC relies on DD which is estimated using thousands of genes and multiple molecular layers.

To demonstrate DDLFC's superiority against standard LFC provided by related deep learning models (LFC_NN) (scVI and MultiVI) and traditional gene expression differential testing (LFC_DT), we asses it in terms of robustness and the biological insights it can offer. For robustness, we simulate 7 scenarios addressing cases of few observations and label imbalance from mESC_Pj and PBMC datasets. Detailed subsampling summary can be found in Table 6.3. To compare the estimators, we initially quantify the DDLFC and LFC_NN distributions by sampling from the trained models conditioned on observations of each of the 7 scenarios. To estimate DDLFC distribution, we decode samples of the difference between latent group representations averaged across molecular layers. For LFC_NN we use differential expression method provided by scVI and MultiVI for mESC_Pj and PBMC respectively. Then using the median DDLFC and LFC_NN in each scenario we compute the range from the smallest to the largest value for each method across the 7 simulated scenarios. This range can be understood as a metric of how resistant are the estimators to label imbalance and sub-sampling.

To understand how DDLFC can aid in downstream analysis, we run Gene Set Enrichment Analysis (GSEA) using DAVID (Huang, 2009) on regions with the strongest change (increase or decrease). In keeping the analysis brief we focus on specific pairs of consecutive days in the mESC datasets. For mESC_Pj we compare GSEA on regions with the strongest DDLFC change against LFC_NN and LFC_DT differentiation. DDLFC results on mESC_Ag are

[14]Briefly DD is the difference vector between two groups of cells with distinct phenotypic or developmental stage label averaged across available modalities. DDLFC is estimated by applying the relevant decoder to DD. Most of the analysis is performed using DDLFC estimated with average normalized gene expression decoder. The only exception is later in the results where we demonstrate some of the biological processes discovered by enrichment analysis using average DNAm and chromatin accessibility DDLF. For formal mathematical introduction of DD and DDLFC please look in Section B.3.

subsequently used for exploring the biology of early development and are compared to LFC_DT.

Exp.	1	2	3	4	5	6	7
set1	No sub.	No sub.	80%	80% sub.	50% sub.	50% sub.	80%
set2	No sub.	80%	No sub.	50% sub.	80% sub.	50% sub.	80%

Table 6.3 Level of cell subsampling in each of the 7 experiments created using mESC_Pj and PBMC to compare DDLFC, LFC_NN and LFC_DT behaviour under label imbalance in gene expression data. Set 1 and 2 are simulated datasets comprised of cells with the same developmental or phenotype label but distinct between datasets. After training scMVI, scVI and MultiVI are conditioned on simulated sets 1 and 2 to estimate median DDLFC (scMVI) and LFC_NN (scVI, MultiVI). Then the average along with the range of medians for DDLFC and LFC_NN across the simulated scenarios are compared.

Figures of Panel 6.12 compare average and range of DDLFC and LFC_NN median estimators across 7 sub-sampling scenarios in mESC_Pj for pairs of embryonic days $E7.0$ to $E7.5$ and $E8.0$ to $E8.5$. Similarly figures comparing DDLFC to LFC_DT can be found in Figure B.22 for the same pairs of embryonic days. Figures 6.12a, 6.12b, B.22a and B.22b indicate strong dependency between DDLFC, LFC_NN and LFC_DT, which is also confirmed by Pearson correlation estimates (0.52, 0.74, 0.38, $0,76$ respectively). Hence GSEA on genomic regions with strong gene expression DDLFC or LFC_NN/LFC_DT will likely yield similar outcomes. DDLFC's superiority comes into play when robustness is considered. In examined Figures 6.12c and 6.12d (along with B.22c and B.22d of Section B.3, it is apparent that subsampling and label imbalance stronger impact both LFC_NN and LFC_DT compared to DDLFC. For the pair where DDLFC is impacted the most on mESC data, in 80% of the genomic regions DDLFC is impacted less than LFC_NN as indicated by Figure 6.12d. For the same embryonic days LFC_DT in 100% of genomic regions DDLFC is impacted less (Figure B.22d). A similar comparison for PBMC data confirmed DDLFC resistance to subsampling, however DDLFC and LFC_NN/LFC_DT values were not correlated (Figures B.21 and B.23 in Section B.3). The lack underlying kinetics in PBMC data are probably to blame for this result, but to get conclusive results more research must be done.

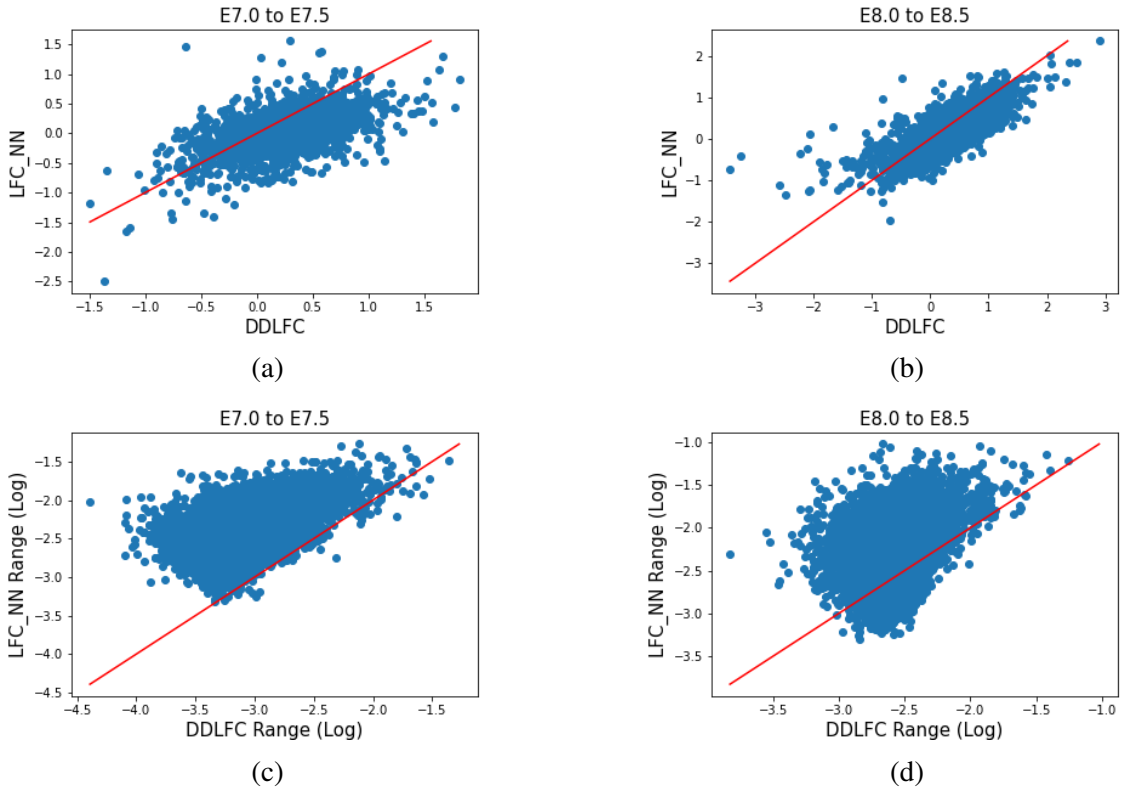


Fig. 6.12 **Median DDLFC and LFC comparison for mESC_Pj. Scatter plots used to asses how robust is DDLFC compared to LFC under various scenarios for specific developmental steps.** Scatter plots summarizing average and range of median DDLFC against average and range of median LFC across simulated scenarios in label pairs where it is less (Embryonic days $E7.0$ vs $E7.5$)(6.12a,6.12c) and most (Embryonic days $E8.0$ vs $E8.5$)(6.12b,6.12d) impacted from subsampling and label imbalance, for mESC_Pj. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range.

We proceed with experiments seeking to understand what biological signal DDLFC, LFC_NN and LFC_DT can extract. For the comparative analysis, GSEAs are performed on 500 genomic regions with the strongest positive/negative median DDLFC values and largest differentiation as estimated by LFC_DT and median LFC_NN, on pairs of embryonic days $E7.0-E7.5/E8.0$-$E8.5$. To identify the processes we allowed a minimum of 10 genes, p-values up to 0.3 and sort them based on enrichment score (see Figures B.24-B.35). Comparing GSEA on the strongest positive and negative values between embryonic days $E8.0-E8.5$ (Figures B.24, B.25,B.26, B.27, B.32 and B.33), we identify many similar processes tied to development and homeostasis. For development, neural tube closure and brain development with enrichment scores above 1.6 for LFC_NN, LFC_DT and DDLFC respectively are detected. Looking for differences, we see that LFC_NN points to

angiogenesis while DDLFC to heart development with enrichment scores 1.5 and 1.4. When it comes to house keeping processes both LFC_NN, LFC_DT and DDLFC present similarities. Cell adhesion is pinpointed by DDLFC, LFC_NN, and LFC_DT with enrichment scores 1.7, 1.5 and 1.5. Interestingly, all approaches identify RNA splicing with enrichment score 1.3 or more, hinting at ties between alternative approaches that measure gene expression kinetics (i.e. RNA velocity). Given the smaller Pearson correlation 0.52 between LFC_NN/LFC_DT and DDFLC in the pair $E7.0 - E7.5$, strongly enriched processes differ. For example both DDLFC and LFC_NN spot negative regulation of gene expression with enrichment score 1.6 and 1.3, but only DDLFC identifies brain development.

Because of the improved sparsity of mESC_Ag data GSEA is equally effective on 250 genomic regions. Similarly to comparative analysis GSEA was performed using strongest positive/negative DDLFC values in consecutive developmental stages. As no other method has been trained on mESC_Ag we compare results with LFC_DT. More can be found in Figures B.36 - B.41 and B.44- B.49. In Figure B.37 we see a plethora of biological processes both linked to house-keeping processes and development having high enrichment score. In the house-keeping side we have multi-cellular organism development, cell differentiation and regulation of transcription with enrichment scores of 2.8, 1.8 and 1.3 respectively. Notable developmental processes in Figure B.37 are heart, brain and in utero embryonic development with strong positive DDLFC genes yielding enrichment scores of 2.7, 2.4 and 1.7 respectively. Heart development is consistent with suggestions of Ivanovitch et al. (2017); Tyser et al. (2016) which explain how cardiac precursors in the splanchnic mesoderm differentiate into cardiomyocytes in embryonic day $E7.5$. Standard differential testing (Figure B.45) also performs well and manages to identify many house keeping and developmental processes indicated by DDLFC. For house-keeping processes multi-cellular organism development, cell differentiation and regulation of transcription the enrichment scores are 3.8, 2.6 and 1.2. Similarly for heart, brain and in utero embryonic development the sores are 3.5, 2.8 and 1.9. Interestingly LFC_DT also identifies kidney development and nervous system development with enrichment scores of 4.1 and 2.7. For the same pair, patterns are also extracted from GSEA using positive DDLFC on DNAm (Figure B.42). The identified developmental processes are brain development, Neural Tube Closure nervous system development with enrichment scores 3.4, 3.8 and 2.7 respectively. Furthermore, positive DNA methylation and chromatin accessibility DDLFC (Figure B.43) recover various house keeping processes like regulation of cell cycle (enrichment scores 2.1 and 1.6), demonstrating the vital role of epigenome in homeostasis.

Turning our attention to genes with the strongest negative DDLFC and LFC_DT values between $E6.5$ and $E7.5$ (Figures B.36 and B.44) we identify many house keeping processes.

DDLFC identifies processes like translation, protein stabilization and negative regulation of gene expression with enrichment scores 5.6, 3.7 and 2.4. LFC_DT points to cell proliferation, regulation of gene expression and regulation of cell cycle 2.1, 1.7 and 1.6. It is also interesting to note that mRNA splicing and RNA splicing biological processes come up between $E4.5-E5.5$ and $E6.5-E7.5$ pairs reinforcing the claim about connections between DDLFC and RNA velocity. Looking at earlier developmental stages (Figures $E4.5$-$E6.5$ B.39- B.41 and B.46- B.49) we notable processes are related to cell cycle, regulation of transcription and translation.

This analysis has demonstrated that DDLFC can become a viable alternative to traditional LFC used in differential analysis. DDLFC is more robust to label imbalance as it utilizes several molecular layers and numerous genes. DDLFC correlates with LFC, hence it can capture important housekeeping and developmental processes. Our analysis on mESC_Ag has shown that DDLFC can extract unique biological insights, often confirmed by literature.

6.5 Discussion

Single cell multi-omics sequencing protocols have recently become a vital tool for exploring the multi-omics landscape. However, inherent limitations in the capturing procedures combined with exponential scaling costs for expanding across layers present major challenges.

In this chapter we argue that inherent single cell mutli-omis protocols limitations, such as limited sensitivity, increase sparsity and undermine downstream analysis. We introduce scMVI, a deep probabilistic model which leverages jointly profiled data from the same cell population to infer a consistent mapping between individual molecular layers and a joint latent representation. This representation is used to accurately impute or integrate partially quantified epigenomic and transcriptomic modalities, post-experimentally addressing protocol limitations. Using features in epigenomic and trascriptomic profiles, scMVI learns a rich latent structure which correlates with phenotype enabling a thorougher and more refined exploratory analysis. By employing dedicated noise models and a deep learning framework, scMVI can model complex genomic region dependencies and account for different sources of technical variability allowing to uncover cell-cell heterogeneity. scMVI's Tensorflow implementation permits scalable inference in existing and future technologies with minimal adaptation.

Our analyses were based on both plate and droplet based protocols to establish that accurate imputation and integration in settings with partially observed data is possible. Furthermore, we introduced DDLFC, a novel alternative to LFC for performing differential

testing. DDLFC combines robustness to label imbalance with the ability to extract unique biological insights. Since DDLFC has been tested only on few data, the introduction of new labelled data will shed light around its strengths and weaknesses.

Chapter 7

Discussion

In the relatively recent past, siblings and mice studies (Dolinoy et al., 2007; Duhl et al., 1994; Kaut et al., 2017) have suggested that an important amount of transcription variation is not attributed to DNA specific effects but to the epigenome. By preventing or facilitating RNA polymerase binding to cis-regulatory regions, the epigenome can affect the differential spatiotemporal expression and eventually control important cell processes like cell differentiation or morphogenesis (Alberts et al., 2007). The advent of single cell multi-omics protocols offers unprecedented opportunities to simultaneously interrogate epigenome and transcriptome at a single cell resolution, allowing for a more in-depth investigation of regulatory mechanisms. This thesis contributes to the development of Bayesian statistical methods to better quantify interactions between components involved in the epigenomic regulation of gene expression.

The technical noise arising in single cell multi-omics assays combined with low coverage, prevent traditional statistical approaches to directly investigate regulatory mechanisms at a genomic region level. To overcome these limitations, Chapter 5 introduces SCRaPL, a Bayesian hierarchical model to infer correlation between epigenomic and transcriptomic layers. Using likelihoods tailored to single cell multi-omics data, SCRaPL accurately quantifies technical variability and accounts for factors like DNA methylation coverage and gene expression zero-inflation. These are often neglected by existing approaches like Pearson correlation. As a result, SCRaPL identifies up to 5 times more genomic regions where the epigenome strongly influences gene regulation compared to traditional tools, while maintaining good control over false positives. When SCRaPL was applied in data from mESC during early gastrulation, we confirmed DNA methylation's strong silencing role in members of the developmental pluripotency family and Dnmt3l, confirming findings of Argelaguet et al. (2019) and Eckersley-Maslin et al. (2019). When the identification of genes with strong regulatory action is combined with enrichment analysis to establish links

with known biological phenomena, a set of biological processes emerge which would be impossible to detect with Pearson correlation, improving interpretative analysis.

Low protocol sensitivity combined with small amount of starting DNA force practitioners to partially or fully discard omics profiles from observations, undermining downstream analysis. In Chapter 6, we introduce scMVI, a multi-modal VAE leveraging jointly profiled data from same cell population to align distinct molecular layers in a joint manifold. Latent cell representations are in turn used to impute or integrate partially profiled cells, remedying the impact of protocol limitations. Dedicated noise models combined with a deep learning framework, enable scMVI to account for multiple sources technical variability allowing to explore cell-cell heterogeneity. By sourcing features from multiple omics modalities, scMVI builds a latent structure which correlates well with phenotype facilitating refined interpretative analyses. Building on latent space's observed properties, we introduce DDLFC a novel estimator alternative to LFC for differential testing. DDLFC blends latent space patterns with robustness to cell imbalance, offering important insights into the underlying biology. A macroscopic investigation of genomic regions with strong DDLFC change using enrichment analysis indicated among others early signs of heart development consistent with findings of Ivanovitch et al. (2017); Tyser et al. (2016), a plethora of house keeping processes and a connection with RNA velocity a prominent method for measuring gene expression kinetics.

SCRaPL and scMVI results lay the baseline for exploring interesting technical and biological questions. In the remaining of this section, we outline potential future directions for expanding this work.

7.1 Future Directions

This thesis has focused on devising models to better understand the regulatory role of epigenome in gene expression. Despite the encouraging results, testing on a broader set of experiments will both help to enhance our understating around complex biological phenomena and models' behaviour. Recently, novel protocols allowed to simultaneously measure histone modification and gene expression in single cell (Moudgil et al., 2020; Rooijers et al., 2019). We could use SCRaPL's Bayesian framework to assess how well histone modification correlates with gene expression and determine if transcription factor binding sites' ability to predict gene expression found in Karlić et al. (2010) holds for single cell data.

Furthermore, both models have relied on a set of assumption which can at times be limiting. For example, SCRaPL ignores complex co-regulation between genomic regions

in favour of scalability while scMVI sacrifices model explainability to improve predictive performance. In the remaining of the section, we expand models' assumptions and suggest potential future directions.

7.1.1 Accounting for genomic region interaction

In Chapter 5, we introduced SCRaPL to robustly quantify regulatory interactions between the epigenome and transcriptome. SCRaPL's genomic region independence has allowed for almost linear scaling but is at odds with gene regulatory mechanisms. In practice, lack of genomic region dependence makes it challenging for SCRaPL to accurately quantify zero-inflation pushing the likelihood to ignore most zero observations. For genes with high average non-zero expression, this is not a significant problem, but there are some edge cases where average expression is inflated potentially affecting posterior correlations. A more realistic assumption would be to allow some information sharing between co-expressed regions, removing some of the influence zero-inflation has on average gene expression.

Early attempts made towards that direction would involve adding a Latent Dirichlet Allocation (Pritchard et al., 2000) module to cluster genomic regions based on their posterior correlation or average DNA methylation and gene expression. However, in our experiments, many of these approaches failed to materialize into meaningful improvements due to a lack of biological insights. More precisely, clustering based on posterior correlation would discriminate between positive and negative correlations. This would mask genes with strong correlation as they were forced into clusters with mildly correlated genes, making any interesting signal extraction impossible. In the model where latent means were clustered, genomic regions with similar average DNA methylation and gene expression would be grouped together sometimes exacerbating the zero-inflation problem.

Numerous approaches have been proposed to model complex regulatory genes interactions. Perhaps the most interesting for our case would be to approximating the gene regulatory network using single cell RNA data (Chan et al., 2017; Jackson et al., 2020) or single cell multi-omics data (Duren et al., 2022; Jin et al., 2021). With minor likelihood adaptations to either make them compatible with plate based protocols or to encourage sparsity, we could create a gene regulatory network to better encapsulate gene dependencies. These could then be encoded into a multi-dimensional Gaussian distribution with block diagonal covariance. To estimate this covariance, we could log-transform the normalized gene expression and estimate relevant mean and covariance terms. In a future extension we could also consider more complex priors involving epigenomic interactions. Assuming J genomic regions, I cells and a subset of co-expressed genomic regions with indices $\{e,u,k,t\}$, then SCRaPL is summarized in Figure 7.1.

Estimating this graphical model will require significantly more computational and memory resources. To effectively explore posteriors with NUTS and avoid memory overload in the naive prior setting, genomic regions would have to be broken in batches such that total inferred parameters would not exceed 85k. In case of ~ 4000 cells per region this would mean the dataset would be broken in batches of ~ 10 genomic regions. Unless strong constrains are placed in the gene regulatory network likelihood to ensure extremely sparse connections, there is no guarantee that the 85k parameter can be satisfied in moderately large datasets with 10000 cells per region. A potential way to exceed this theoretical limit would be to fit one normal per parameter in its unconstrained support using Variational Inference. When we tried using TensorFlow's built in surrogate posterior library in SCRaPL, inference took a fraction of time (~ 15 minutes for Argelaguet et al. (2019) mESC) but the resulting approximations' accuracy negatively correlated with height in the hierarchy. As variables lower in the hierarchy are the most resource intensive to acquire and approximate samples are accurate, we could devise a rejection sampling scheme for extended prior SCRaPL with proposals the Variational approximations.

7.1.2 Transferring knowledge across protocols

In Chapter 6, we introduced scMVI a multi-view deep probabilistic model used for downstream analysis on partially and fully quantified single cell multi-omics data. In the datasets used, scMVI tends to produce latent structures that correlate with cell phenotype, but the large number of parameters and hidden layers prohibit any insight as to what pattern the model has extracted. The large number of parameters also make cross platform generalization hard as it is easier for scMVI to over-fit training data instead of extracting patterns related to underlying biology. Furthermore, the white noise assumption for the latent space is in contrast with meaningful pattern extraction. To avoid some of the problems outlines above for single cell RNA data, Svensson (2020) proposed Linearly Decoded Variational Autoencoder (LDVAE) a factor model based on scVI (Lopez et al., 2018). LDVAE restricts decoder to the set of linear functions allowing result interpretation and reducing the number of parameters.

Introducing a multi-modal model with complex encoders but linear decoders or single cell Linearly Decoder Multi View Inference (scLDMVI) could provide a deep learning framework capable of addressing partially observed data while facilitating explainable pattern extraction. Training scLDMVI will be a two-step procedure. Initially a deep conditional variational autoencoder, similar to scVI, with non-linear encoder and decoder will be trained on single cell RNA data from healthy donors. Such cells could be sourced with droplet or plate based, single or multi-omics protocols and their purpose is to create a rich latent representation. The model will account for dataset differences by conditioning on cell type and sequencing

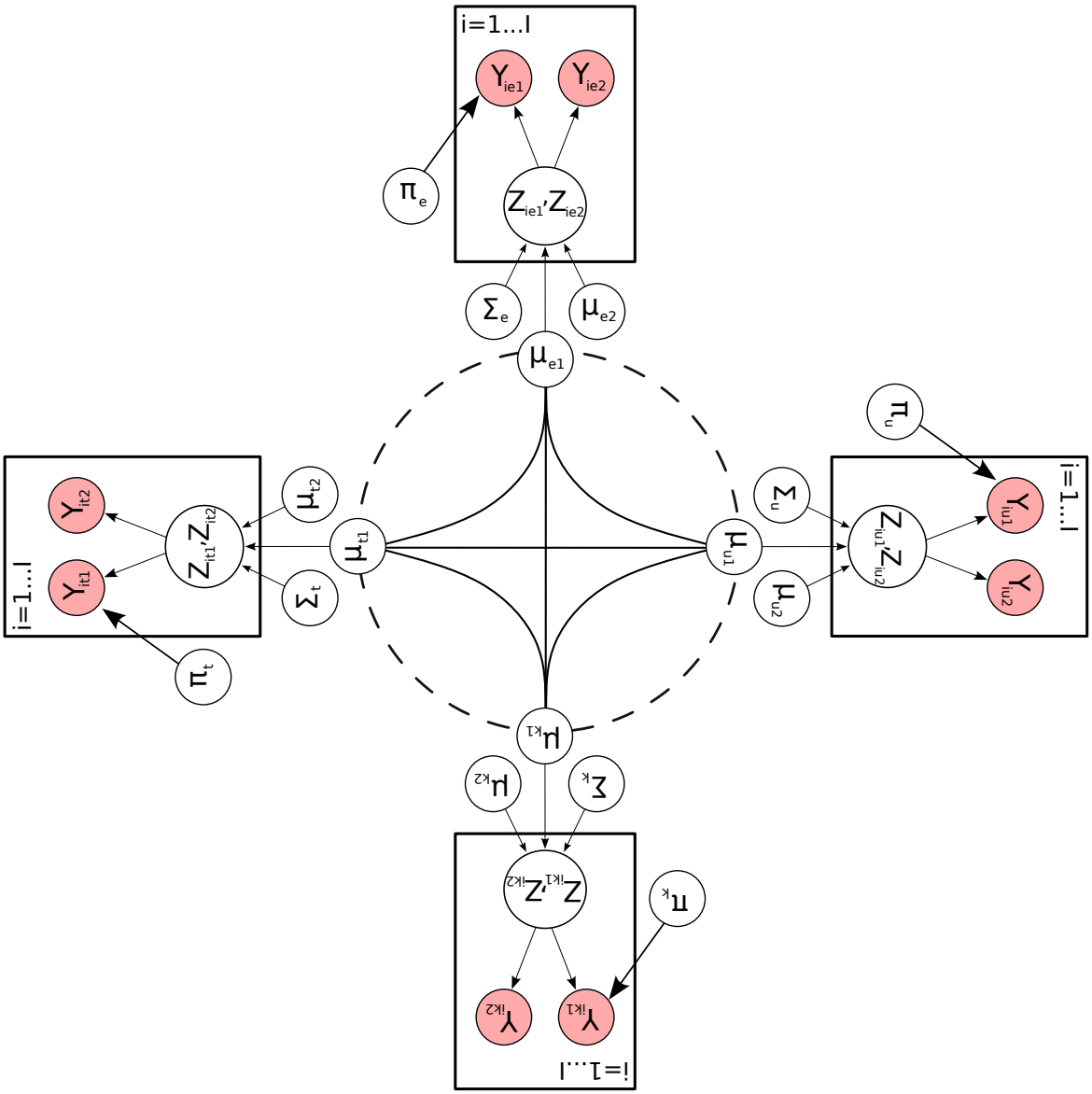


Fig. 7.1 **Schematic summary of SCRaPL with more complex latent mean expression prior. Future work.** Schematic summary of SCRaPL with more complex latent mean expression prior. Dashed circle replaces standard plate notation implying that nodes on the circumference are independent unless explicitly mentioned. For a subset of genomic regions with correlated latent means and indices $\{e,u,k,t\}$, dependencies are marked with solid lines.

technology. Encoder depth and number of parameters will be tuned such that the latent space structure arranges cells based on their phenotypic properties as in scMVI. In the encoder pretraining step we avoid the linear decoder to allow for more complex latent structure. In turn, this complex encoder is used as prior to guide training of epigenomic conditional VAEs

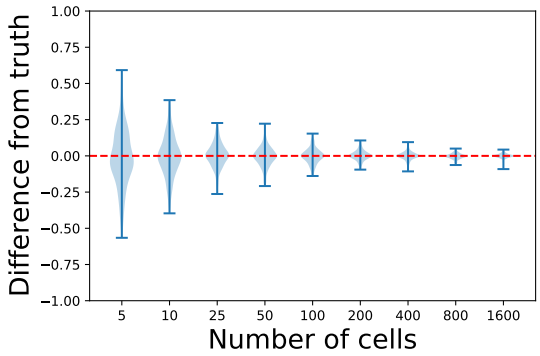
along with the transcriptomic linear decoder. For that step training will be restricted only on jointly profiled data.

The suggested training strategy has multiple advantages compared to scMVI or related methods (i.e. MultiVI,scMM and scMVAE) training. Building a prior using that approach leverages information from a substantially larger set which is impossible to get from existing single cell multi-omics data. Furthermore, it allows new single cell RNA data integration without the need for expensive retraining or share restricted data as demonstrated in Lotfollahi et al. (2022). Since many multi-omics protocols assay gene expression, having a prior that encapsulates information from multiple sources allows for a series of interesting tasks such as cross technology transfer and imputation. Currently methods for plate based epigenomic data rely on Binomial distribution for inference (Argelaguet et al., 2020; Kapourani and Sanguinetti, 2019; Maniatis et al., 2022), while droplet use zero inflated Negative Binomial (i.e. scVI, MultiVI, scMM). To the best of our knowledge there has not been any published research proposing likelihoods that could work with both sequencing technologies. A potential work-around for cross technology information transfer and imputation would be to train two versions of scLDMVI, one for droplet and another for plate based data with both having a common transcriptomic layer prior. Then chromatin accessibility quantified using a droplet based multi-omics protocol for some cell is projected using relevant scLDMVI encoder. In turn, plate based decoder conditioned on the correct type of cell could be used to achieve our goal.

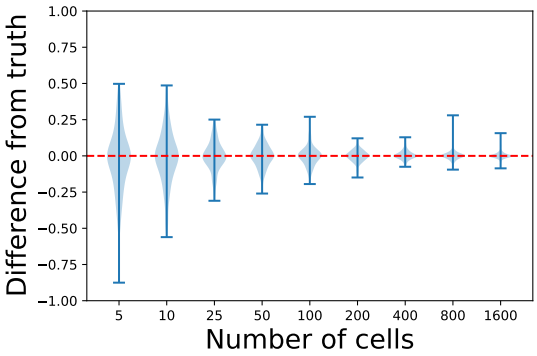
Appendix A

Supplementary information for SCRaPL

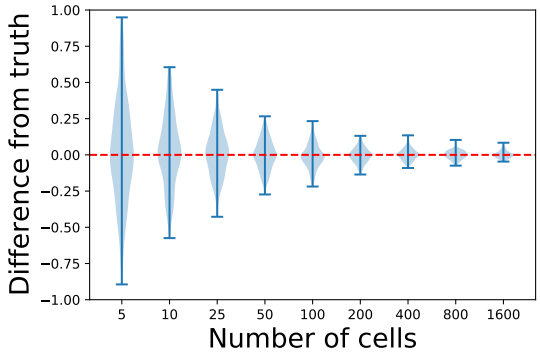
A.1 Synthetic Data



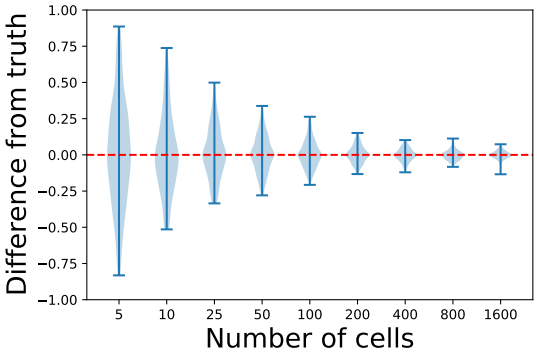
(a)



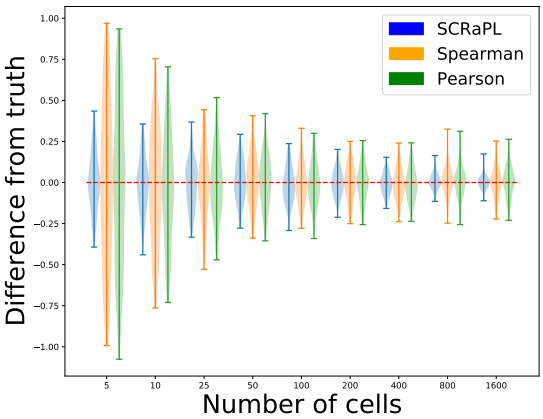
(b)



(c)

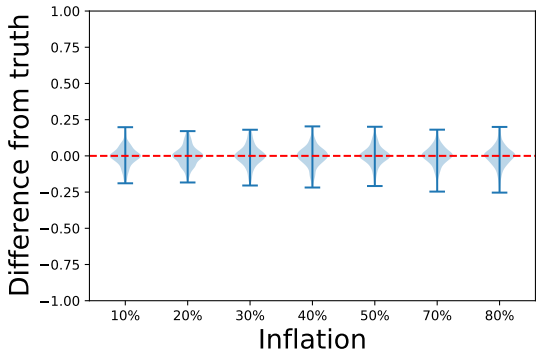


(d)

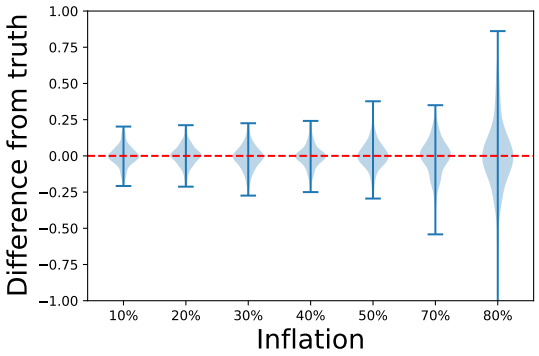


(e)

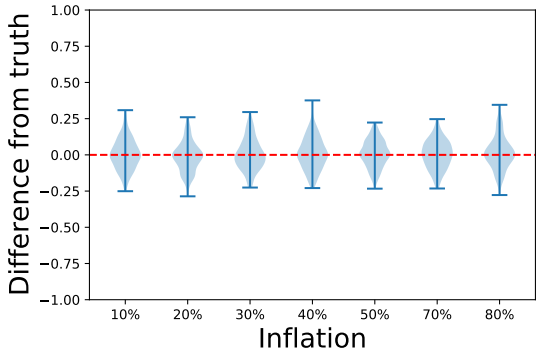
Fig. A.1 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 1. Figures used to demonstrate model behaviour under synthetic data generated from the model.**. Differences of posterior estimates from data generating parameters as functions of the number of cells across features for: (A.1a) methylation mean; (A.1b) expression mean; (A.1c) methylation standard deviation; (A.1d) expression standard deviation; (A.1e) correlation between expression and methylation.



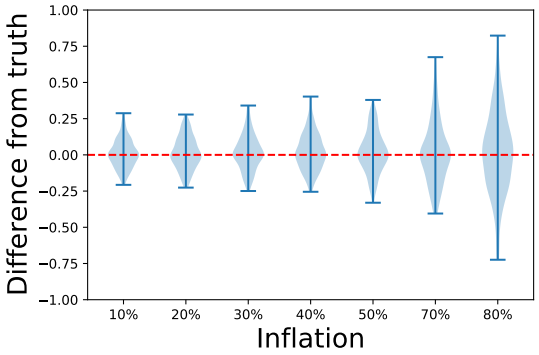
(a)



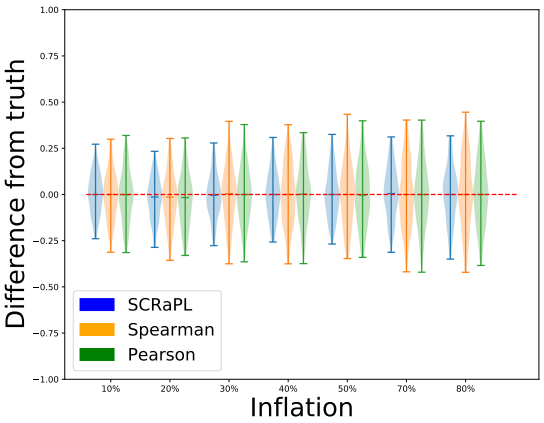
(b)



(c)



(d)



(e)

Fig. A.2 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 2. Figures used to demonstrate model behaviour under synthetic data generated from the model.**. Differences of posterior estimates from data generating parameters as functions of average expression inflation across features for: (A.2a) methylation mean; (A.2b) expression mean; (A.2c) methylation standard deviation; (A.2d) expression standard deviation; (A.2e) correlation between expression and methylation.

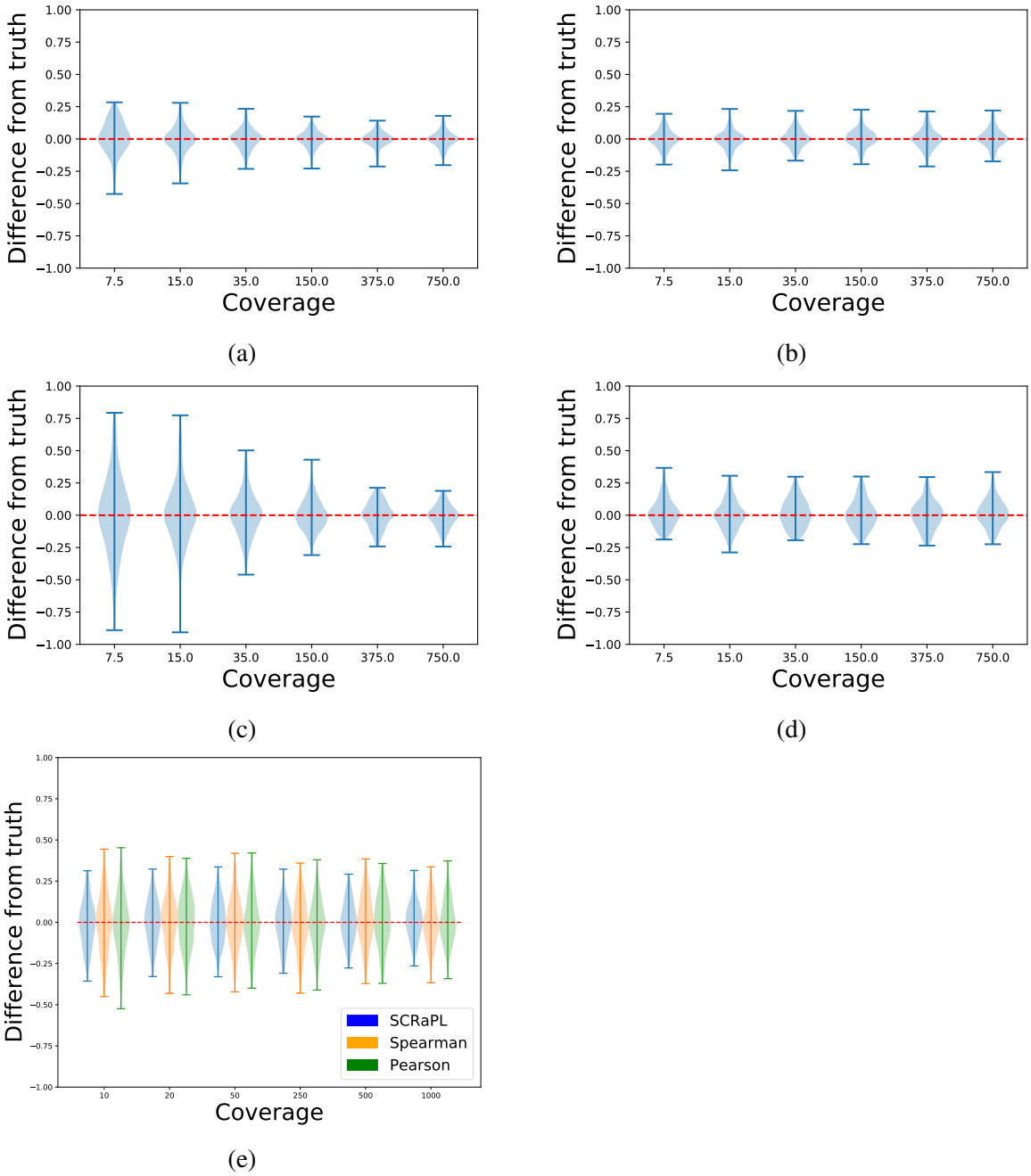


Fig. A.3 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 3. Figures used to demonstrate model behaviour under synthetic data generated from the model.**. Differences of posterior estimates from data generating parameters as functions of average methylation coverage across features for: (A.3a) methylation mean; (A.3b) expression mean; (A.3c) methylation standard deviation; (A.3d) expression standard deviation; (A.3e) correlation between expression and methylation.

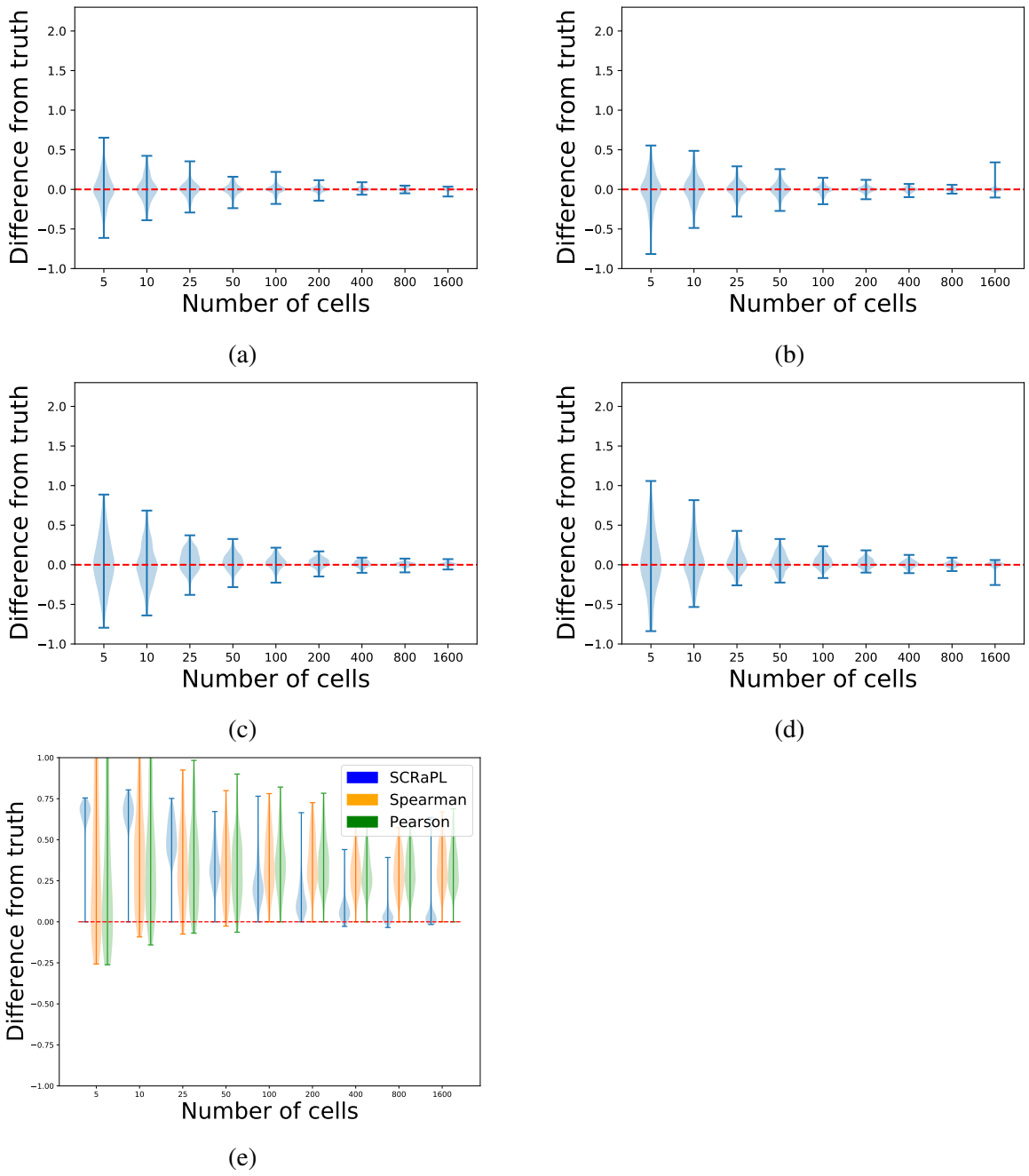
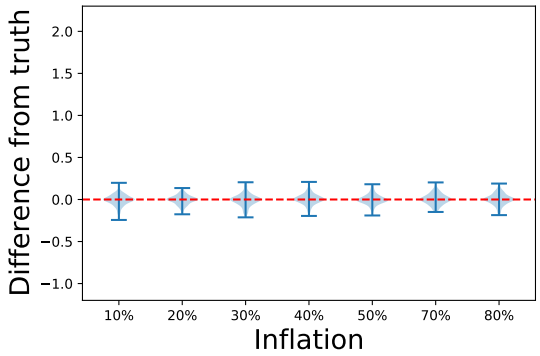
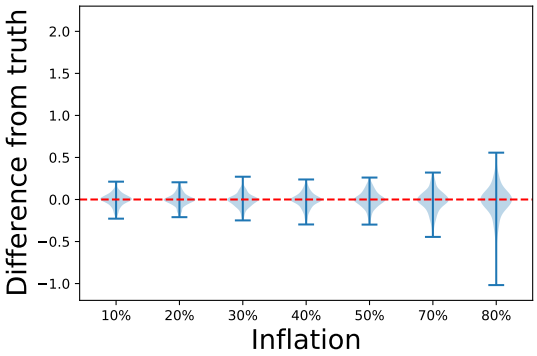


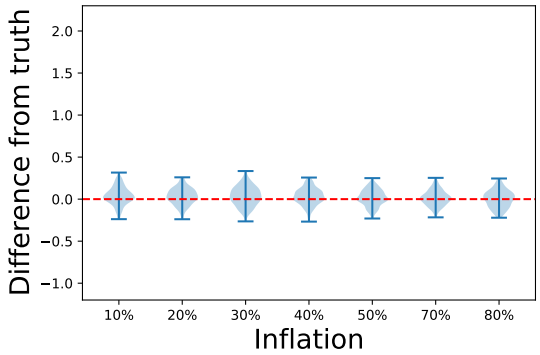
Fig. A.4 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 4. Figures used to demonstrate model behaviour under synthetic data generated from the model with different prior.**. Differences of posterior estimates from data generating parameters as functions of the number of cells across features for: (A.4a) methylation mean; (A.4b) expression mean; (A.4c) methylation standard deviation; (A.4d) expression standard deviation; (A.4e) correlation between expression and methylation.



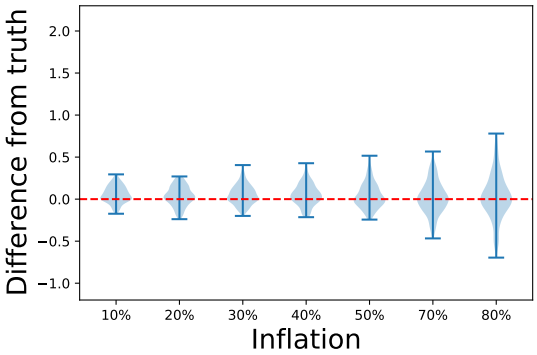
(a)



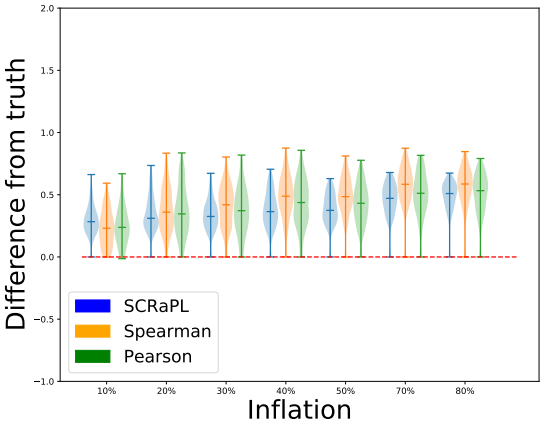
(b)



(c)



(d)



(e)

Fig. A.5 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 5. Figures used to demonstrate model behaviour under synthetic data generated from the model with different prior.**. Differences of posterior estimates from data generating parameters as functions of average expression inflation across features for: (A.5a) methylation mean; (A.5b) expression mean; (A.5c) methylation standard deviation; (A.5d) expression standard deviation; (A.5e) correlation between expression and methylation.

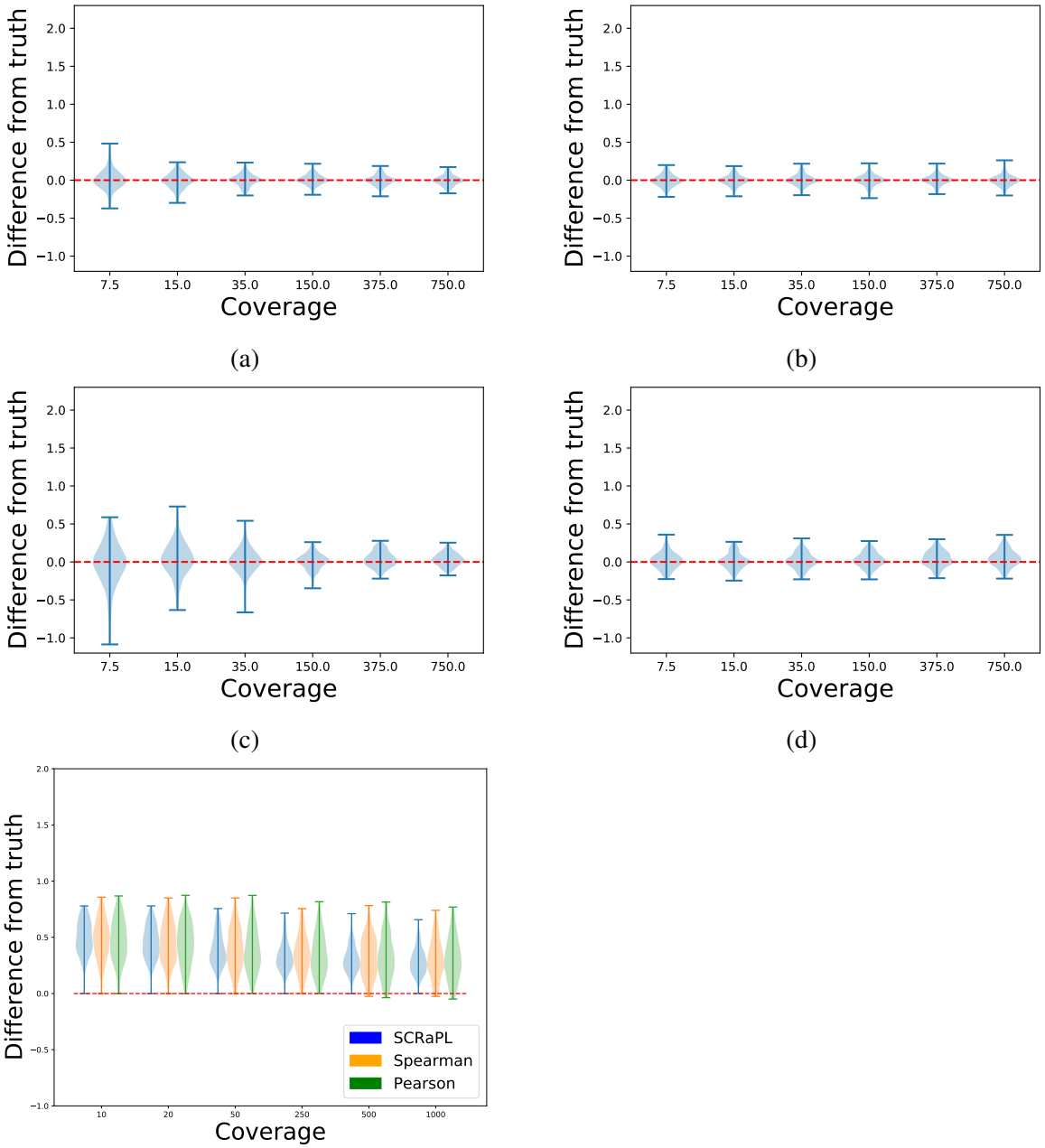
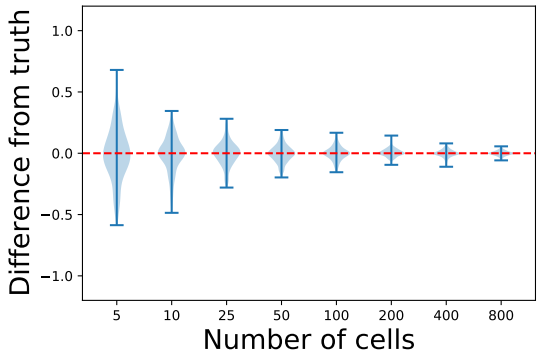
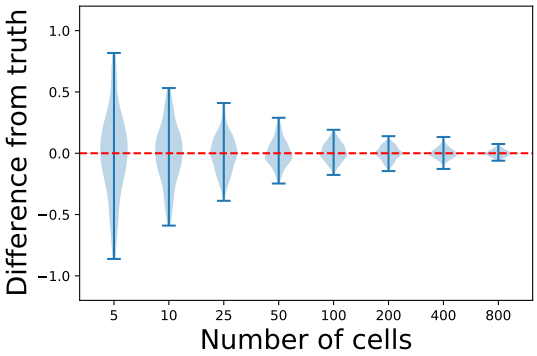


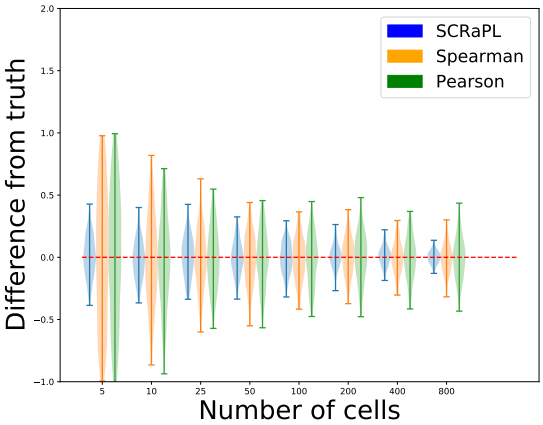
Fig. A.6 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 6. Figures used to demonstrate model behaviour under synthetic data generated from the model with different prior.**. Differences of posterior estimates from data generating parameters as functions of average methylation coverage across features for: (A.6a) methylation mean; (A.6b) expression mean; (A.6c) methylation standard deviation; (A.6d) expression standard deviation; (A.6e) correlation between expression and methylation.



(a)

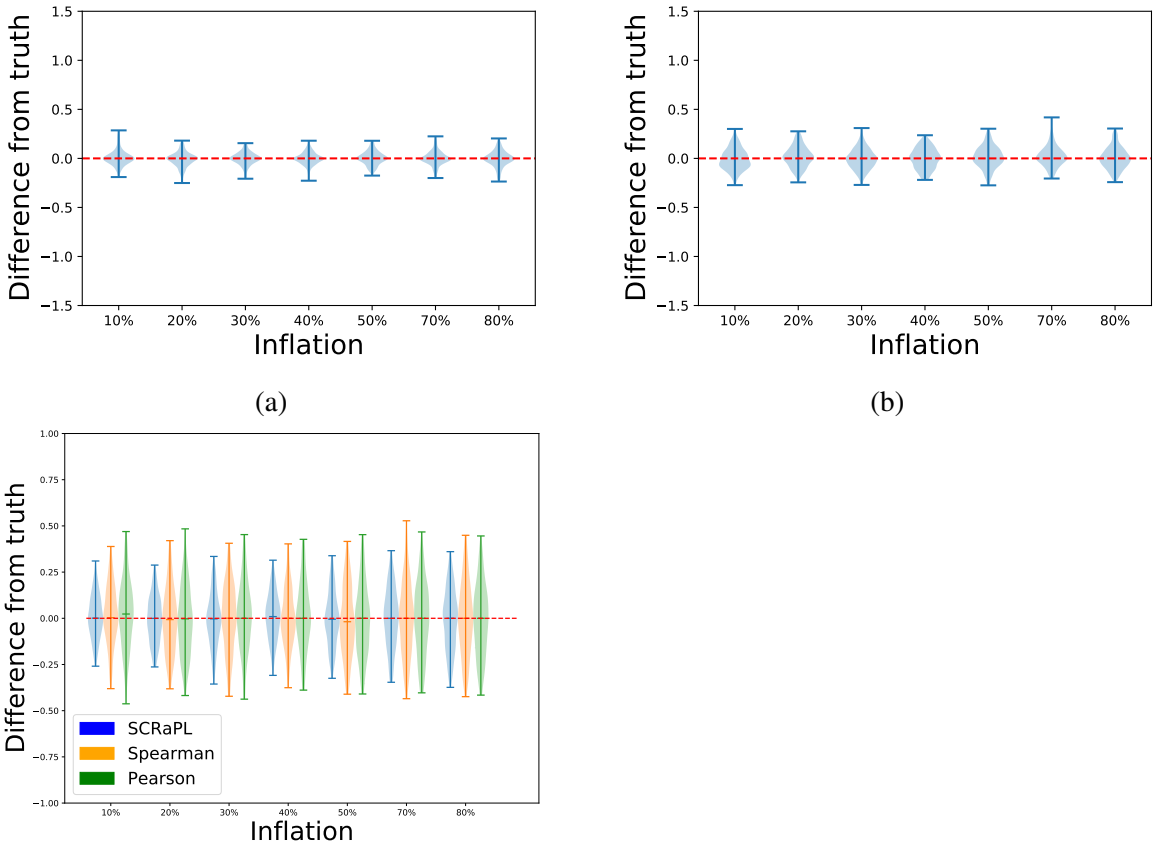


(b)



(c)

Fig. A.7 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 7. Figures used to demonstrate model behaviour under synthetic data generated under model mis-specification.** . Differences of posterior estimates from data generating parameters as functions of the number of cells across features for: (A.7a) methylation mean; (A.7b) methylation standard deviation; (A.7c) correlation between expression and methylation.



(a) (b)

(c)

Fig. A.8 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 8. Figures used to demonstrate model behaviour under synthetic data generated under model mis-specification.** . Differences of posterior estimates from data generating parameters as functions of average expression inflation across features for: (A.8a) methylation mean; (A.8b) methylation standard deviation; (A.8c) correlation between expression and methylation.

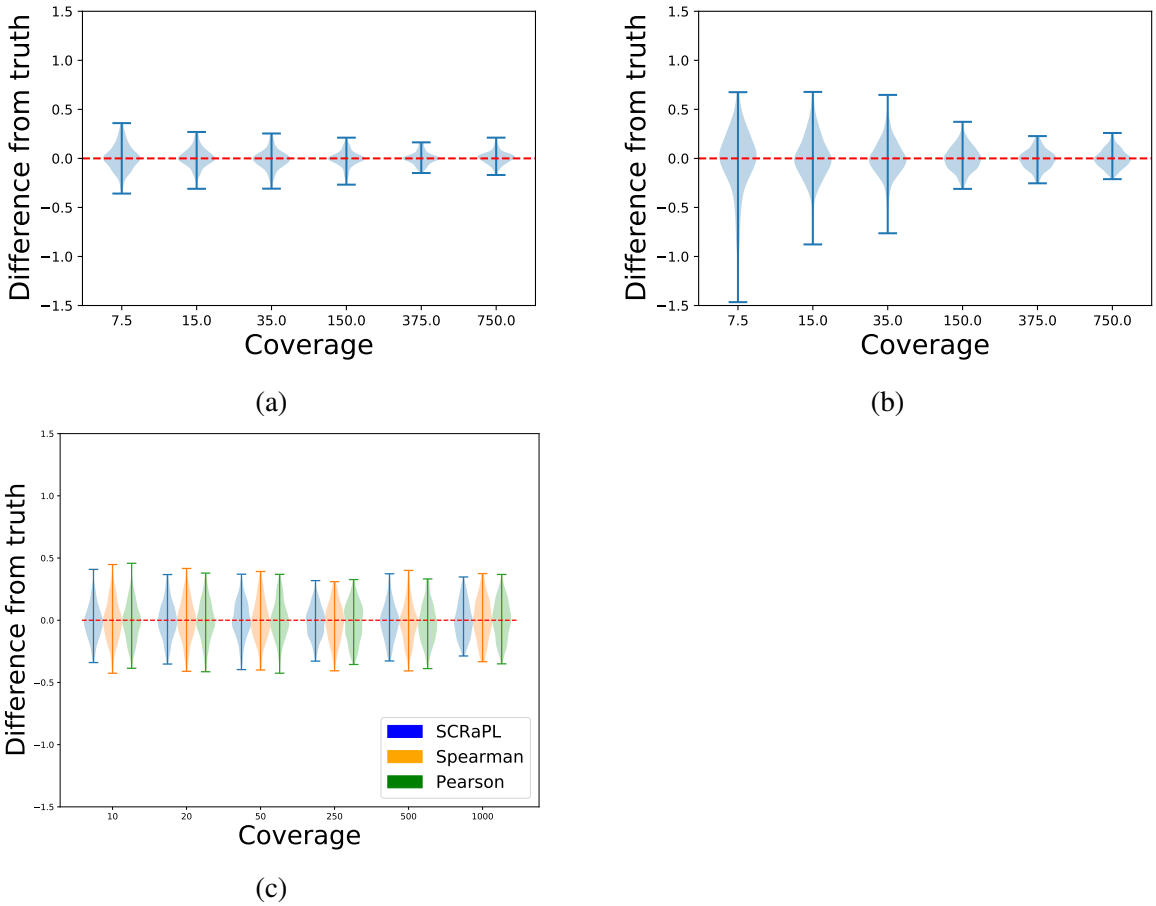


Fig. A.9 **Violin plots summarizing difference of posterior mean from data generating parameters across features for Experiment 9. Figures used to demonstrate model behaviour under synthetic data generated under model mis-specification.** . Differences of posterior estimates from data generating parameters as functions of average methylation coverage across features for: (A.6a) methylation mean; (A.9b) methylation standard deviation; (A.6e) correlation between expression and methylation.

A.2 Negative control experiments

He we provide additional negative control experiments omited from the main text. More precisely we include negative control experiments on synthetic and chromatin accessibility/-gene expression layers of mESC data. Along with that, we include Bayesian and frequentist volcano plots for the mESC and mEBC data, created using Spearman correlation.

	Orig.	Neg.1	Neg.2	Neg.3	Neg.4	Neg.5
SCRaPL	0	0	0	0	0	0
Pearson	0	0	0	0	0	0
Spearman	0	0	0	0	0	0

Table A.1 **Number of detected features for original and negative control data with SCRaPL, Pearson and Spearman correlation in accessibility-expression modalities of mESC data data when** $\gamma = 0.28$

	Orig.	Neg.1	Neg.2	Neg.3	Neg.4	Neg.5
SCRaPL	715	0	1	4	0	0
Pearson	359	0	0	0	0	0
Spearman	408	0	0	0	0	0

Table A.2 **Number of detected features for original and negative control data with SCRaPL, Pearson and Spearman correlation for synthetic data generated from SCRaPL when** $\gamma = 0.115$

In accessibility-expression layers of mESC data there does not seem to be enough signal for SCRaPL to yield any useful results.

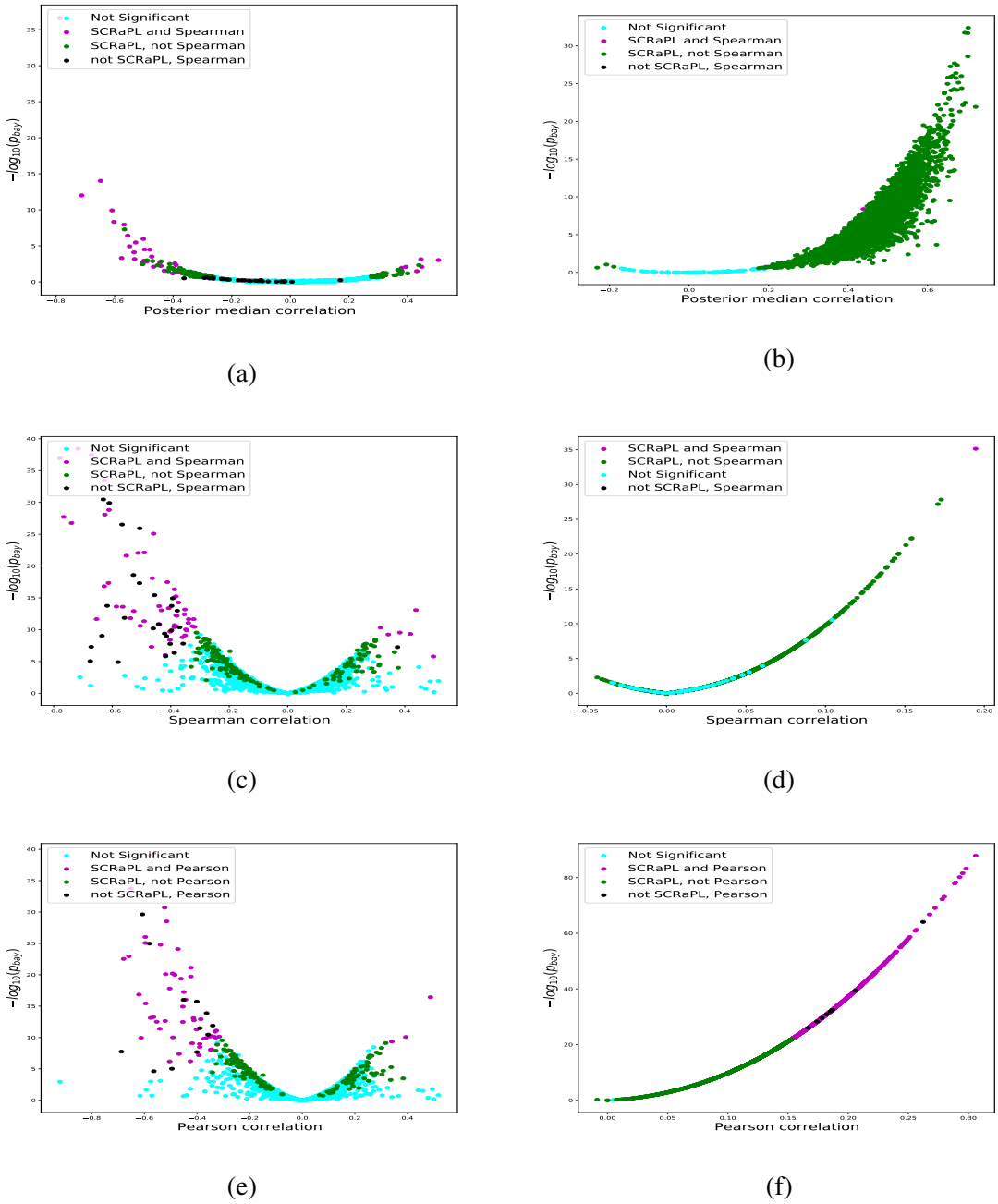


Fig. A.10 **Bayesian and frequentist versions of volcano plots for the mESC and mEBC data. Plots demonstrating that SCRaPL can significantly improve detection rates, by choosing features picked up by Pearson and Spearman correlation while avoiding spurious correlations.** Each dot represents a feature and is marked with different color depending the method that labels it as a significant association. (A.10a) Bayesian volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mESC data. (A.10b) Bayesian volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mEBC data. (A.10c) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mESC data. (A.10d) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Spearman correlation in mEBC data. (A.10e) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Pearson correlation in mESC data. (A.10f) Frequentist volcano plot summarizing significant features as detected by SCRaPL and Pearson correlation in mEBC data.

A.3 Extended comparisons between SCRaPL, Pearson and Spearman predictions.

In the process of understanding SCRaPL's behavior, we analyzed results in a series of features. We consider the set of associations which are called as significant by at least one method, and split it into 3 categories: agreement between predictions, association labeling as significant by SCRaPL, but not Pearson, and vice-versa. Here we go through some examples omitted from the main text, which demonstrate SCRaPL's superior performance. Since for Spearman correlation there is only only one significant feature, in the relevant subsections we only rely on Pearson for comparison.

A.3.1 Classified as significant by SCRaPL and Pearson/Spearman

In this section we present some cases where SCRaPL agrees with Pearson in both mESC and mEBC data. For many of the examples we have good coverage for both methylation and relatively high accessibility/expression. Furthermore in Argelaguet et al. (2019) we tend to have observations for a wide range of methylation and expression values.

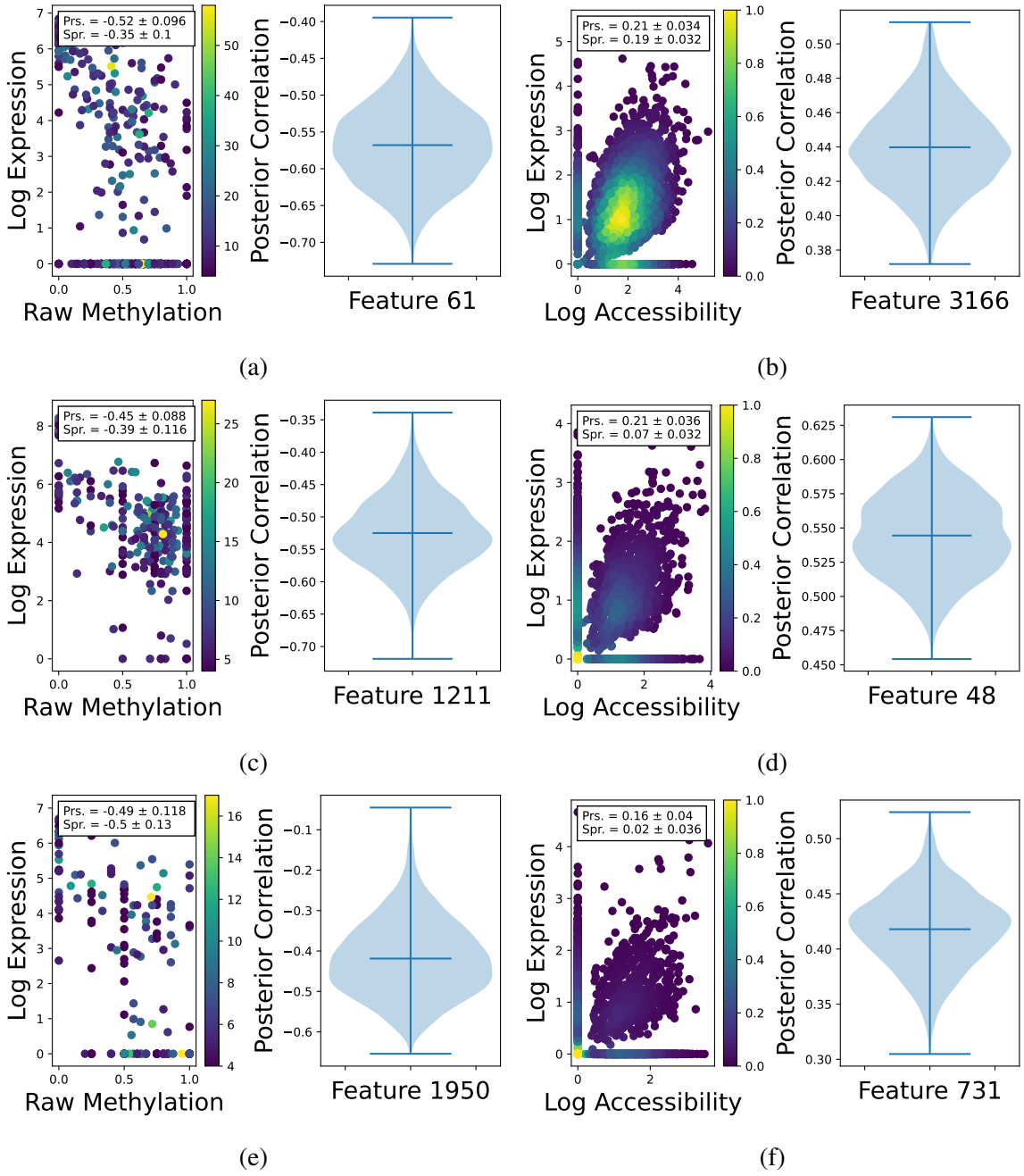


Fig. A.11 **Example where both SCRaPL and Pearson identify feature's association as significant in mESC and mEBC data. Examples used to identify common patterns between SCRaPL and Pearson correlation.** In all Figures the left part is a scatter plot of feature's raw data and the right part is the posterior correlation as inferred by SCRaPL. In features from mESC data (i.e. (A.11a), (A.11c) and (A.11e)) each dot of the scatter plot represents a cell reading, color-coded by CpG coverage and the expression axis is in $log(1+x)$ scale. In features from mEBC data (i.e. (A.11b), (A.11d) and (A.11f)) each dot of the scatter plot represents a cell reading, color-coded by space occupation and the accessibility/expression axis are in $log(1+x)$ scale.

A.3.2 Classified as insignificant by SCRaPL and significant by Pearson

In this subsection we generally have features with observations leading to spurious correlations. Some of these problems encountered in observations include, low expression/ number of observations, extremely low coverage and large number of zeros. For many of these features we see that are large areas where we do not have any observation for both methylation and expression. In mEBC data, we see an imbalanced ratio of readings with zero accessibility and non-zero expression.

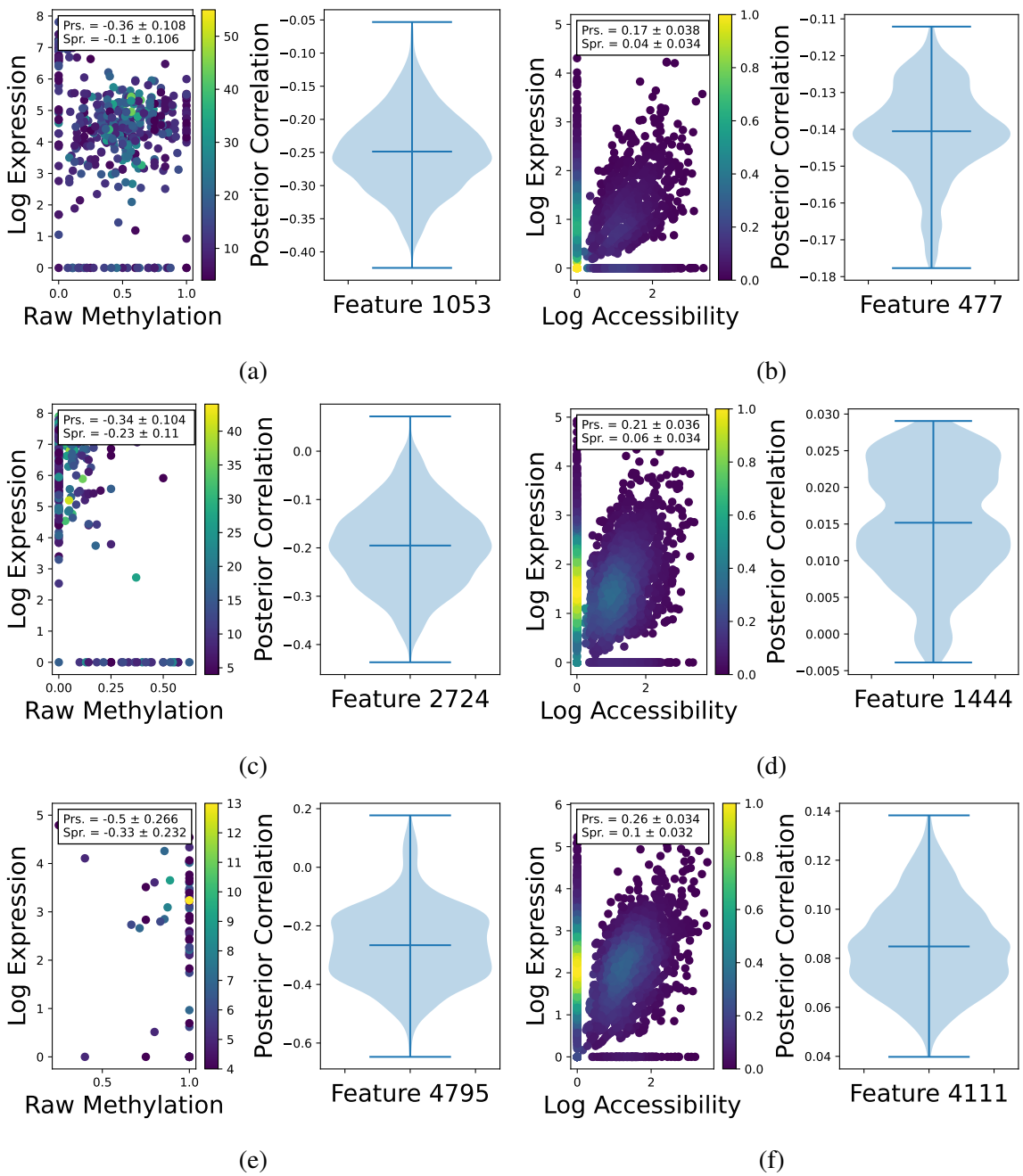


Fig. A.12 **Example where only Pearson identifies the feature's association as significant in mESC and mEBC data. Examples used to understand why SCRaPL avoids Pearson correlation limitations.** In all Figures the left part is a scatter plot of feature's raw data and the right part is the posterior correlation as inferred by SCRaPL. In features from Argelaguet et al. (2019) data (i.e. (A.12a), (A.12c) and (A.12e)) each dot of the scatter plot represents a cell reading, color-coded by CpG coverage and the expression axis is in $log(1+x)$ scale. In features from mEBC data (i.e. (A.12b), (A.12d) and (A.12f)) each dot of the scatter plot represents a cell reading, color-coded by space occupation and the accessibility/expression axis are in $log(1+x)$ scale.

A.3.3 Classified as significant by EFDR and insignificant by FDR

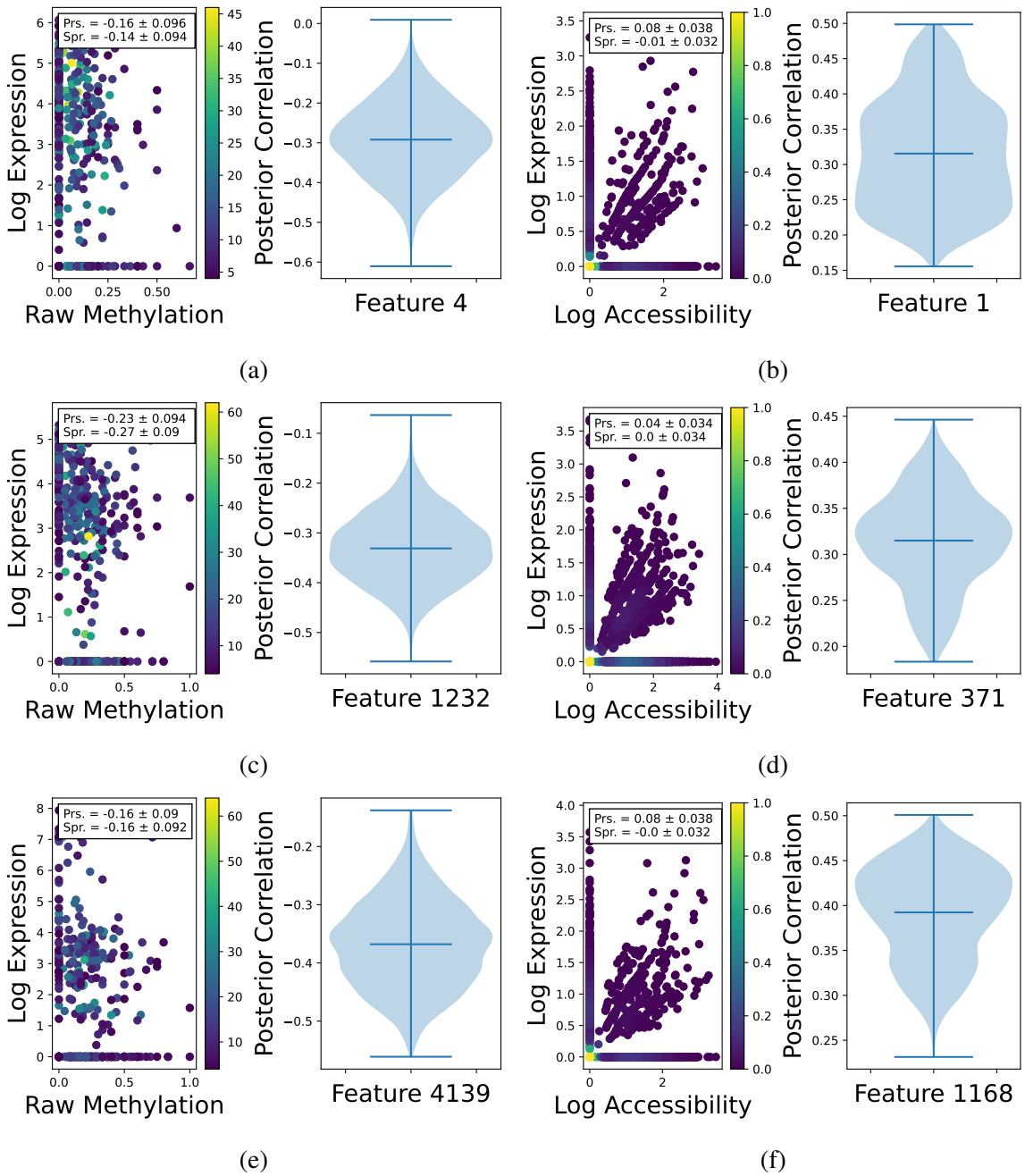


Fig. A.13 **Example where only SCRaPL identifies the feature's association as significant in mESC and mEBC data. Examples used to understand why SCRaPL avoids Pearson correlation limitations.** In all Figures the left part is a scatter plot of feature's raw data and the right part is the posterior correlation as inferred by SCRaPL. In features from Argelaguet et al. (2019) data (i.e. (A.13a), (A.13c) and (A.13e)) each dot of the scatter plot represents a cell reading, color-coded by CpG coverage and the expression axis is in $log(1+x)$ scale. In features from mEBC data (i.e. (A.13b), (A.13d) and (A.13f)) each dot of the scatter plot represents a cell reading, color-coded by space occupation and the accessibility/expression axis are in $log(1+x)$ scale.

A.3.4 Posterior correlation inference for common genes in bibliography

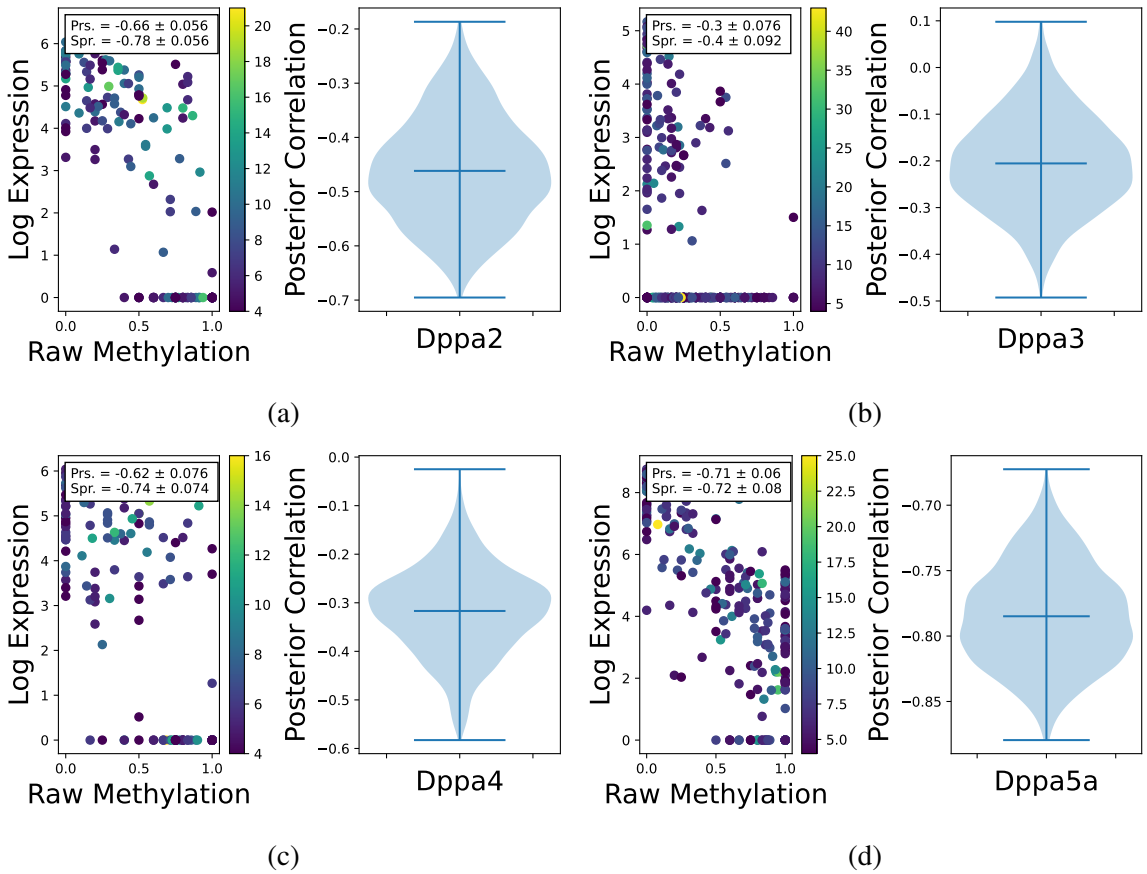
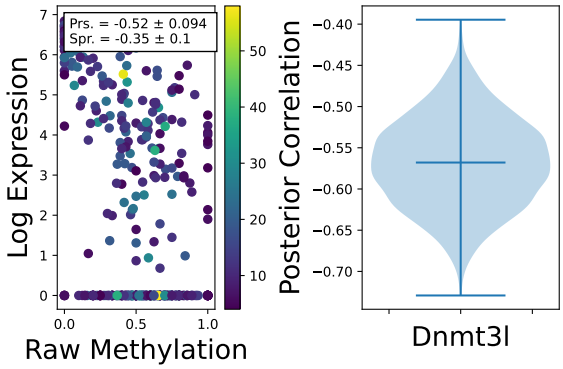
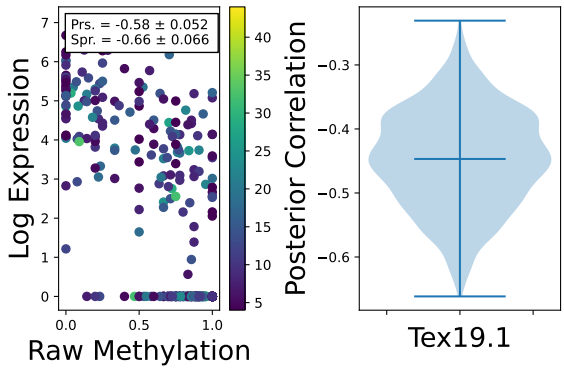


Fig. A.14 **Raw data and posterior gene correlation for selected members of the Dppa gene family, inferred with SCRaPL. Examples used to demonstrate that SCRaPL identifies genes with known role in early development.**

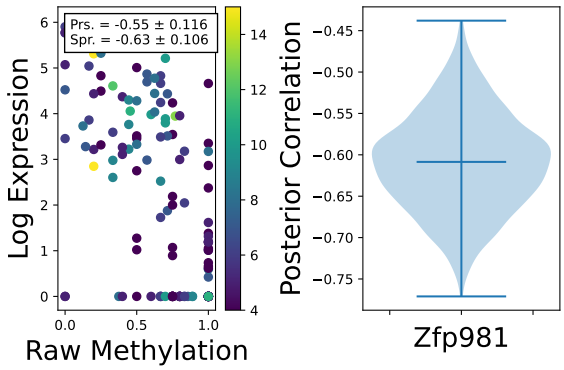


(a)

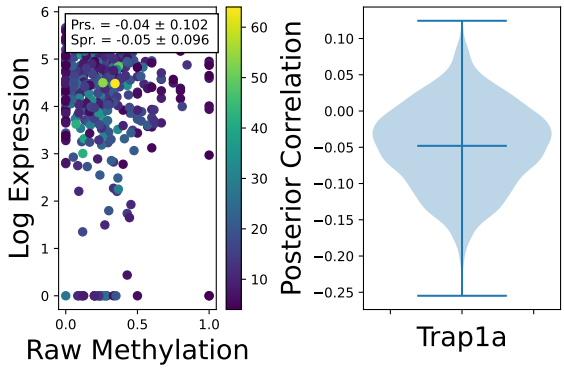
Fig. A.15 **Raw data and posterior gene correlation for selected members of the Dnmt gene family, inferred with SCRaPL. Example used to demonstrate that SCRaPL identifies genes with known role in early development.**



(a)



(b)



(c)

Fig. A.16 **Raw data and posterior gene correlation for other genes in Argelaguet et al. (2019) that have been partially or not studied, inferred with SCRaPL. Examples where SCRaPL has identified genes of potential interest.**

A.4 Gene Set Enrichment Analysis.

To get a rounded understanding of the genes with strong regulatory action identified by alternative approaches, we carry our Gene Set Enrichment Analysis (GSEA) analysis using DAVID (Huang, 2009). This process helps us to identify important biological processes by looking at over-represented genes in a starting pool of genetic markers. Using SCRaPL correlation outcomes we can determine a lists of biological outcomes that practitioners would discover. The γ threshold for SCRaPL was set to 0.205 or 90% quantile of the folded distribution constructed by correlation samples of the permuted dataset.

Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Background_2
214 DAVID IDs

Options

Thresholds: Count 7 EASE 0.3
Display: ☑ Fold Enrichment ☐ Bonferroni ☐ Benjamini ☐ FDR ☐ Fisher Exact ☐ LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

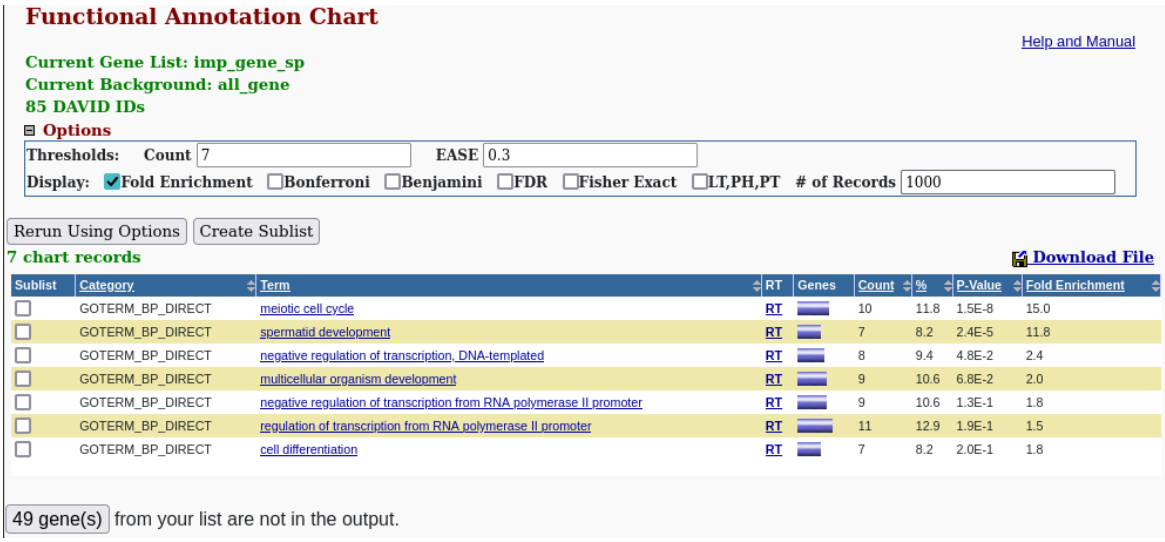
15 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	angiogenesis	RT		7	3.3	5.4E-2	2.6
☐	GOTERM_BP_DIRECT	proteolysis	RT		13	6.1	1.4E-2	2.2
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		10	4.7	5.4E-2	2.1
☐	GOTERM_BP_DIRECT	ion transport	RT		8	3.7	1.3E-1	1.9
☐	GOTERM_BP_DIRECT	spermatogenesis	RT		8	3.7	1.6E-1	1.8
☐	GOTERM_BP_DIRECT	response to drug	RT		7	3.3	2.1E-1	1.7
☐	GOTERM_BP_DIRECT	cell adhesion	RT		7	3.3	2.3E-1	1.7
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		13	6.1	8.0E-2	1.7
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		10	4.7	1.5E-1	1.7
☐	GOTERM_BP_DIRECT	metabolic process	RT		10	4.7	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	RNA splicing	RT		8	3.7	2.5E-1	1.6
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		20	9.3	5.9E-2	1.5
☐	GOTERM_BP_DIRECT	oxidation-reduction process	RT		13	6.1	2.1E-1	1.4
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		10	4.7	2.8E-1	1.4
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		39	18.2	1.4E-1	1.2

89 gene(s) from your list are not in the output.

(a)

Fig. A.17 **GO analysis with features detected by SCRaPL. Figures used to demonstrate early developmental stage biological processes identified by SCRaPL.**



Functional Annotation Chart

Help and Manual

Current Gene List: imp_gene_sp
Current Background: all_gene
85 DAVID IDs

Options

Thresholds: Count 7 EASE 0.3
Display: ☑ Fold Enrichment ☐ Bonferroni ☐ Benjamini ☐ FDR ☐ Fisher Exact ☐ LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

7 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	meiotic cell cycle	RT		10	11.8	1.5E-8	15.0
☐	GOTERM_BP_DIRECT	spermatid development	RT		7	8.2	2.4E-5	11.8
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		8	9.4	4.8E-2	2.4
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		9	10.6	6.8E-2	2.0
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		9	10.6	1.3E-1	1.8
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		11	12.9	1.9E-1	1.5
☐	GOTERM_BP_DIRECT	cell differentiation	RT		7	8.2	2.0E-1	1.8

49 gene(s) from your list are not in the output.

(a)

Fig. A.18 **GO analysis with features detected by Spearman correlation. Figures used to demonstrate early developmental stage biological processes identified by Spearman correlation.**

The detected biological presented in Figures A.17,A.18 have at 7 or more genes associated with them and maximum p-value 0.3. Then processes were sorted based on their enrichment score (i.e. how many times larger is the detected set compared to a random subset of the pool of genes linked to a particular biological process). SCRaPL has detected many processes directly linked to regulation of transcription and regulation of transcription in promoter regions (something expected as we are looking at promoters of a methylation-expression pair at early development where methylation plays a crucial role). Apart from them there are also biological processes linked to development, like in utero embryonic development and angiogenesis with high enrichment score. Spearman also detects processes linked to transcription regulation like negative regulation of transcription or meiotic cell cycle as seen in Figure A.18. However no process with immediate links to development appears. For the exact same filtering parameter and genes detected with Pearson correlation, the enrichment would link genes to "regulation of transcription" with enrichment score 1.5.

A.5 Mathematical Derivations

In the field of transcriptomics it is widely accepted that over-dispersion is an important feature of state of the art models (Svensson, 2020). Following the example of several papers (e.g, He et al. (2021); Svensson (2020); Vallejos et al. (2016)) where they provide evidence

that their model can handle excessive zeros by using over-dispersion, we demonstrate that zero-inflated Poisson is also a valid alternative as the over-dispersion property exists. For the sake of completeness we provide mean and covariance formulas for the error models we used.

Let Y_{ij1} and Y_{ij2} be the variables used to model transcriptomic and epigenomic molecular layers respectively. Then we have latent variables Z_{ij1}, Z_{ij2} which associate with Y_{ij1}, Y_{ij2}, feature-specific model parameters (μ_j, Σ_j) and parameter π_j specific to the noise model that is assigned to RNA expression data and captures zero inflation. Parameters s_i and n_{ij} are secondary parameters which we leave their definition for later in the section.

$$\begin{aligned} &\mathbb{P}\left(Y_{ij1}, Y_{ij2} | s_i, n_{ij}\right) = \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{P}\left(Y_{ij1}, Z_{ij1} | s_i\right) \mathbb{P}\left(Y_{ij2}, Z_{ij2} | n_{ij}\right) \mathbb{N}\left(Z_{ij}, \mu_j, \Sigma_j\right) dZ_{ij1} dZ_{ij2}, \\ &s_i \geq 0, n_{ij} > 0 \end{aligned} \tag{A.1}$$

By marginalizing Y_{ij1} and Y_{ij2} we get the following conditionals.

$$\mathbb{P}\left(Y_{ij1} | s_i\right) = \int_{-\infty}^{\infty} \mathbb{P}\left(Y_{ij1}, Z_{ij1} | s_i\right) \mathbb{N}\left(Z_{ij1}, \mu_{j1}, \sigma_{j1}^2\right) dZ_{ij1} \tag{A.2}$$

$$\mathbb{P}\left(Y_{ij2} | n_{ij}\right) = \int_{-\infty}^{\infty} \mathbb{P}\left(Y_{ij2}, Z_{ij2} | s_i\right) \mathbb{N}\left(Z_{ij2}, \mu_{j2}, \sigma_{j2}^2\right) dZ_{ij2} \tag{A.3}$$

Up to this point we have only assumed that Y_{ij1}, Y_{ij2} depend on s_i and n_{ij}, but it is important to note that depending the error model we might change n_{ij} with $\tilde{s}_i$. We start the analysis by considering an error model directed for NMT data. This means for expression data $(\mathbb{P}\left(Y_{ij1}, X_{ij1} | s_i\right))$ following a ZIP model with exponential link and DNAm $(\mathbb{P}\left(Y_{ij2}, X_{ij2} | n_{ij}\right))$ a binomial distribution we get the following conditional expectations and variances:

$$\mathbb{E}\left(Y_{ij1} | s_i\right) = s_i \left(1 - \pi_j\right) e^{\mu_{j1} + \frac{\sigma_{j1}^2}{2}} \tag{A.4}$$

$$\mathbb{V}\left(Y_{ij1} | s_i\right) = \mathbb{E}\left(Y_{ij1} | s_i\right) \left(1 + \mathbb{E}\left(Y_{ij1} | s_i\right) \frac{e^{\sigma_{j1}^2} - 1 + \pi_j}{1 - \pi_j}\right) \tag{A.5}$$

$$\mathbb{E}\left(\frac{Y_{ij2}}{n_{ij}} | n_{ij}\right) = \Phi\left(\frac{\mu_{j2}}{\sqrt{1 + \sigma_{j2}^2}}\right) \tag{A.6}$$

$$\mathbb{V}\left(\frac{Y_{ij2}}{n_{ij}}|n_{ij}\right) = \Phi\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)\left[1-\Phi\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)\right] \\ -2\left(1-\frac{1}{n_{ij}}\right)\mathbb{T}\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}, \frac{1}{\sqrt{1+2\sigma_{j2}^2}}\right) \tag{A.7}$$

Using Equation A.1 we estimate the off diagonal term of the covariance matrix:

$$\mathbb{CV}\left(Y_{ij1}, \frac{Y_{ij2}}{n_{ij}}|s_i, n_{ij}\right) = s_i\left(1-\pi_j\right)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}}\left[\Phi\left(\frac{\mu_{j2}+\rho_j\sigma_{j1}\sigma_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)-\Phi\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)\right] \tag{A.8}$$

For the error model aimed at 10X data, we get the following Expectations/Covariace for expression ($\mathbb{P}\left(Y_{ij1}, X_{ij1}|s_i\right)$) and accessibility ($\mathbb{P}\left(Y_{ij2}, X_{ij2}|\tilde{s}_i\right)$) data.

$$\mathbb{E}\left(Y_{ij1}|s_i\right) = s_i\left(1-\pi_j\right)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}} \tag{A.9}$$

$$\mathbb{E}\left(Y_{ij2}|\tilde{s}_i\right) = \tilde{s}_i\left(1-\tilde{\pi}_j\right)e^{\mu_{j2}+\frac{\sigma_{j2}^2}{2}} \tag{A.10}$$

$$\mathbb{V}\left(Y_{ij1}|s_i\right) = \mathbb{E}\left(Y_{ij1}|s_i\right)\left(1+\mathbb{E}\left(Y_{ij1}|s_i\right)\frac{e^{\sigma_{j1}^2}-1+\pi_j}{1-\pi_j}\right) \tag{A.11}$$

$$\mathbb{V}\left(Y_{ij2}|\tilde{s}_i\right) = \mathbb{E}\left(Y_{ij2}|\tilde{s}_i\right)\left(1+\mathbb{E}\left(Y_{ij2}|\tilde{s}_i\right)\frac{e^{\sigma_{j2}^2}-1+\tilde{\pi}_j}{1-\tilde{\pi}_j}\right) \tag{A.12}$$

$$\mathbb{CV}\left(Y_{ij1}, Y_{ij2}|s_i, \tilde{s}_i\right) = s_i\tilde{s}_i\left(1-\pi_j\right)\left(1-\tilde{\pi}_j\right)e^{\mu_{j1}+\mu_{j2}+\frac{\sigma_{j1}^2+\sigma_{j2}^2}{2}}\left[e^{\rho_j\sigma_{j1}\sigma_{j2}}-1\right] \tag{A.13}$$

From conditional to total moments.The conditional moments estimated above, paint a partial picture of the phenomenon we have observed in the synthetic and real experiments. The reason is that for each feature cell specific normalization constants are an extra source of variation which conditioning does not account for. To account for this source of variation we use laws of total mean, variance and covariance.

$$\mathbb{E}\left(Y_{ij1}\right) = \mathbb{E}(s_i)\left(1-\pi_j\right)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}} \tag{A.14}$$

$$\mathbb{E}\left(Y_{ij2}\right) = \Phi\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right) \tag{A.15}$$

$$\mathbb{V}\left(Y_{ij1}\right) = \left(1-\pi_j\right)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}}\left[\mathbb{E}(s_i)+\mathbb{E}(s_i^2)(1-\pi_j)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}}\frac{e^{\sigma_{j1}^2}-1+\pi_j}{1-\pi_j}\right] \tag{A.16}$$

$$\begin{aligned}\mathbb{V}\left(\frac{Y_{ij2}}{n_{ij}}\right) &= \Phi\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)\left[1-\Phi\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)\right] \\ &-2\left(1-\mathbb{E}\left(\frac{1}{n_{ij}}\right)\right)\mathbb{T}\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}},\frac{1}{\sqrt{1+2\sigma_{j2}^2}}\right)\end{aligned} \tag{A.17}$$

$$\begin{aligned}\mathbb{CV}\left(Y_{ij1},\frac{Y_{ij2}}{n_{ij}}\right) &= \mathbb{E}\left(s_i\right)\left(1-\pi_j\right)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}} \\ &\left[\Phi\left(\frac{\mu_{j2}+\rho_j\sigma_{j1}\sigma_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)-\Phi\left(\frac{\mu_{j2}}{\sqrt{1+\sigma_{j2}^2}}\right)\right]\end{aligned} \tag{A.18}$$

Equations A.16,A.17 and A.18 suggest that as expression inflation $\pi_j \to 1$ then Pearson correlation $\to 0$. Furthermore as n_{ij} increases, then $\mathbb{V}\left(\frac{Y_{ij2}}{n_{ij}}\right)$ drops slightly improving correlation. For 10X we proceed with the assumption that normalization constants for each modality come from a joint distribution yielding the following results.

$$\mathbb{E}\left(Y_{ij1}\right) = \mathbb{E}(s_i)\left(1-\pi_j\right)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}} \tag{A.19}$$

$$\mathbb{E}\left(Y_{ij2}\right) = \mathbb{E}(\tilde{s}_i)\left(1-\tilde{\pi}_j\right)e^{\mu_{j2}+\frac{\sigma_{j2}^2}{2}} \tag{A.20}$$

$$\mathbb{V}\left(Y_{ij1}\right) = \left(1-\pi_j\right)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}}\left[\mathbb{E}(s_i)+\mathbb{E}(s_i^2)(1-\pi_j)e^{\mu_{j1}+\frac{\sigma_{j1}^2}{2}}\frac{e^{\sigma_{j1}^2}-1+\pi_j}{1-\pi_j}\right] \tag{A.21}$$

$$\mathbb{V}\left(Y_{ij2}\right) = \left(1-\tilde{\pi}_j\right)e^{\mu_{j2}+\frac{\sigma_{j2}^2}{2}}\left[\mathbb{E}(\tilde{s}_i)+\mathbb{E}(\tilde{s}_i^2)(1-\tilde{\pi}_j)e^{\mu_{j2}+\frac{\sigma_{j2}^2}{2}}\frac{e^{\sigma_{j2}^2}-1+\tilde{\pi}_j}{1-\tilde{\pi}_j}\right] \tag{A.22}$$

$$\mathbb{CV}\left(Y_{ij1},Y_{ij2}\right) = \left(1-\pi_j\right)\left(1-\tilde{\pi}_j\right)e^{\mu_{j1}+\mu_{j2}+\frac{\sigma_{j1}^2+\sigma_{j2}^2}{2}}\left[\mathbb{E}\left(s_i\tilde{s}_i\right)\left(e^{\rho_j\sigma_{j1}\sigma_{j2}}-1\right)+\mathbb{CV}(\tilde{s}_i,s_i)\right] \tag{A.23}$$

From Equations A.21, A.22 and A.23 we can see that as accessibility or expression inflation $\pi_j, \tilde{\pi}_j \to 1$ then Pearson correlation $\to 0$.

A.6 Efficiency Analysis

We performed efficiency analysis to understand model's scalability. In preliminary implementations we experimented with both Markov Chain Monte Carlo (MCMC) and Variational Inference (VI). Unfortunately VI did not yield accurate results so we abandoned it in favor of MCMC. Hence execution time as a function of scaling is summarized in the figure below.

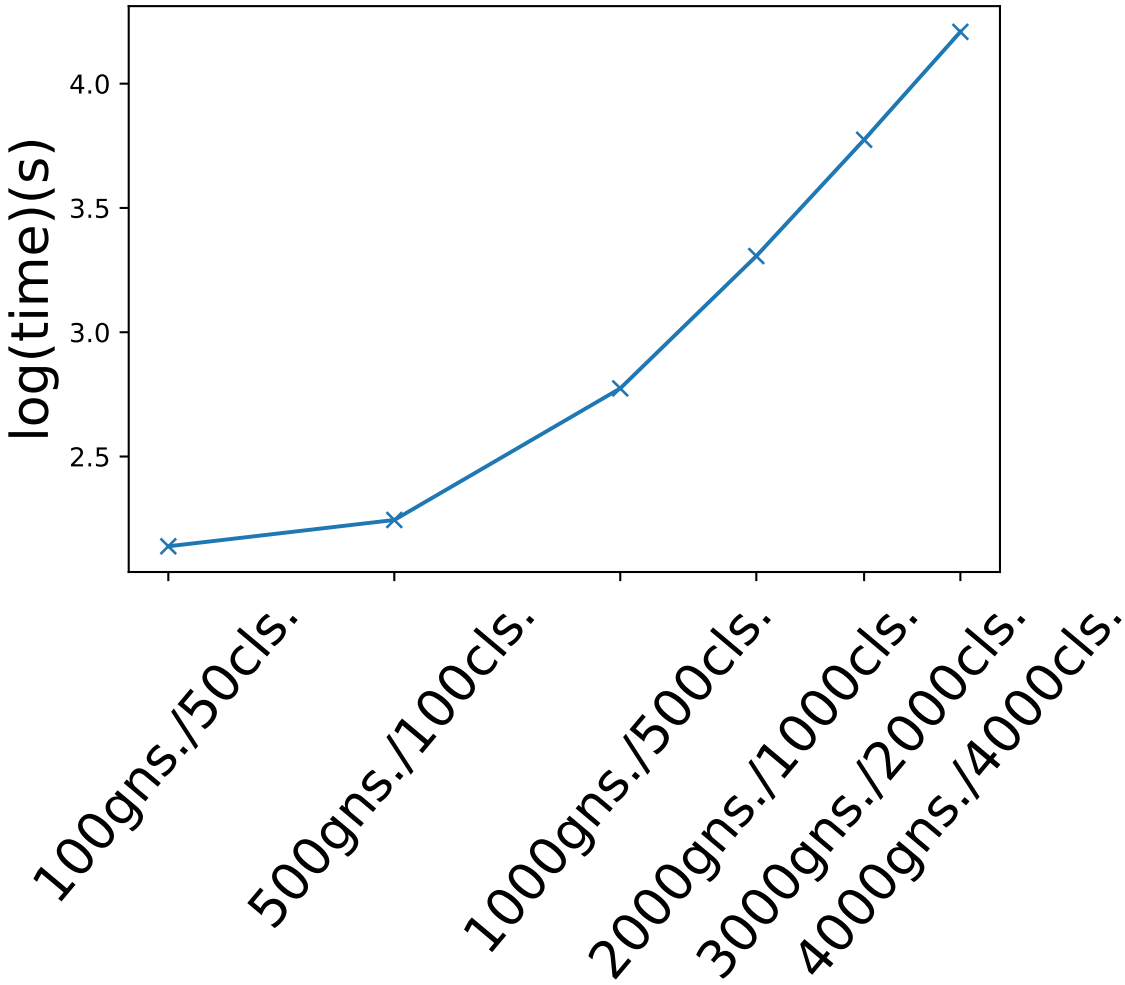


Fig. A.19 **Execution times as a function (features/cell numbers). Graph used to asses how SCRaPL scales with dimension. For large datasets it scales almost linearly.**

A.7 Macroscopic analysis of SCRaPL's behavior compared to Pearson and Spearman Correlation.

Here we provide more evidence in favor of the arguments presented in Section 5.4.2, about the differences between SCRaPL and different frequentist approaches considered in this work. More precisely, we argued that high CpG coverage and low inflation contribute in the reduction of discrepancy between SCRaPL's posterior correlation and Pearson/Spearman correlations. By regressing correlation discrepancy against these parameters we generate Figure A.20 along with Figure 5.9d presented in the main text which perfectly demonstrate the point.

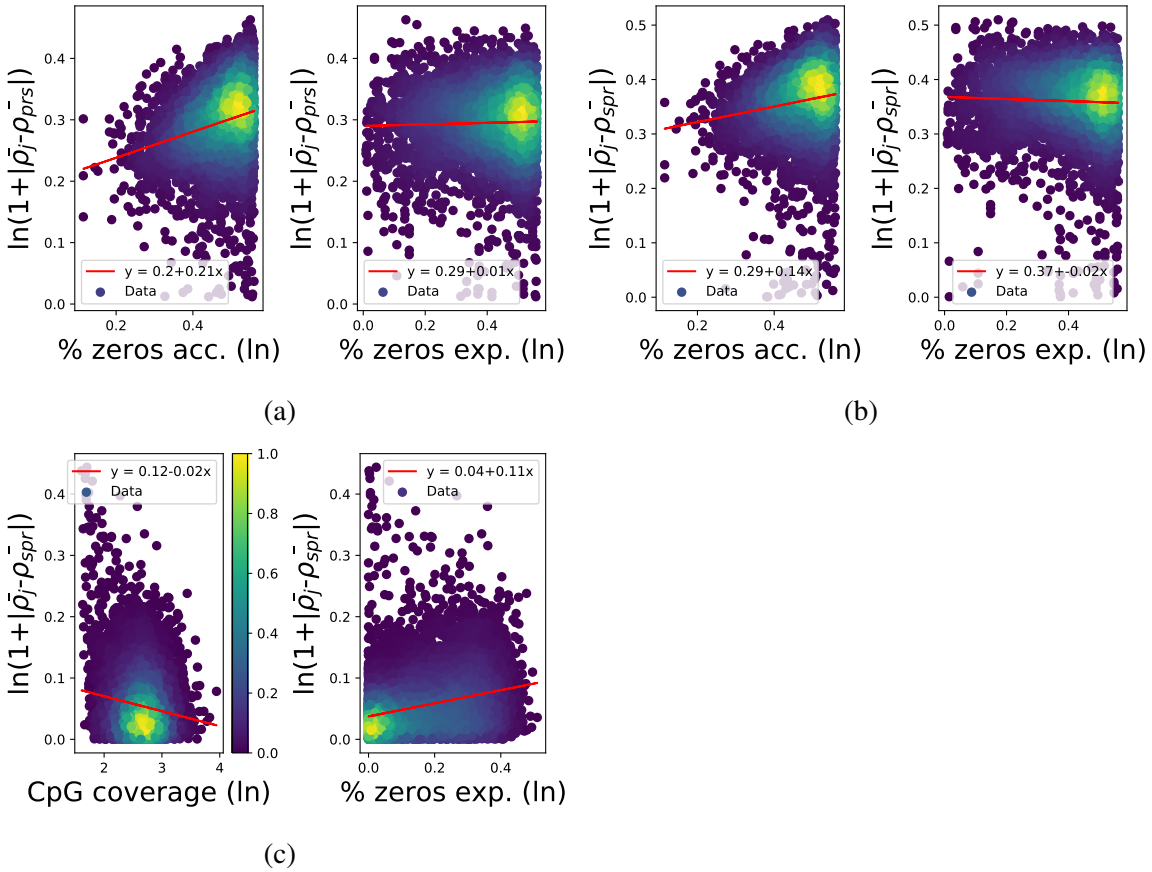


Fig. A.20 **SCRaPL's behavior compared to Pearson/Spearman correlation in macro scale across. Scatter plots used to asses how Spearman/Pearson predictions vary from SCRaPL's with zero inflation and CpG coverage.**. (A.20a) Scatterplots demonstrating the positive relationship between alternative correlation estimates (Pearson,Spearman and SCRaPL) and % zeros in accessibility and expression across features. (A.20b) Scatterplots demonstrating the positive relationship between alternative correlation estimates (Spearman and SCRaPL) and % zeros in accessibility and expression across features. (A.20c) Scatter plots demonstrating the negative/positive relationship between alternative correlation estimates and CpG coverage/% zeros in expression across features.

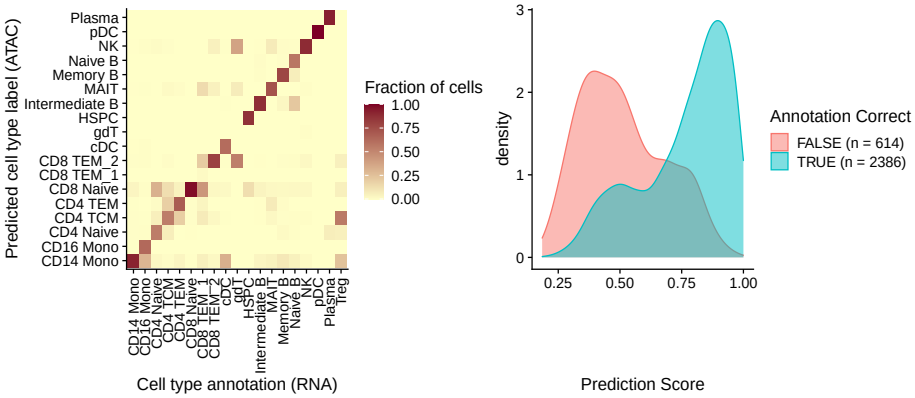
A.8 Using SCRaPL as a data denoising tool

The detection of associations between layers is only one of the many possible analyses which can be performed on multi-omics data sets. A substantial line of research has recently emerged around the topic of data integration, which aims to combine data from multiple layers measured in different cells obtained from the same biological system. The goal of such analyses is to enhance our understanding of cellular identity and function (Stuart et al., 2019). Popular platforms like Seurat (Hao et al., 2021) implement data integration via

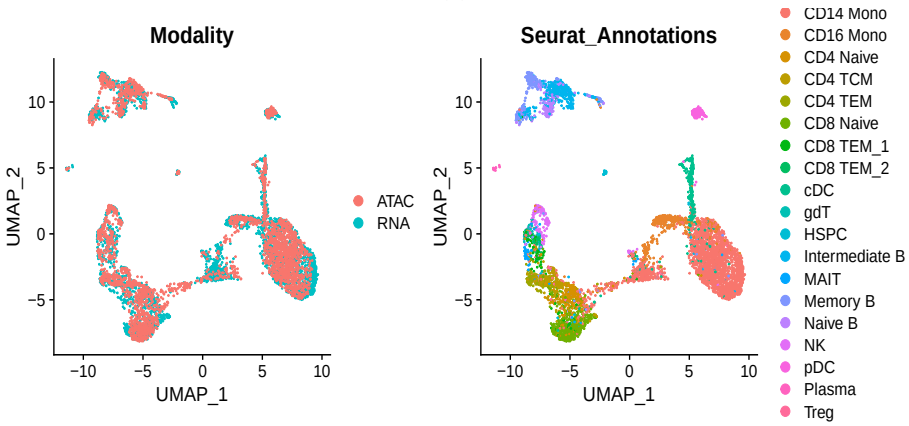
a dimensionality reduction approach based on Canonical Correlation Analysis (CCA), a technique based on Singular Value Decomposition of empirical correlation matrices (Loan, 1976). Despite its proven capabilities, CCA is not designed to handle count data. We therefore wondered whether SCRaPL's likelihoods tailored on specific data formats could under certain cases provide a valuable addition to the integration pipeline. Specifically, here we use SCRaPL as a denoising tool, and perform data integration at the level of the latent variables, rather than the raw data. In this subsection we follow the vignettes provided by Seurat's authors (Hoffman and Collaborators, 2021) and compare the results with and without SCRaPL's denoising. We note that this analysis is only a proof of concept as SCRaPL uses multi-omics data collected in the same cells as input and therefore such integration is not required.

For comparison between SCRaPL denoised and raw data we looked at peripheral blood mononuclear cells (PBMC) data (10X, 2020). This dataset contains expression and accessibility for 12000 PBMCs gathered from a healthy 25 year old donor. For illustration purposes, we downsampled the dataset to 3000 cells by keeping only the ones with the highest sum across peaks. Then peaks and genes were reduced from 180000 to 30000 and from 36000 to 10000 respectively, based on their variability. Then the to $60k$ features in association magnitude were used by SCRaPL.

To perform data integration, we remove cell specific noise by sampling latent space accessibility/expression from the respective posterior distributions obtained from SCRaPL. In cases of peaks mapped to multiple genes, readings were averaged. These data were integrated by Seurat (Hao et al., 2021), ignoring TF-IDF(Seurat's accessibility data preprocessing) and scRNA normalization (aimed at expression data) steps. Standard performance monitoring plots like label transfer between single cell expression and accessibility data as well as integration plots are presented in Figures A.21 and A.22. In general epigenomic and transcriptomic layers have integrated well for both raw and SCRaPL preprocessed data as suggested from Figure A.21b and A.22b. This picture remains consistent across multiple other trials summarized in Section A.8. The integration metrics found in Figure A.21a and A.22b show a comparable performance between SCRaPL preprocessing and raw data. This is in stark contrast with the results of the preceding sections, which demonstrated a consistent superiority of SCRaPL in detecting associations at the level of individual features. The reason for this is probably to be found in the dimensionality reduction performed by Seurat: canonical components found by CCA are obtained via an averaging process which already does an excellent job at filtering out noise, much in the way that robust principal components can often be extracted also from noisy data.

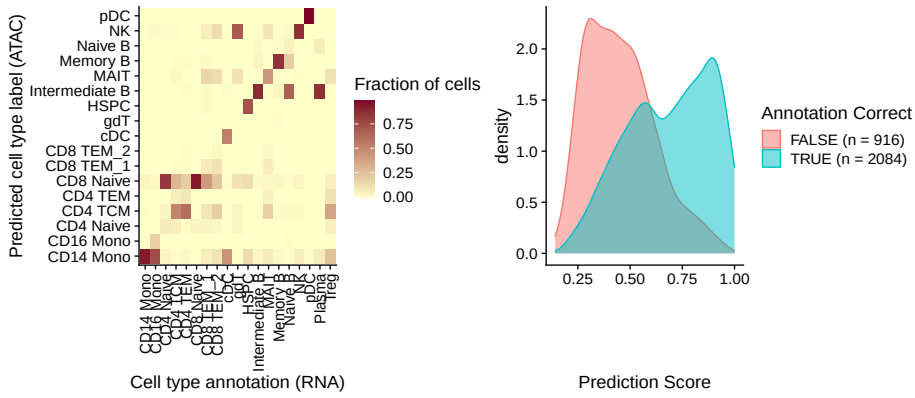


(a)

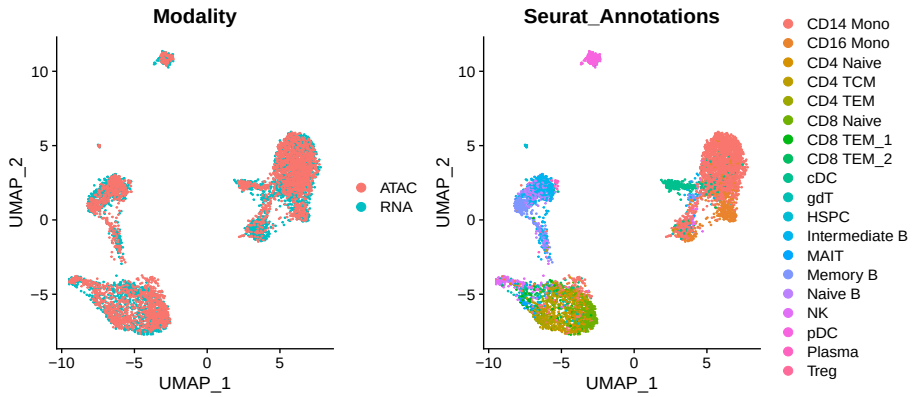


(b)

Fig. A.21 **Cell label transfer from expression to accessibility data for raw data A.21a. (A.21b) Visualization of sc-RNA and scATAC data on the same plot for raw data. Serves as baseline to compare raw data integration with Seurat.**

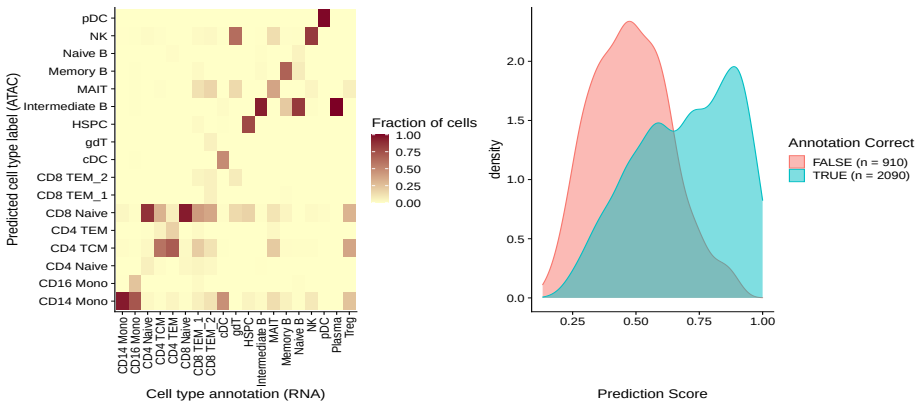


(a)

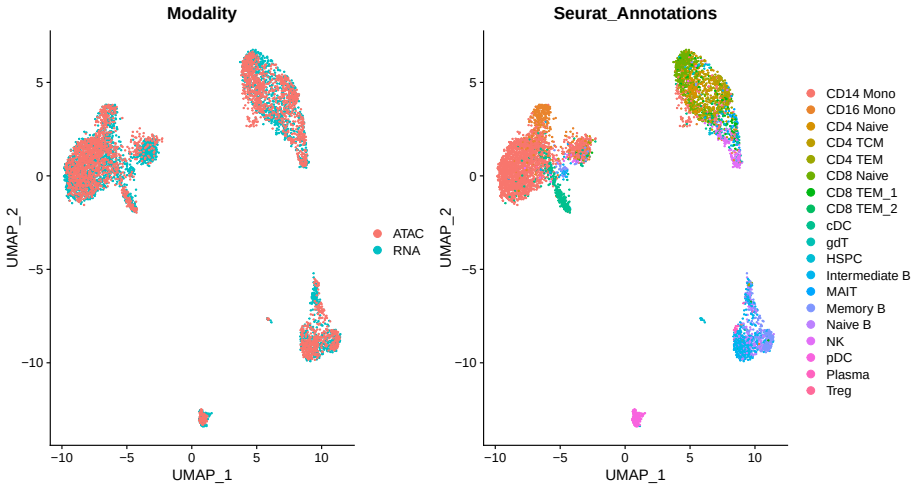


(b)

Fig. A.22 **Cell label transfer from expression to accessibility data for SCRaPL preprocessed data A.22a. (A.22b)Visualization of sc-RNA and scATAC data on the same plot for SCRaPL preprocessed data. Used to compare SCRaPL preprocessed data and raw data when integration is done with Seurat.**

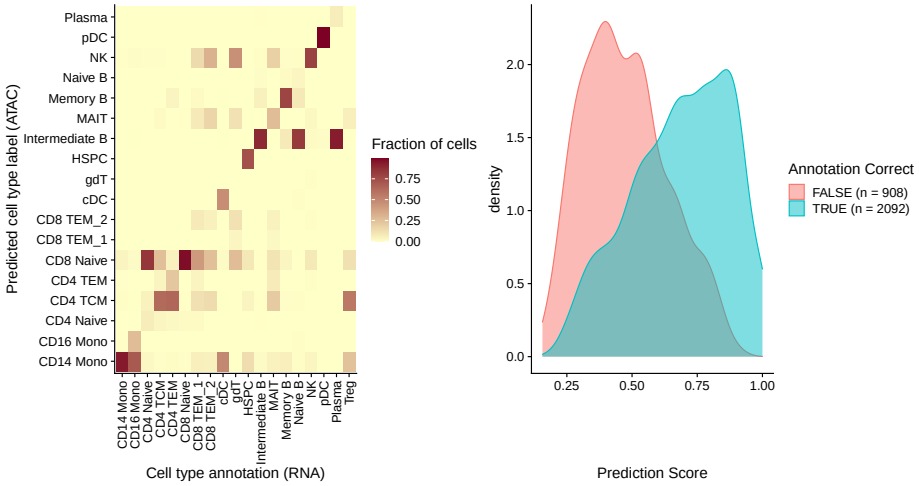


(a)

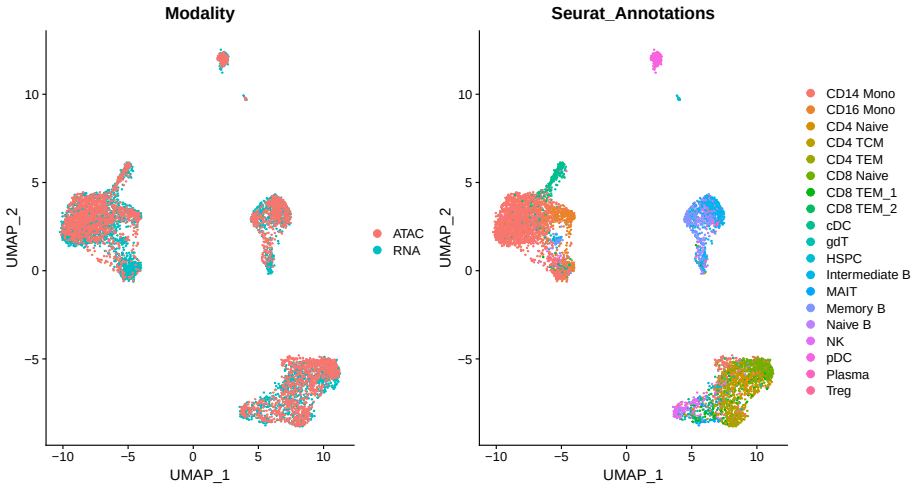


(b)

Fig. A.23 **Stability analysis of cell integration using SCRaPL as a denoising tool. Used to asses stability of Seurat's integration for SCRaPL preprocessed data.** Panel summarizing label transfer from expression to accessibility in trial 2(A.23a) along with respective latent plot visualization (A.23b) of sc-RNA and scATAC.

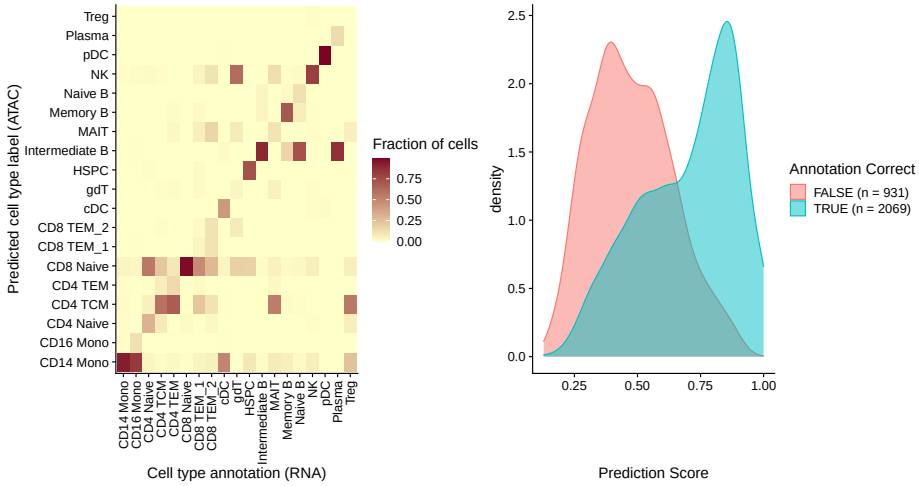


(a)

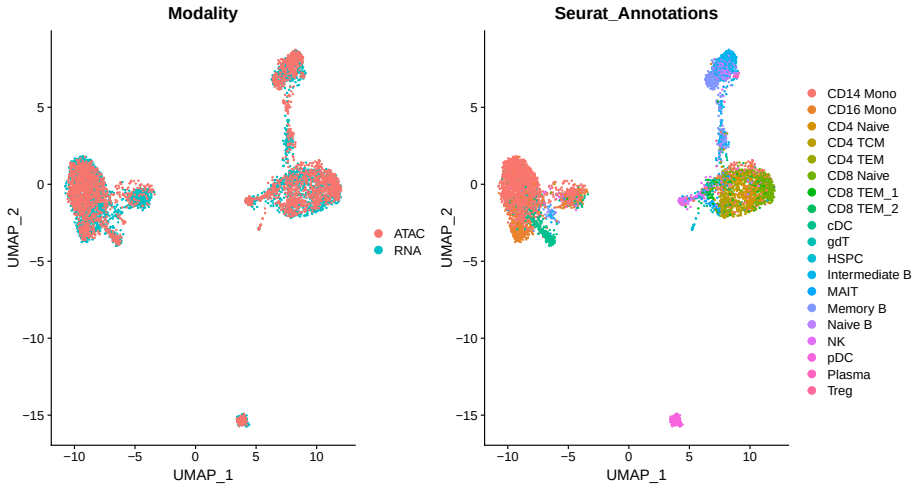


(b)

Fig. A.24 **Stability analysis of cell integration using SCRaPL as a denoising tool. Used to asses stability of Seurat's integration for SCRaPL preprocessed data.** Panel summarizing label transfer from expression to accessibility in trial 3 (A.24a) along with respective latent plot visualization (A.24b) of sc-RNA and scATAC.



(a)



(b)

Fig. A.25 **Stability analysis of cell integration using SCRaPL as a denoising tool. Used to asses stability of Seurat's integration for SCRaPL preprocessed data.** Panel summarizing label transfer from expression to accessibility in trial 4 (A.25a) along with respective latent plot visualization (A.25a) of sc-RNA and scATAC.

A.9 Convergence Analysis

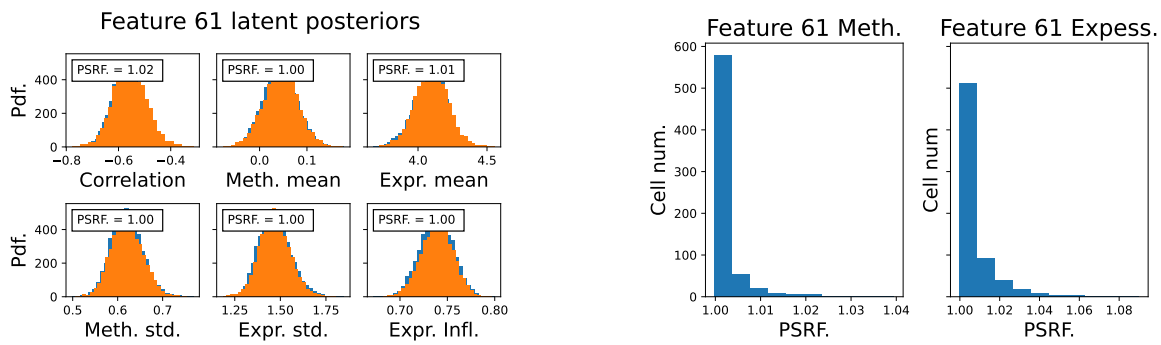
Convergence analysis in MCMC is an important tool for assessing whether the Markov chain has reached its equilibrium distribution and whether the samples are representative of the target distribution. If the Markov chain has not reached its equilibrium distribution, the samples generated by the chain may be biased. To asses convergence we rely on Gelman-Rubin criterion (Gelman and Rubin, 1992) also known as potential scale reduction factor

(PSRF). To demonstrate that MCMC has converged two MCMC chains have been inferred per parameter along with the corresponding PSRF. This was done for mESC and mEBC data for models with and without inflation.

In the remaining of Section A.9 we present chain convergence evidence. Convergence of latent epigenomic and transcriptomic states is demonstrated by poltting the distribution of PSRF values. For the remaining latent parameters like means and covariance convergence is assessed using PSRF and visually inspecting target distribution for the two MCMC chains.

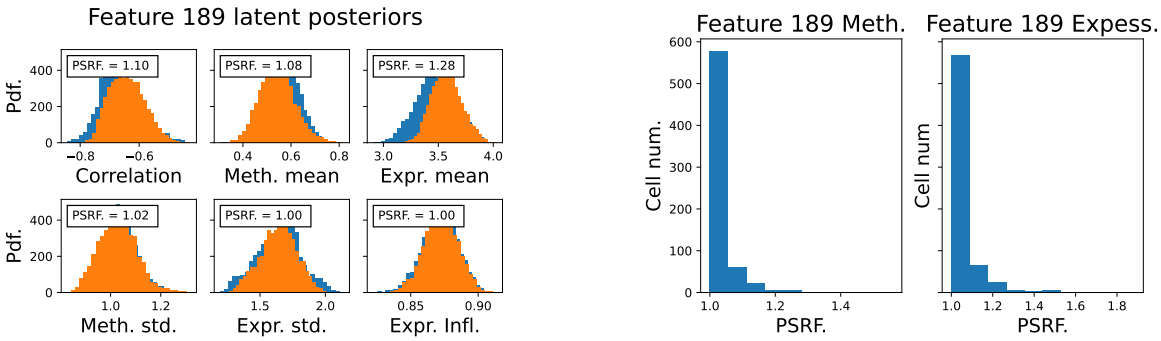
A.9.1 SCRaPL with Inflation on mESC

In this section we present evidence for parameter convergence when SCRaPL with inflation was applied on mESC data.



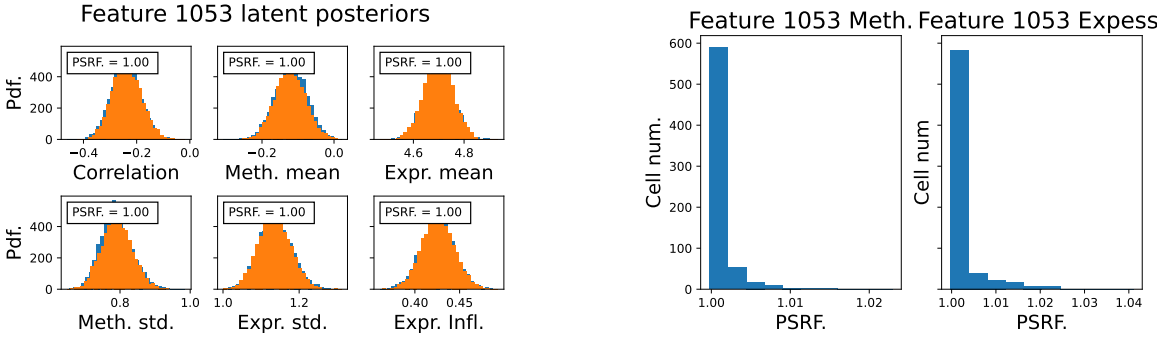
(a) (b)

Fig. A.26 **Convergence analysis for feature 61.** Convergence analysis of SCRaPL latent parameters for feature 61 when SCRaPL is applied with inflation on mESC data. (A.26a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.26b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



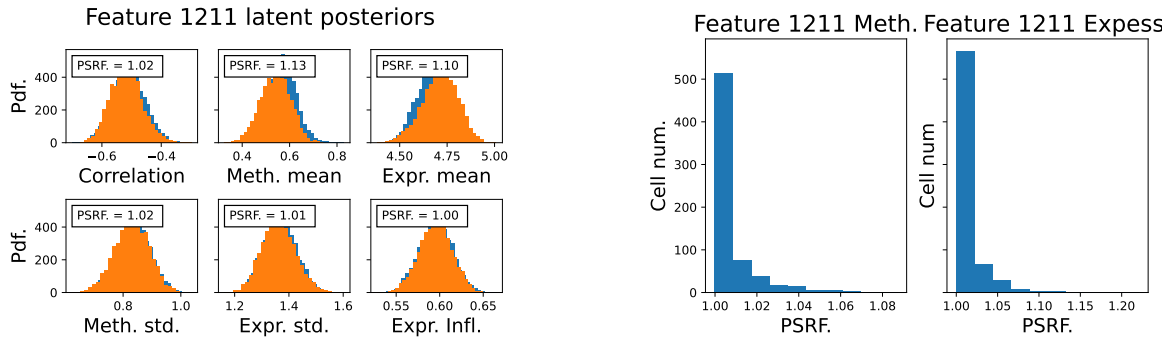
(a) (b)

Fig. A.27 **Convergence analysis for feature 189.** Convergence analysis of SCRaPL latent parameters for feature 189 when SCRaPL is applied with inflation on mESC data. (A.27a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.27b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



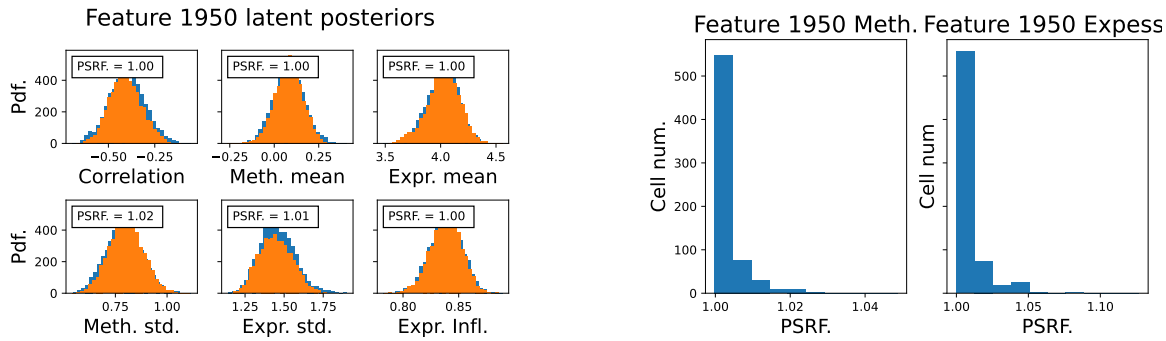
(a) (b)

Fig. A.28 **Convergence analysis for feature 1053.** Convergence analysis of SCRaPL latent parameters for feature 1053 when SCRaPL is applied with inflation on mESC data. (A.28a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.28b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



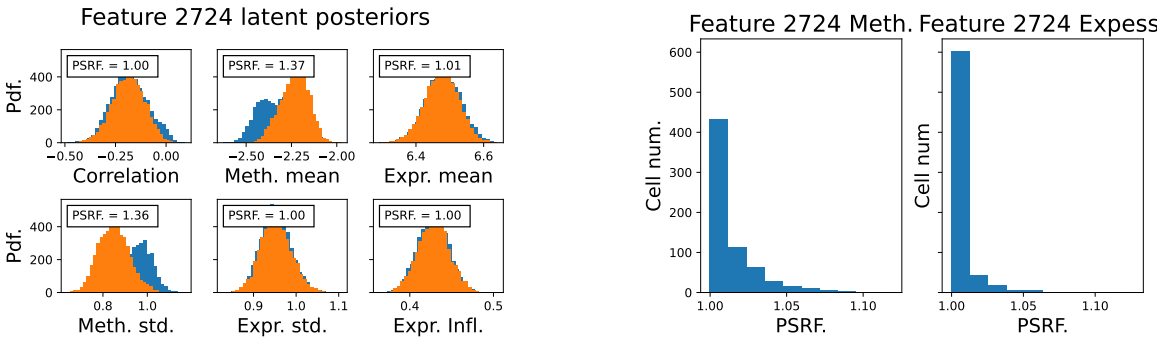
(a) (b)

Fig. A.29 **Convergence analysis for feature 1211.** Convergence analysis of SCRaPL latent parameters for feature 1211 when SCRaPL is applied with inflation on mESC data. (A.29a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.29b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



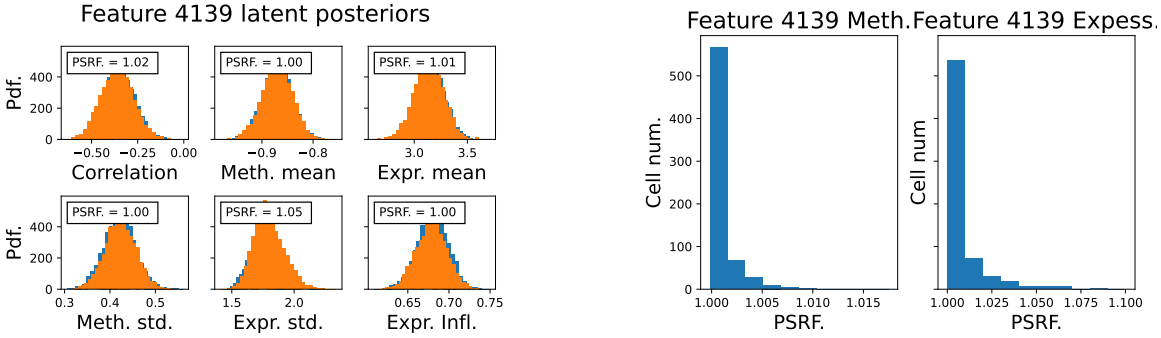
(a) (b)

Fig. A.30 **Convergence analysis for feature 1950.** Convergence analysis of SCRaPL latent parameters for feature 1950 when SCRaPL is applied with inflation on mESC data. (A.30a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.30b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



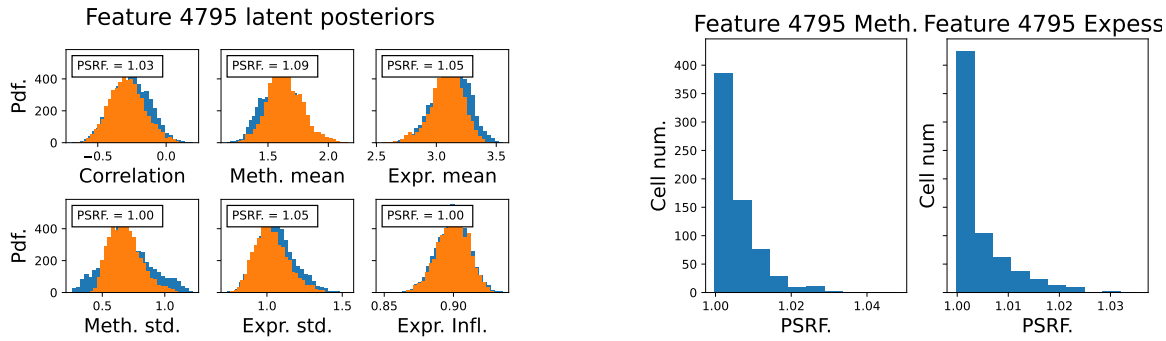
(a) (b)

Fig. A.31 **Convergence analysis for feature 2724.** Convergence analysis of SCRaPL latent parameters for feature 2724 when SCRaPL is applied with inflation on mESC data. (A.31a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.31b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



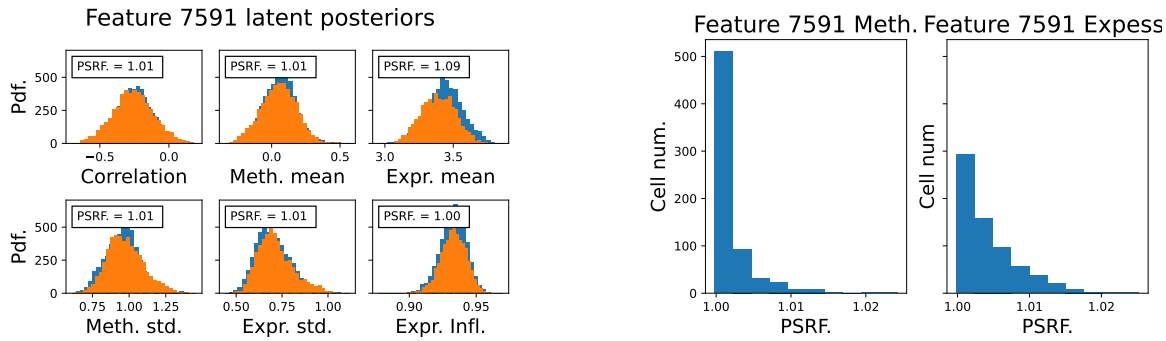
(a) (b)

Fig. A.32 **Convergence analysis for feature 4139.** Convergence analysis of SCRaPL latent parameters for feature 4139 when SCRaPL is applied with inflation on mESC data. (A.32a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.32b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



(a) (b)

Fig. A.33 **Convergence analysis for feature 4795.** Convergence analysis of SCRaPL latent parameters for feature 4795 when SCRaPL is applied with inflation on mESC data. (A.33a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.33b) PSRF. distributions for epigenetic and transcriptomic latent parameters.

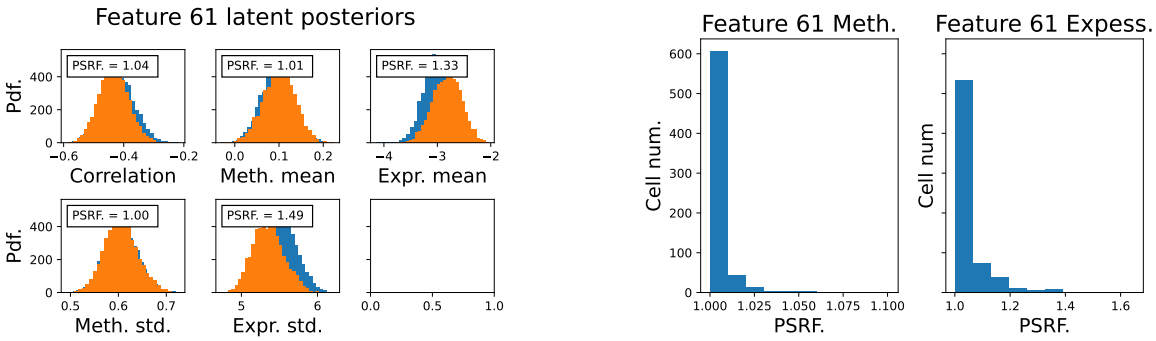


(a) (b)

Fig. A.34 **Convergence analysis for feature 7591.** Convergence analysis of SCRaPL latent parameters for feature 7591 when SCRaPL is applied with inflation on mESC data. (A.34a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.34b) PSRF. distributions for epigenetic and transcriptomic latent parameters.

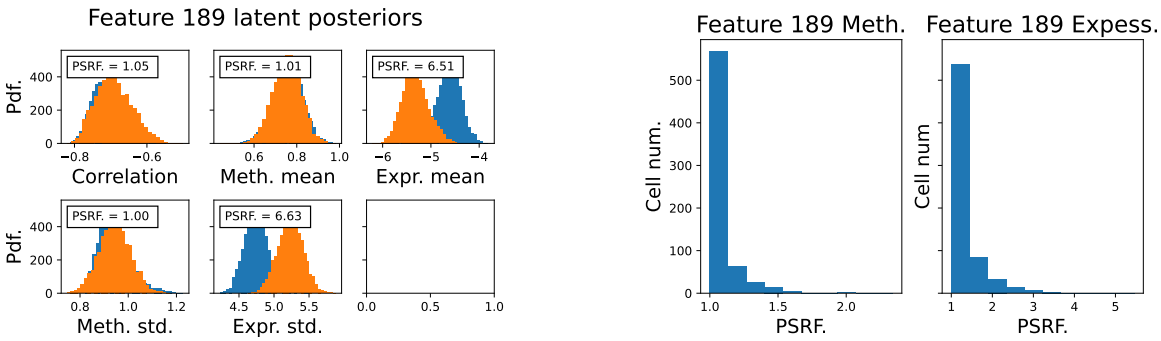
A.9.2 SCRaPL without Inflation on mESC

In this section we present evidence for parameter convergence when SCRaPL without inflation was applied on mESC data.



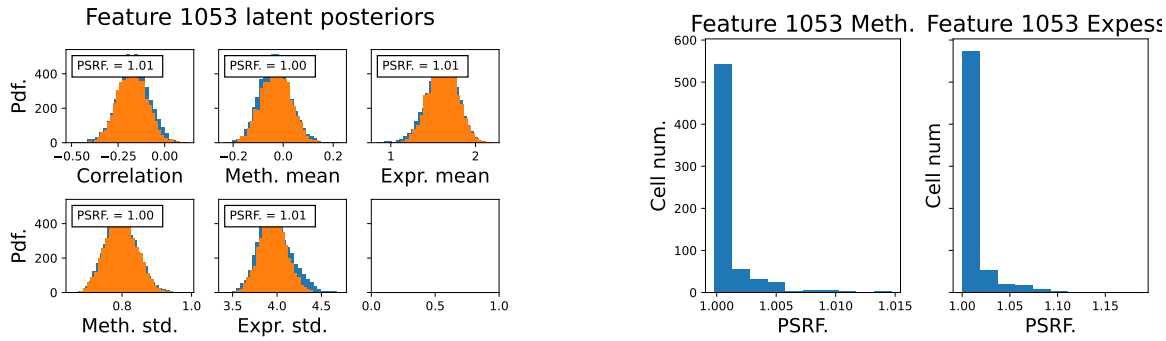
(a) (b)

Fig. A.35 **Convergence analysis for feature 61.** Convergence analysis of SCRaPL latent parameters for feature 61 when SCRaPL is applied without inflation on mESC data. (A.35a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.35b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



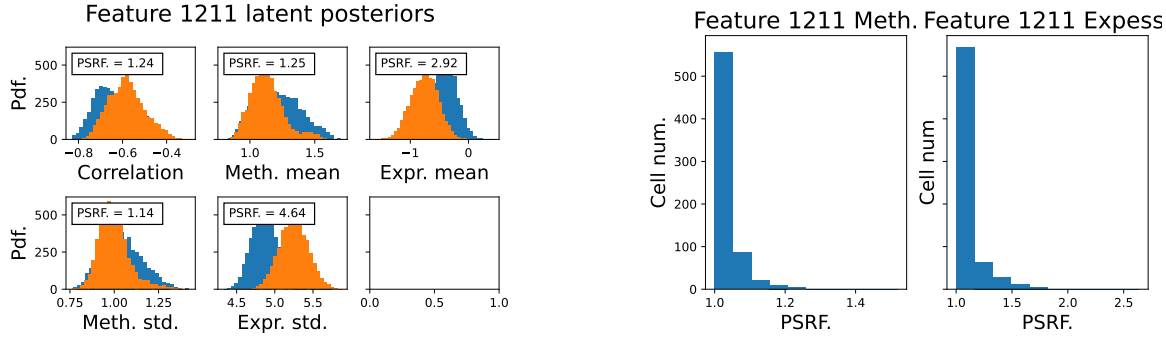
(a) (b)

Fig. A.36 **Convergence analysis for feature 189.** Convergence analysis of SCRaPL latent parameters for feature 189 when SCRaPL is applied without inflation on mESC data. (A.36a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.36b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



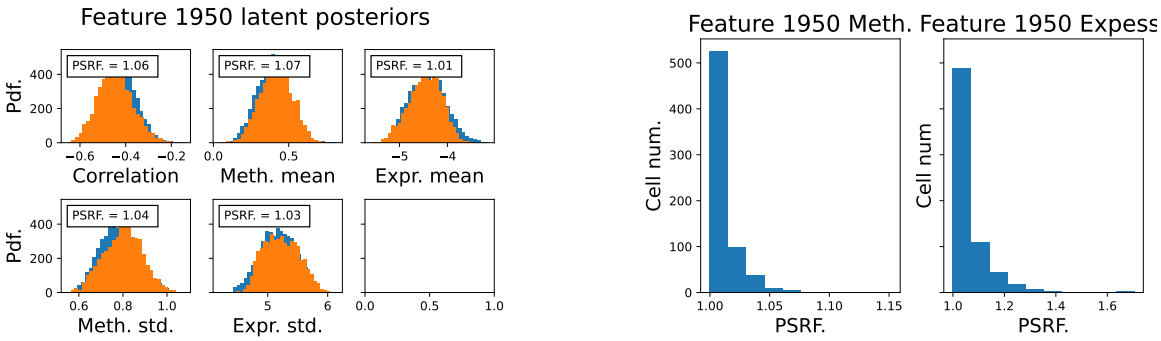
(a) (b)

Fig. A.37 **Convergence analysis for feature 1053.** Convergence analysis of SCRaPL latent parameters for feature 1053 when SCRaPL is applied without inflation on mESC data. (A.37a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.37b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



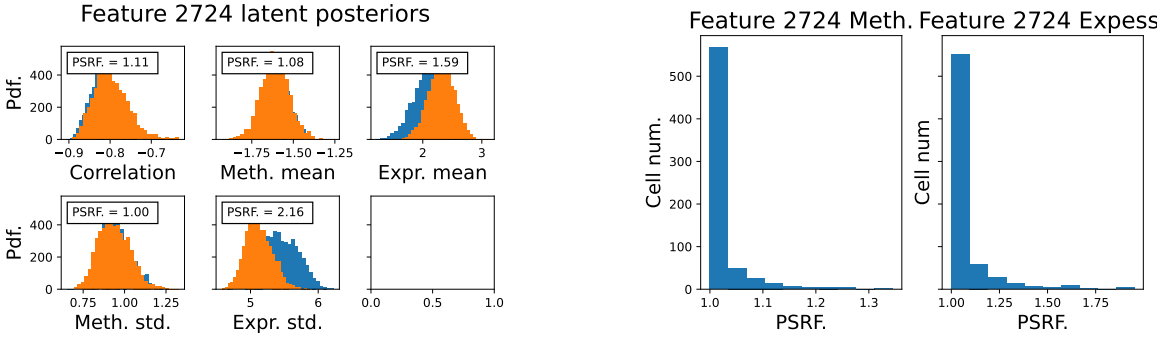
(a) (b)

Fig. A.38 **Convergence analysis for feature 1211.** Convergence analysis of SCRaPL latent parameters for feature 1211 when SCRaPL is applied without inflation on mESC data. (A.38a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.38b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



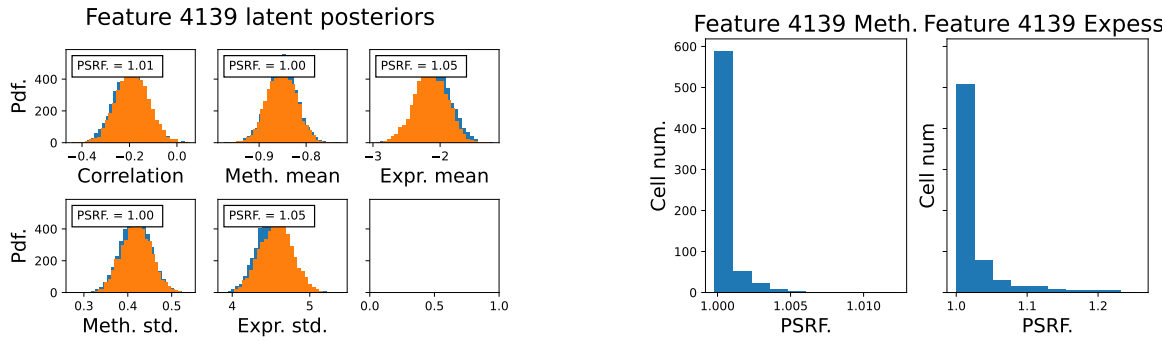
(a) (b)

Fig. A.39 **Convergence analysis for feature 1950** Convergence analysis of SCRaPL latent parameters for feature 1950 when SCRaPL is applied without inflation on mESC data. (A.39a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.39b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



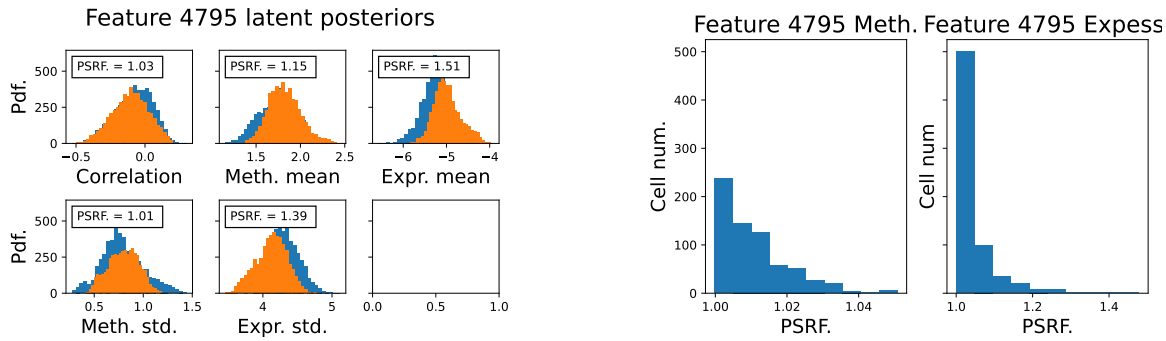
(a) (b)

Fig. A.40 **Convergence analysis for feature 2724.** Convergence analysis of SCRaPL latent parameters for feature 2724 when SCRaPL is applied without inflation on mESC data. (A.40a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.40b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



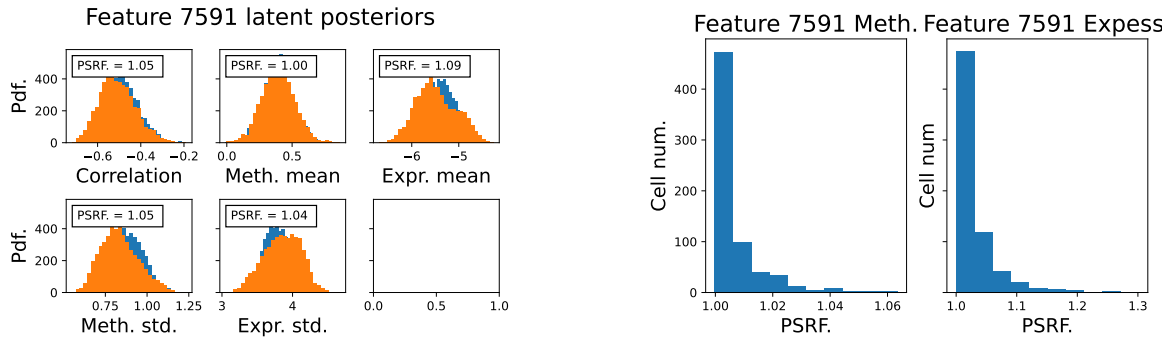
(a) (b)

Fig. A.41 **Convergence analysis for feature 4139.** Convergence analysis of SCRaPL latent parameters for feature 4139 when SCRaPL is applied without inflation on mESC data. (A.41a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.41b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



(a) (b)

Fig. A.42 **Convergence analysis for feature 4795.** Convergence analysis of SCRaPL latent parameters for feature 4795 when SCRaPL is applied without inflation on mESC data. (A.42a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.42b) PSRF. distributions for epigenetic and transcriptomic latent parameters.

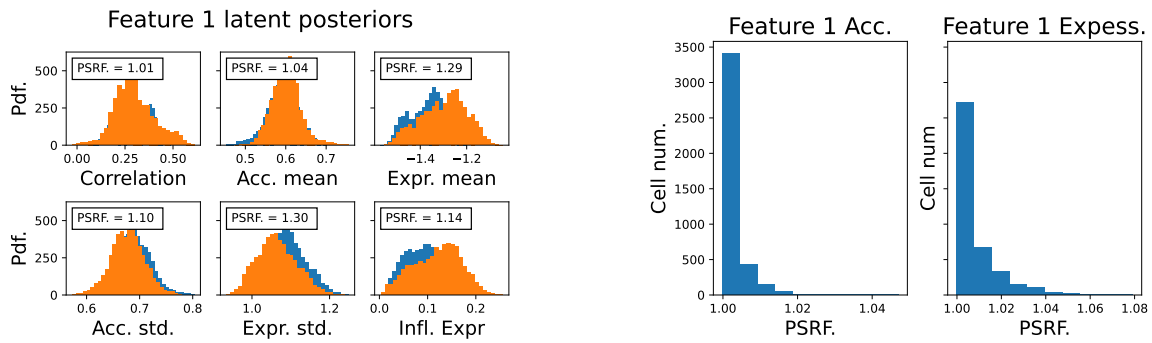


(a) (b)

Fig. A.43 **Convergence analysis for feature 7591.** Convergence analysis of SCRaPL latent parameters for feature 7591 when SCRaPL is applied without inflation on mESC data. (A.43a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.43b) PSRF. distributions for epigenetic and transcriptomic latent parameters.

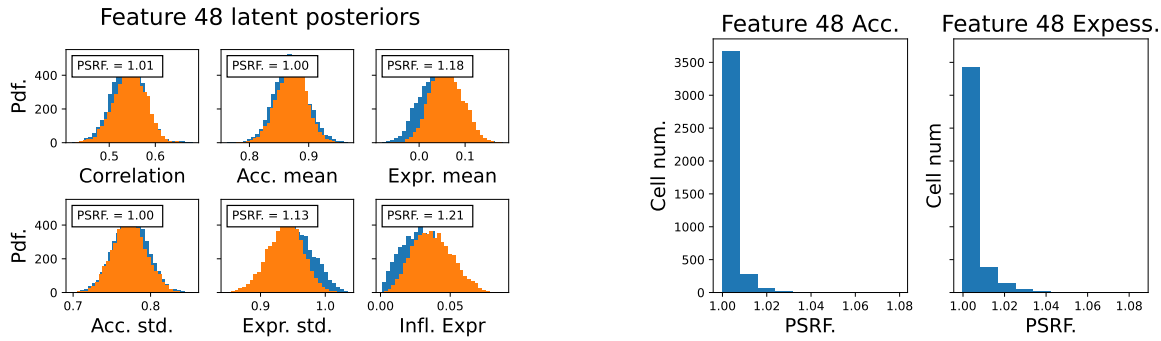
A.9.3 SCRaPL with Inflation on mEBC

In this section we present evidence for parameter convergence when SCRaPL with inflation was applied on mEBC data.



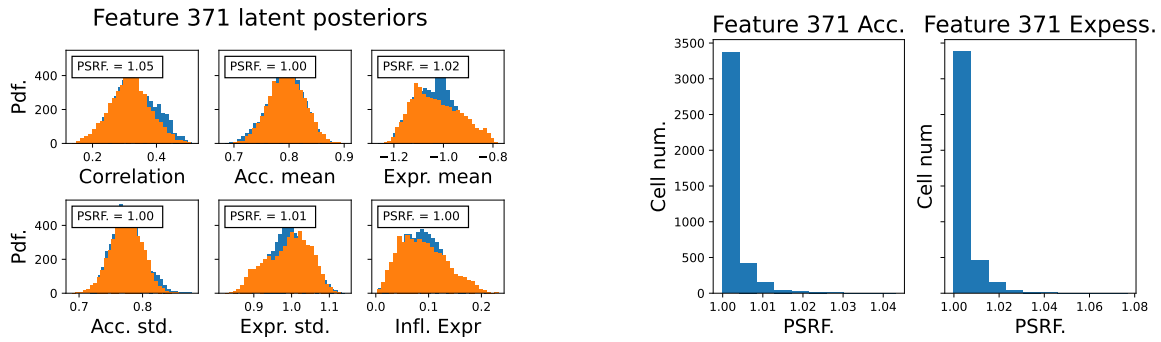
(a) (b)

Fig. A.44 **Convergence analysis for feature 1.** Convergence analysis of SCRaPL latent parameters for feature 1 when SCRaPL is applied with inflation on mEBC data. (A.44a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.44b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



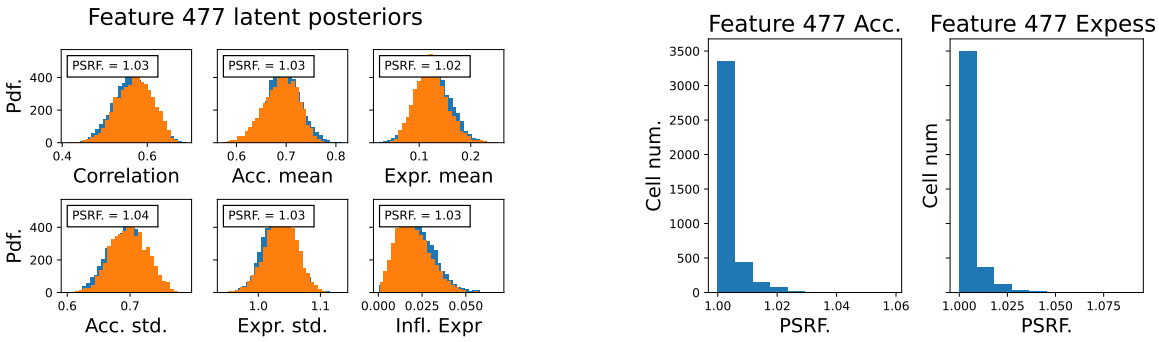
(a) (b)

Fig. A.45 **Convergence analysis for feature 48.** Convergence analysis of SCRaPL latent parameters for feature 48 when SCRaPL is applied with inflation on mEBC data. (A.45a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.45b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



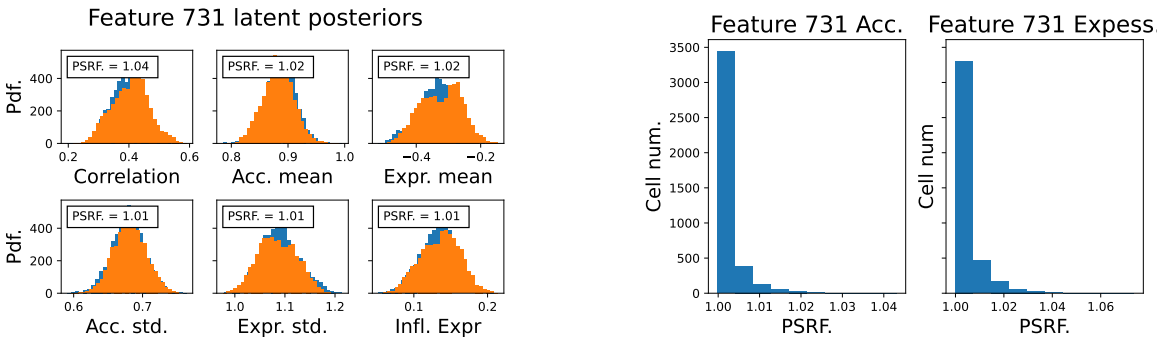
(a) (b)

Fig. A.46 **Convergence analysis for feature 371.** Convergence analysis of SCRaPL latent parameters for feature 371 when SCRaPL is applied with inflation on mEBC data. (A.46a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.46b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



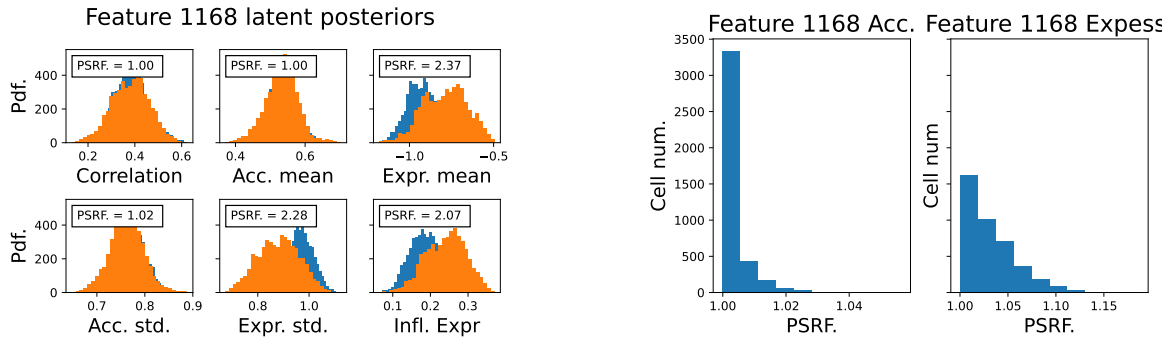
(a) (b)

Fig. A.47 **Convergence analysis for feature 477.** Convergence analysis of SCRaPL latent parameters for feature 477 when SCRaPL is applied with inflation on mEBC data. (A.47a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.47b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



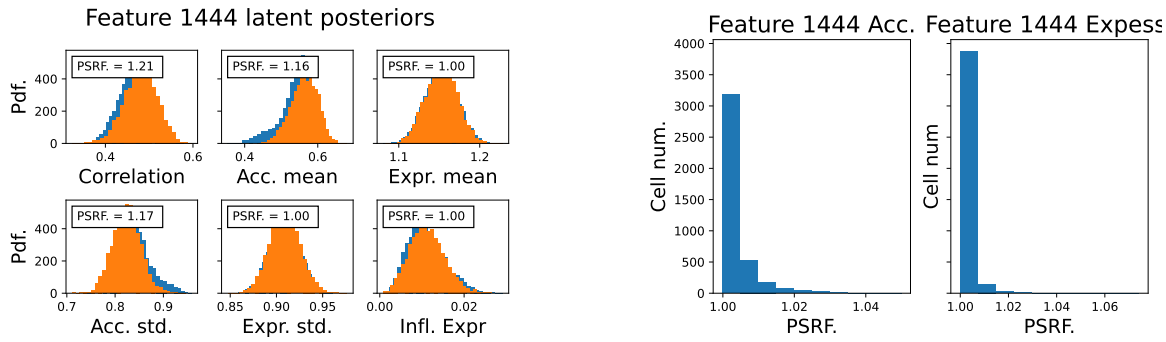
(a) (b)

Fig. A.48 **Convergence analysis for feature 731.** Convergence analysis of SCRaPL latent parameters for feature 731 when SCRaPL is applied with inflation on mEBC data. (A.48a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.48b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



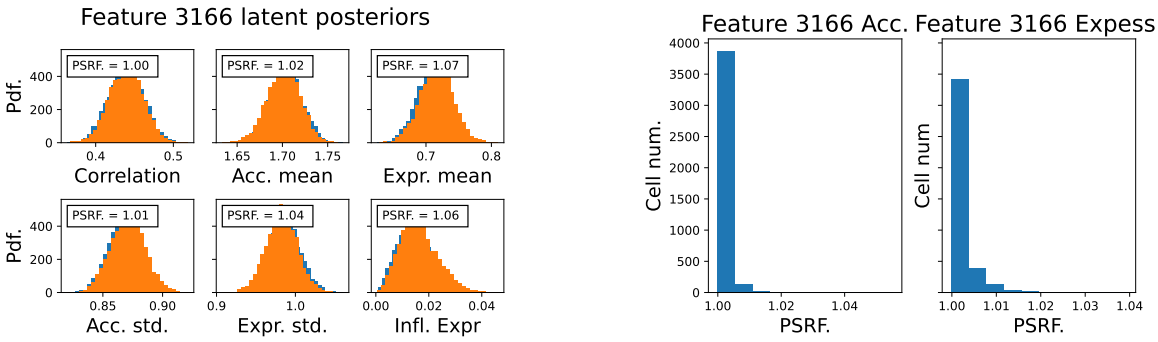
(a) (b)

Fig. A.49 **Convergence analysis for feature 1168.** Convergence analysis of SCRaPL latent parameters for feature 1168 when SCRaPL is applied with inflation on mEBC data. (A.49a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.49b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



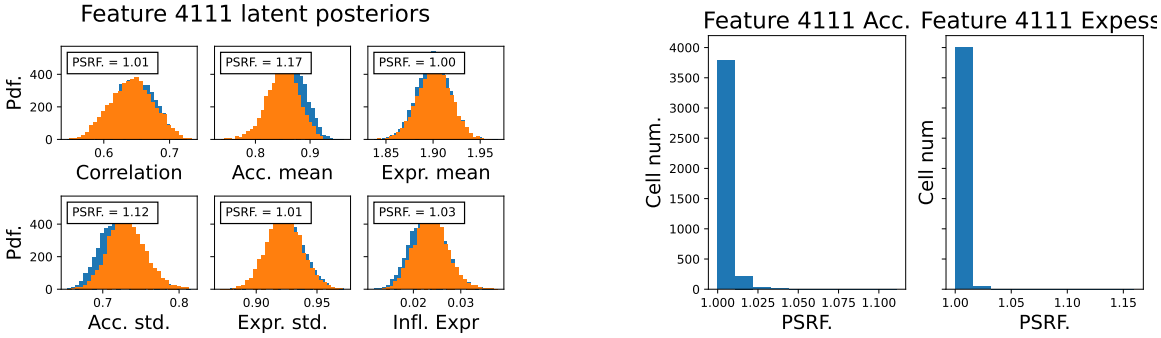
(a) (b)

Fig. A.50 **Convergence analysis for feature 1444.** Convergence analysis of SCRaPL latent parameters for feature 1444 when SCRaPL is applied with inflation on mEBC data. (A.50a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.50b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



(a) (b)

Fig. A.51 **Convergence analysis for feature 3166.** Convergence analysis of SCRaPL latent parameters for feature 3166 when SCRaPL is applied with inflation on mEBC data. (A.51a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.51b) PSRF. distributions for epigenetic and transcriptomic latent parameters.

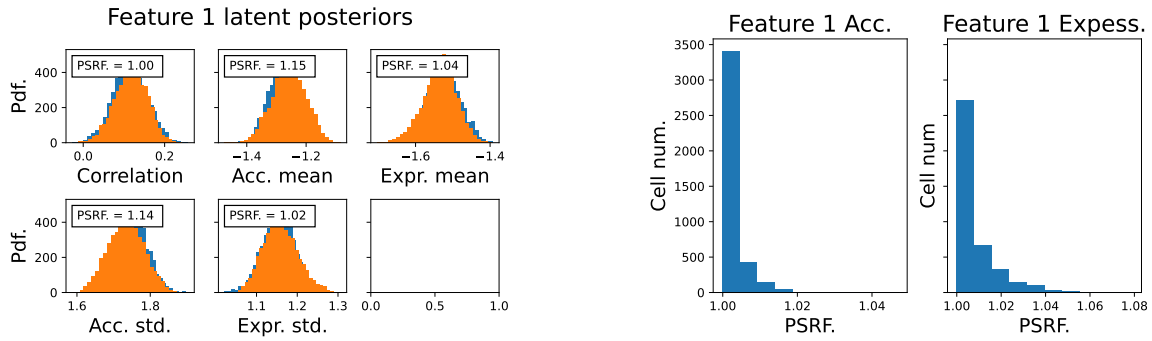


(a) (b)

Fig. A.52 **Convergence analysis for feature 4111.** Convergence analysis of SCRaPL latent parameters for feature 4111 when SCRaPL is applied with inflation on mEBC data. (A.52a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.52b) PSRF. distributions for epigenetic and transcriptomic latent parameters.

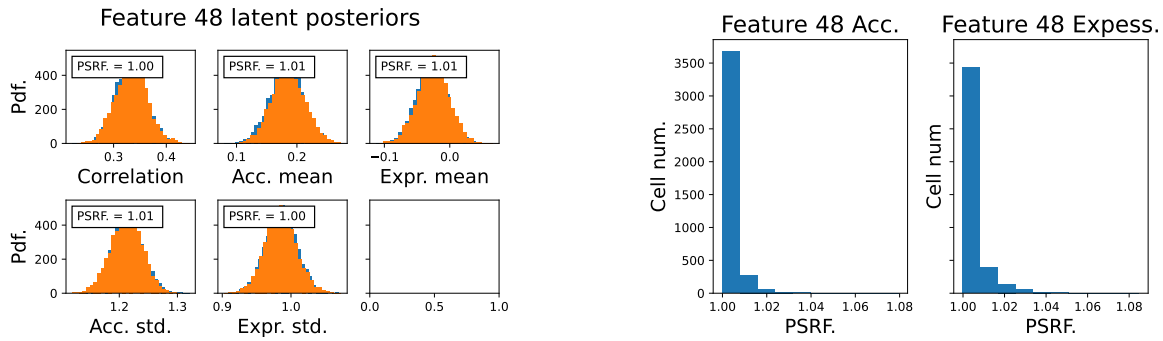
A.9.4 SCRaPL without Inflation on mEBC

In this section we present evidence for parameter convergence when SCRaPL without inflation was applied on mEBC data.



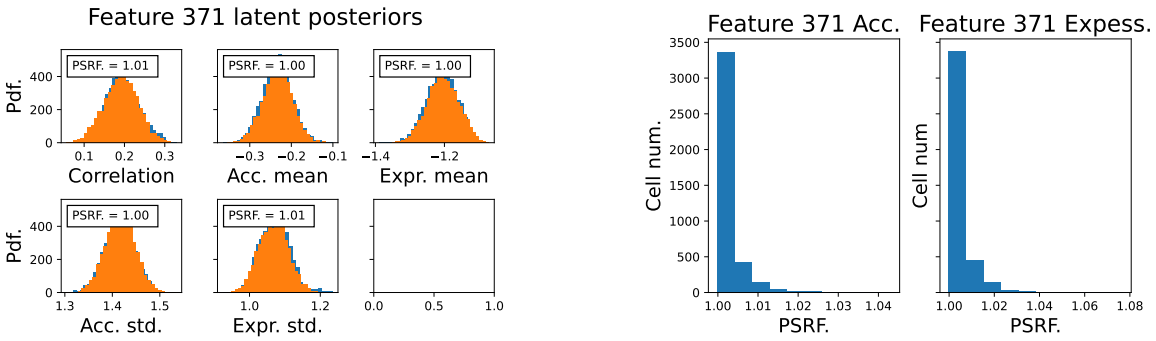
(a) (b)

Fig. A.53 **Convergence analysis for feature 1.** Convergence analysis of SCRaPL latent parameters for feature 1 when SCRaPL is applied without inflation on mEBC data. (A.53a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.53b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



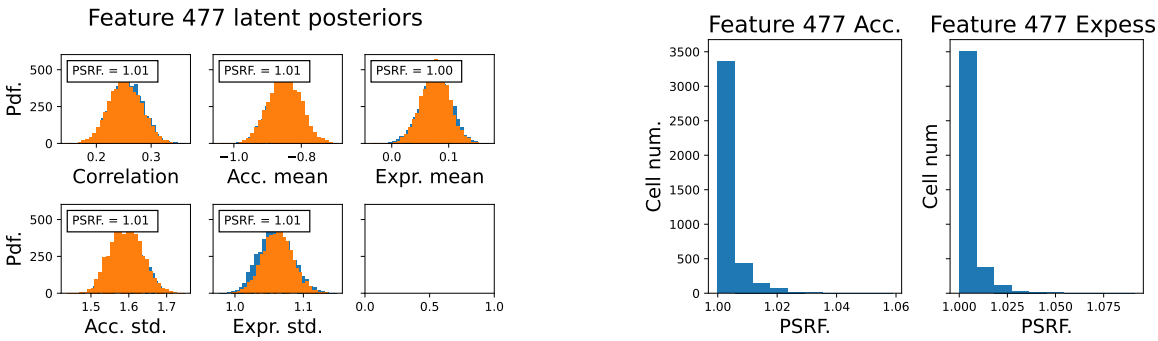
(a) (b)

Fig. A.54 **Convergence analysis for feature 48.** Convergence analysis of SCRaPL latent parameters for feature 48 when SCRaPL is applied without inflation on mEBC data. (A.54a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.54b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



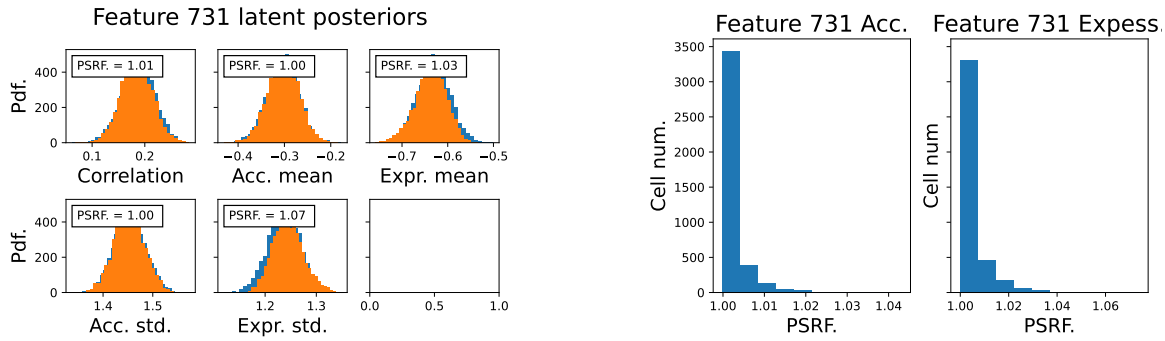
(a) (b)

Fig. A.55 **Convergence analysis for feature 371.** Convergence analysis of SCRaPL latent parameters for feature 371 when SCRaPL is applied without inflation on mEBC data. (A.55a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.55b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



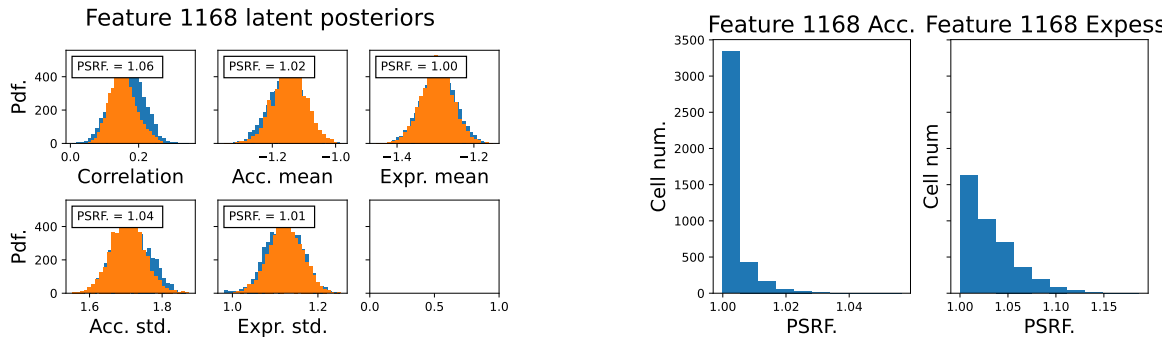
(a) (b)

Fig. A.56 **Convergence analysis for feature 477.** Convergence analysis of SCRaPL latent parameters for feature 477 when SCRaPL is applied without inflation on mEBC data. (A.56a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.56b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



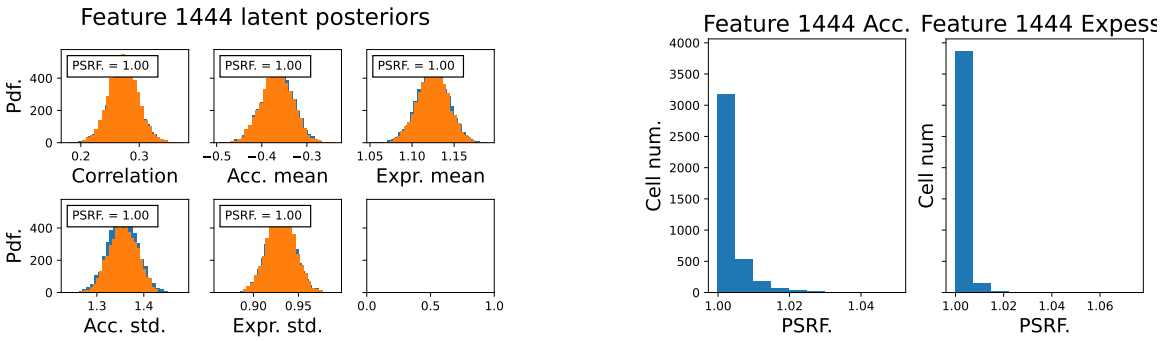
(a) (b)

Fig. A.57 **Convergence analysis for feature 731.** Convergence analysis of SCRaPL latent parameters for feature 731 when SCRaPL is applied without inflation on mEBC data. (A.57a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.57b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



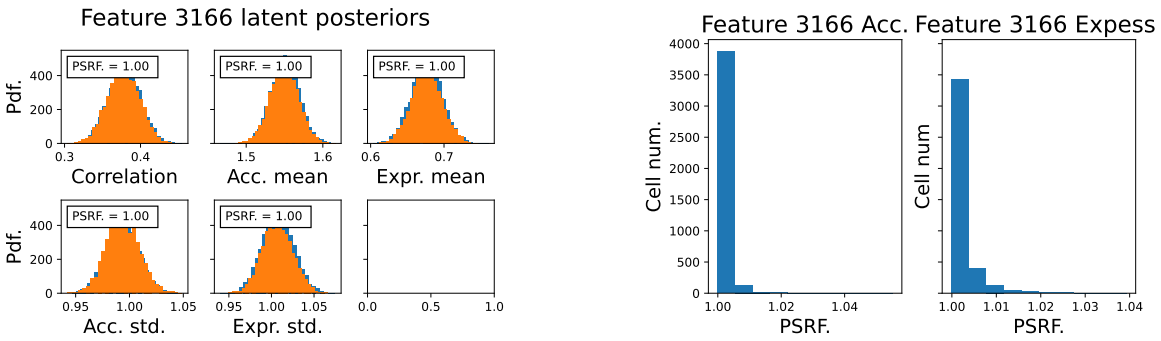
(a) (b)

Fig. A.58 **Convergence analysis for feature 1168.** Convergence analysis of SCRaPL latent parameters for feature 1168 when SCRaPL is applied without inflation on mEBC data. (A.58a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.58b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



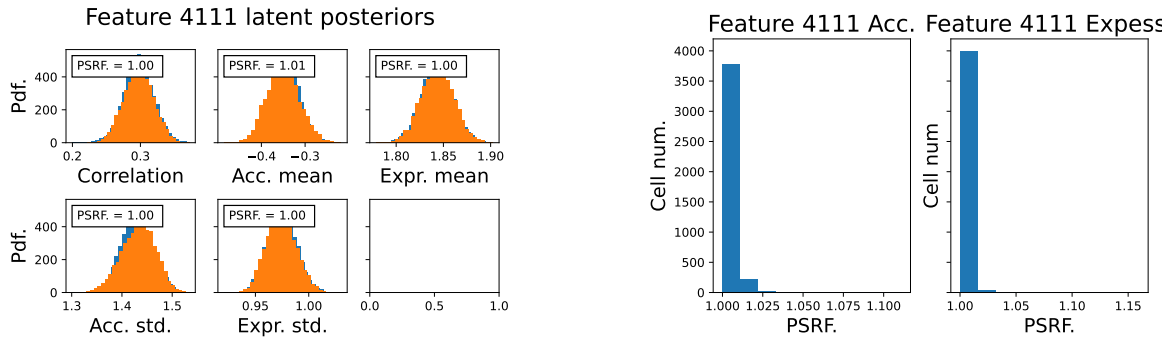
(a) (b)

Fig. A.59 **Convergence analysis for feature 1444.** Convergence analysis of SCRaPL latent parameters for feature 1444 when SCRaPL is applied without inflation on mEBC data. (A.59a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.59b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



(a) (b)

Fig. A.60 **Convergence analysis for feature 3166.** Convergence analysis of SCRaPL latent parameters for feature 3166 when SCRaPL is applied without inflation on mEBC data. (A.60a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.60b) PSRF. distributions for epigenetic and transcriptomic latent parameters.



(a) (b)

Fig. A.61 **Convergence analysis for feature 4111.** Convergence analysis of SCRaPL latent parameters for feature 4111 when SCRaPL is applied without inflation on mEBC data. (A.61a) Panel summarizing two MCMC chains for posteriors latent means and covariance. (A.61b) PSRF. distributions for epigenetic and transcriptomic latent parameters.

A.10 Posterior predictive checks

Posterior predictive checks are an approach in Bayesian statistics for assessing the fit of a statistical model to observed data. The goal of posterior predictive checks is to compare the ground truth and model generated data, in order to evaluate whether the model is consistent with the observed data and to identify potential deficiencies in the model.

In this section we simulate data from the fitted model with and without inflation for mESC and mEBC data. Then, for selected features we qualitatively compare real and simulated data for respective models and datasets by drawing transcriptome against epigenome scatter-plots. More precisely, to generate simulated scatter-plots we feed SCRaPL with posterior mean values for the fitted parameters and simulate mESC and mEBC raw data.

A.10.1 Posterior predictive checks for mESC

In this section we present posterior predictive checks for SCRaPL with and without inflation in mESC data.

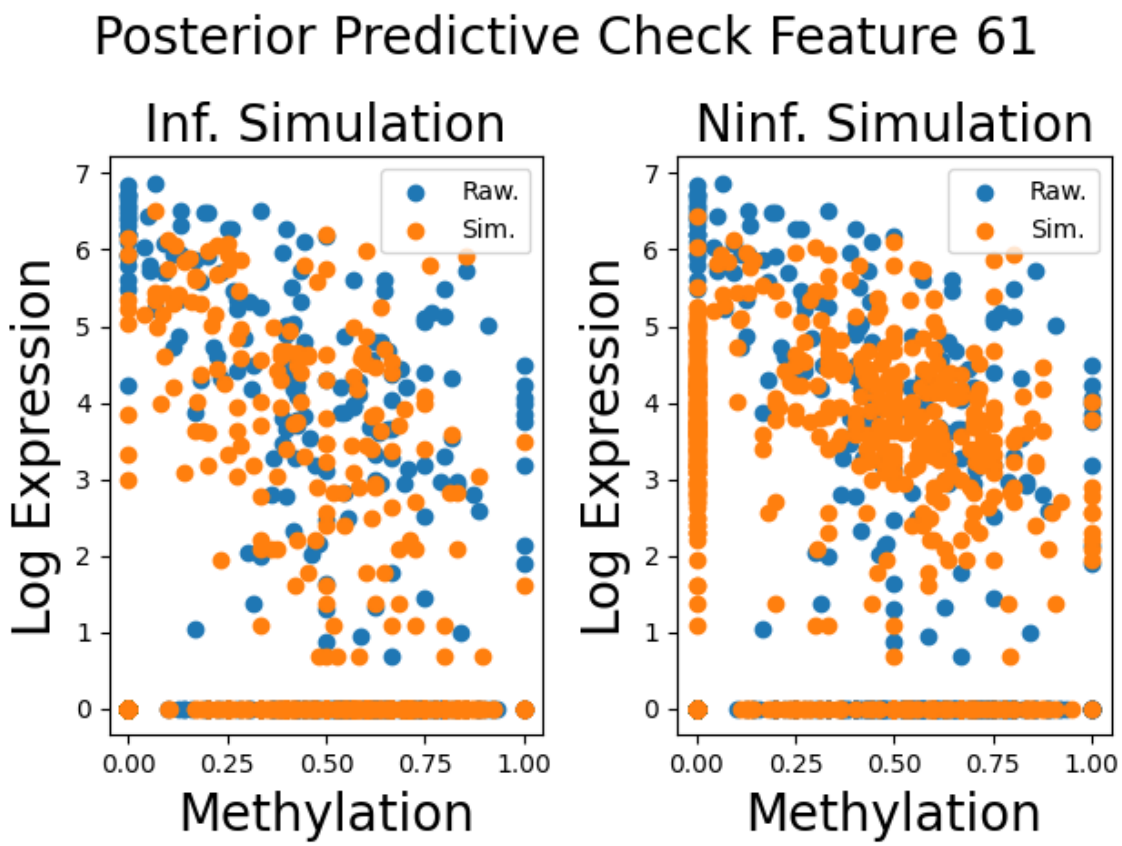


Fig. A.62 **Posterior predictive check for feature** 61 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 61 in mESC. Each dot represents a cell.

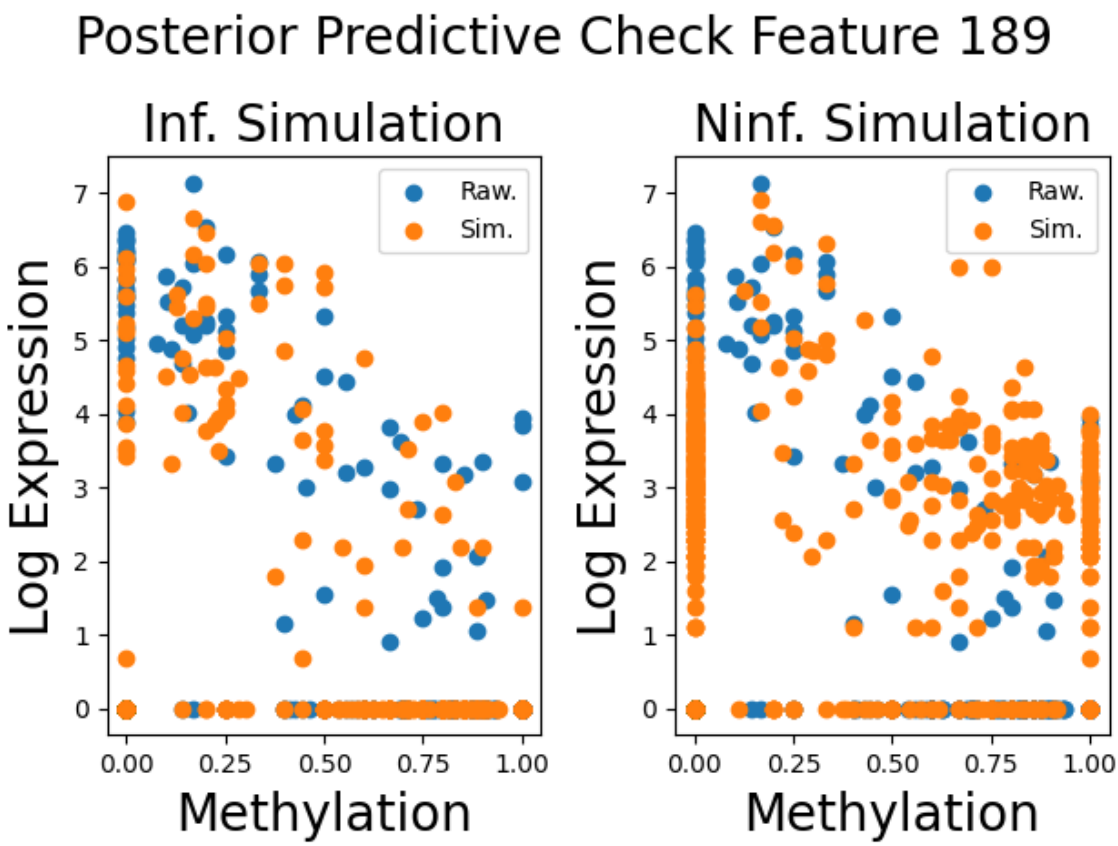


Fig. A.63 **Posterior predictive check for feature** 189 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 189 in mESC. Each dot represents a cell.

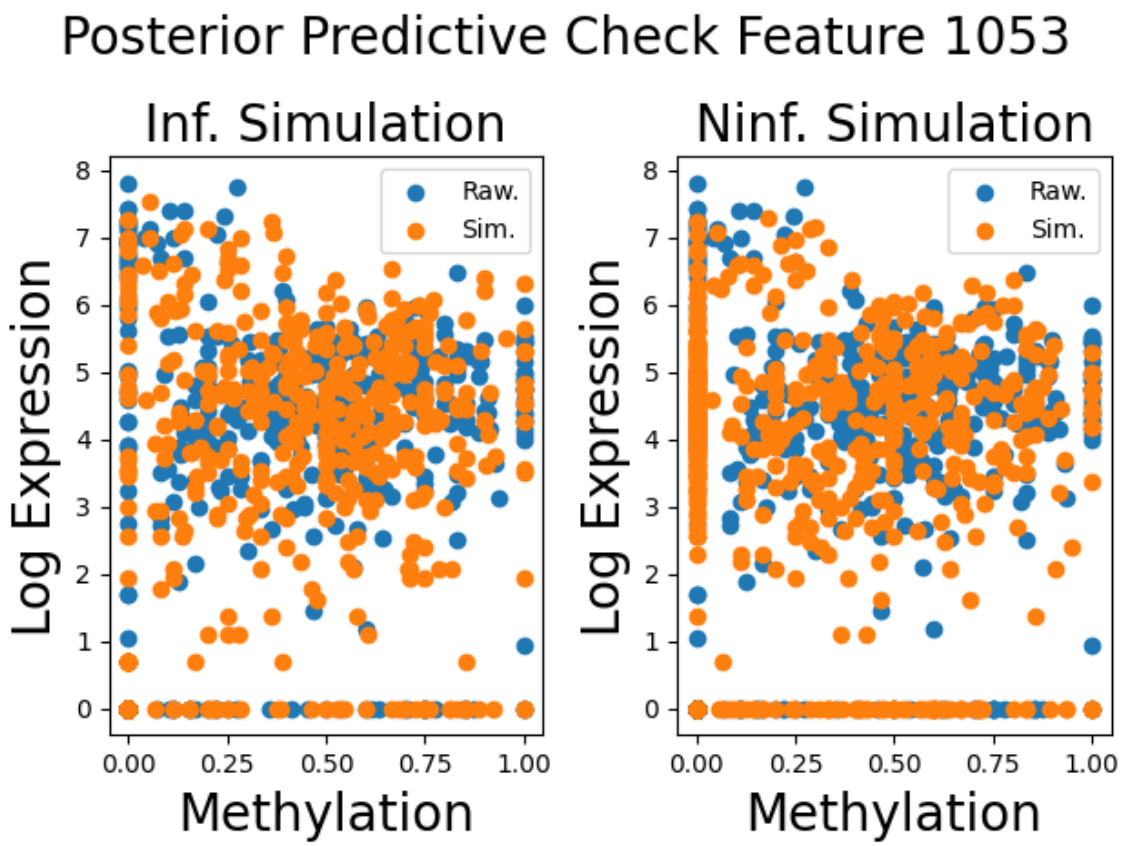


Fig. A.64 **Posterior predictive check for feature** 1053 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1053 in mESC. Each dot represents a cell.

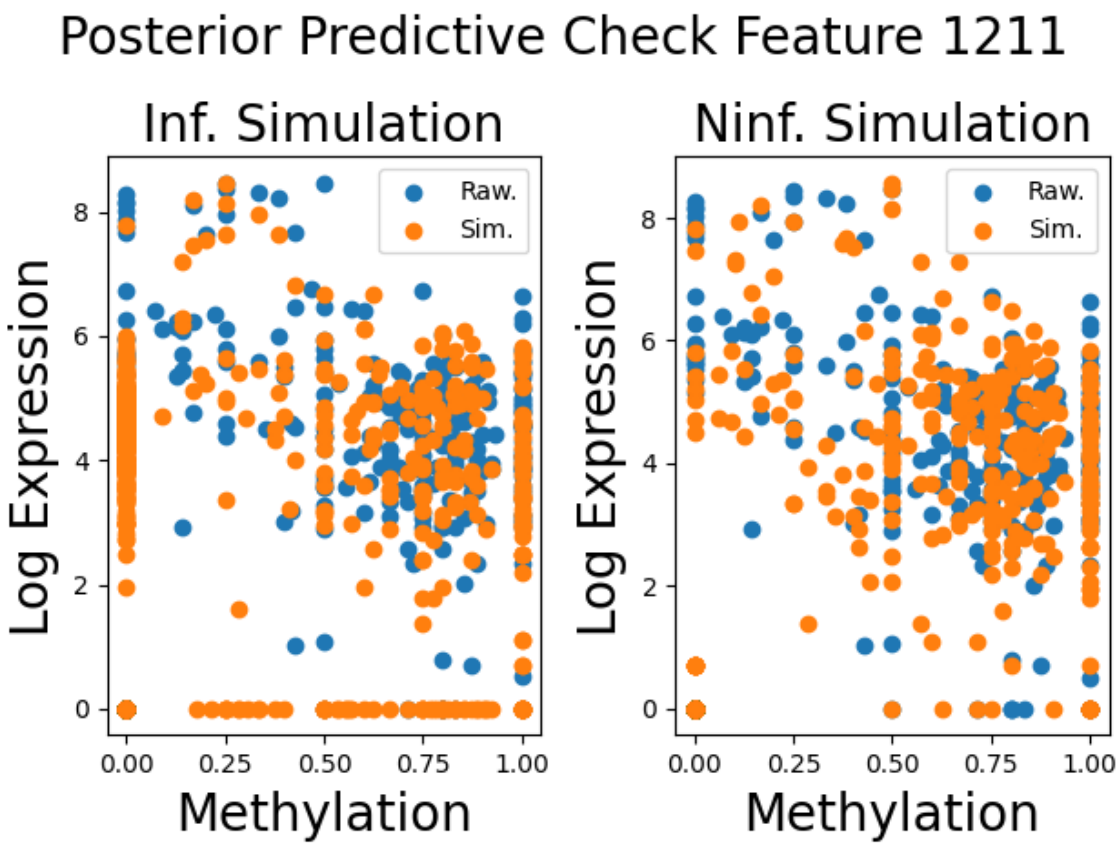


Fig. A.65 **Posterior predictive check for feature** 1211 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1211 in mESC. Each dot represents a cell.

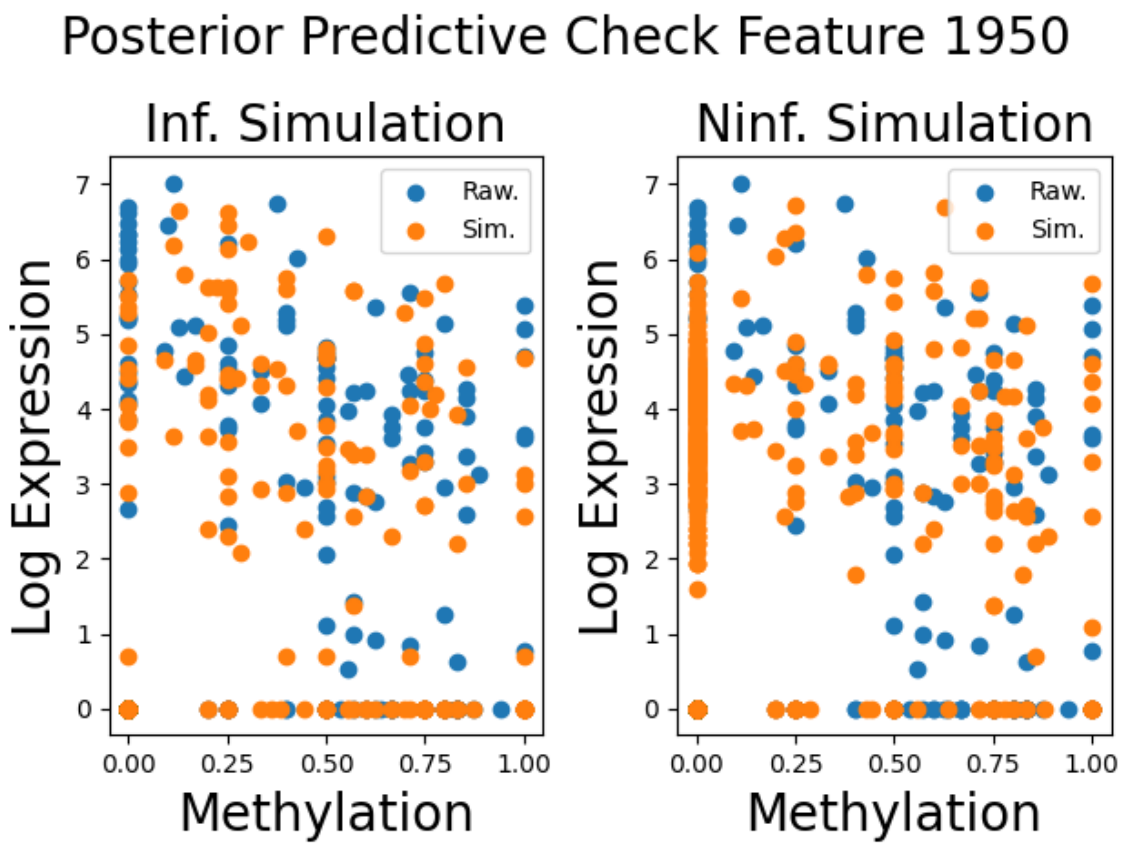


Fig. A.66 **Posterior predictive check for feature** 1950 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1950 in mESC. Each dot represents a cell.

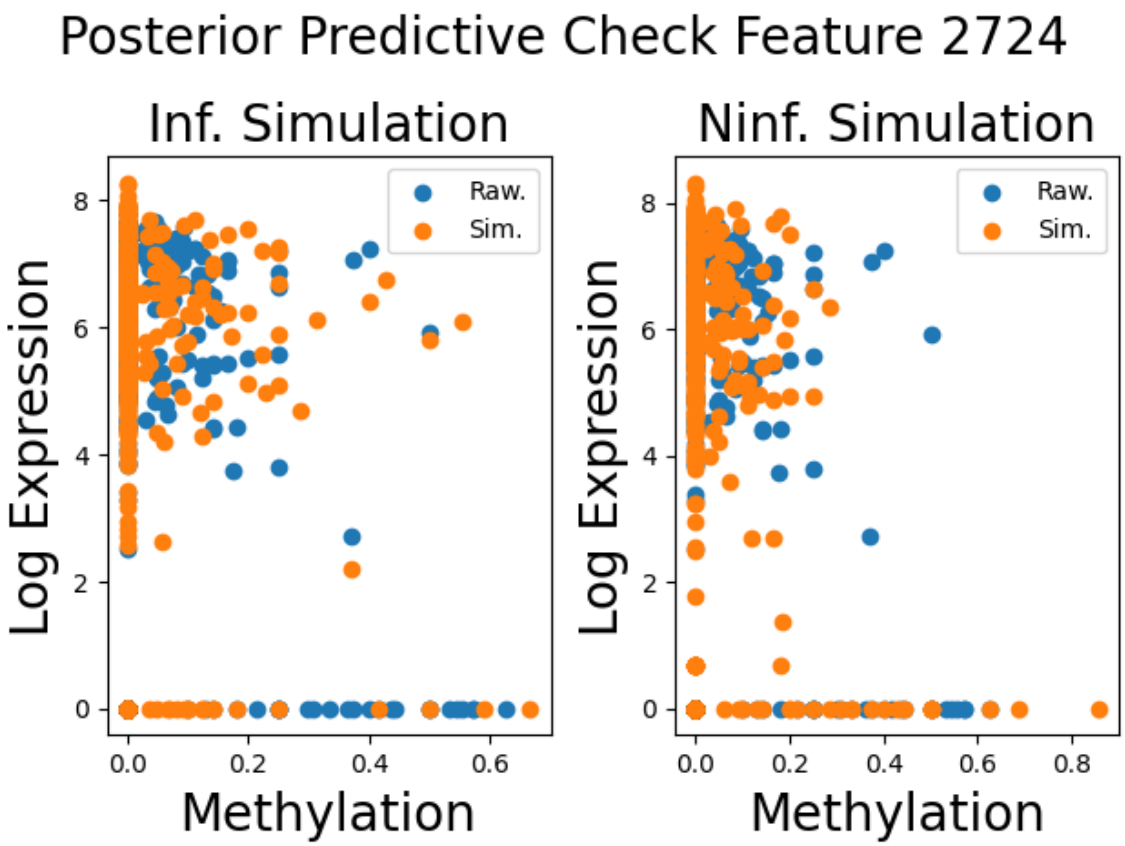


Fig. A.67 **Posterior predictive check for feature** 2724 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 2724 in mESC. Each dot represents a cell.

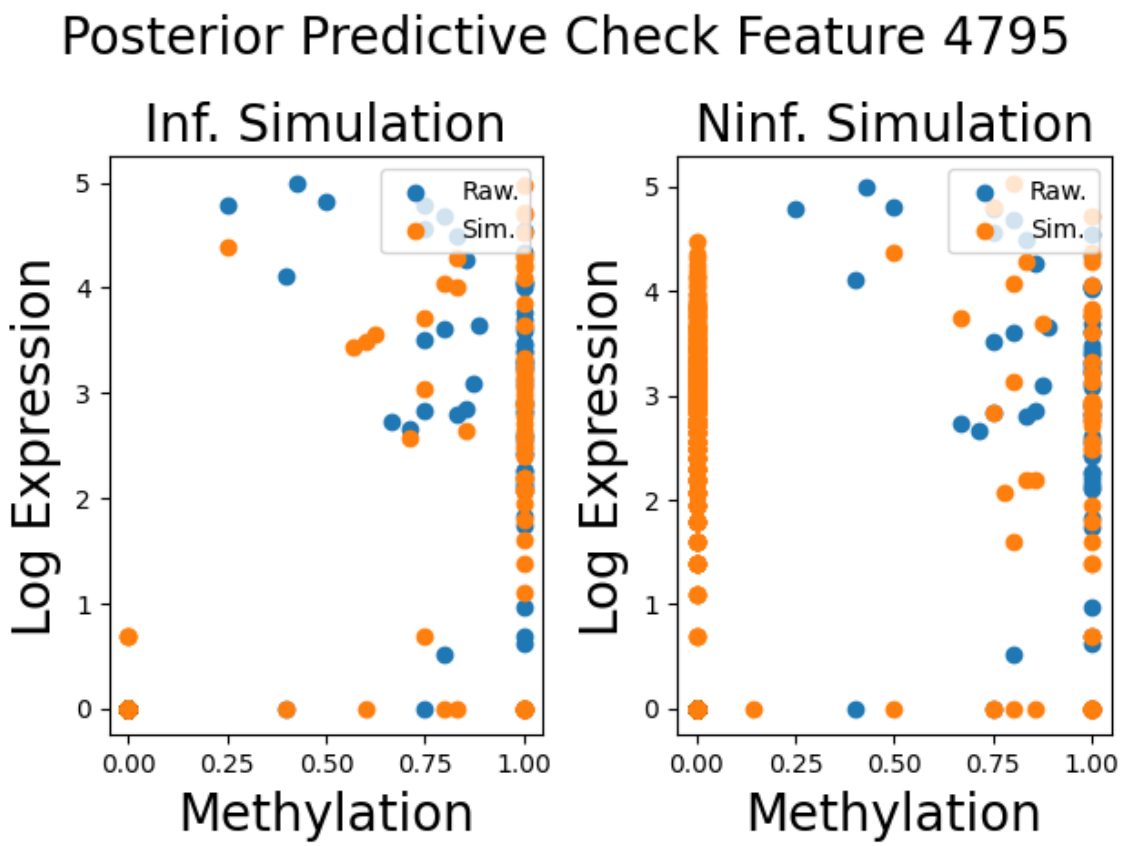


Fig. A.68 **Posterior predictive check for feature** 4795 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 4795 in mESC. Each dot represents a cell.

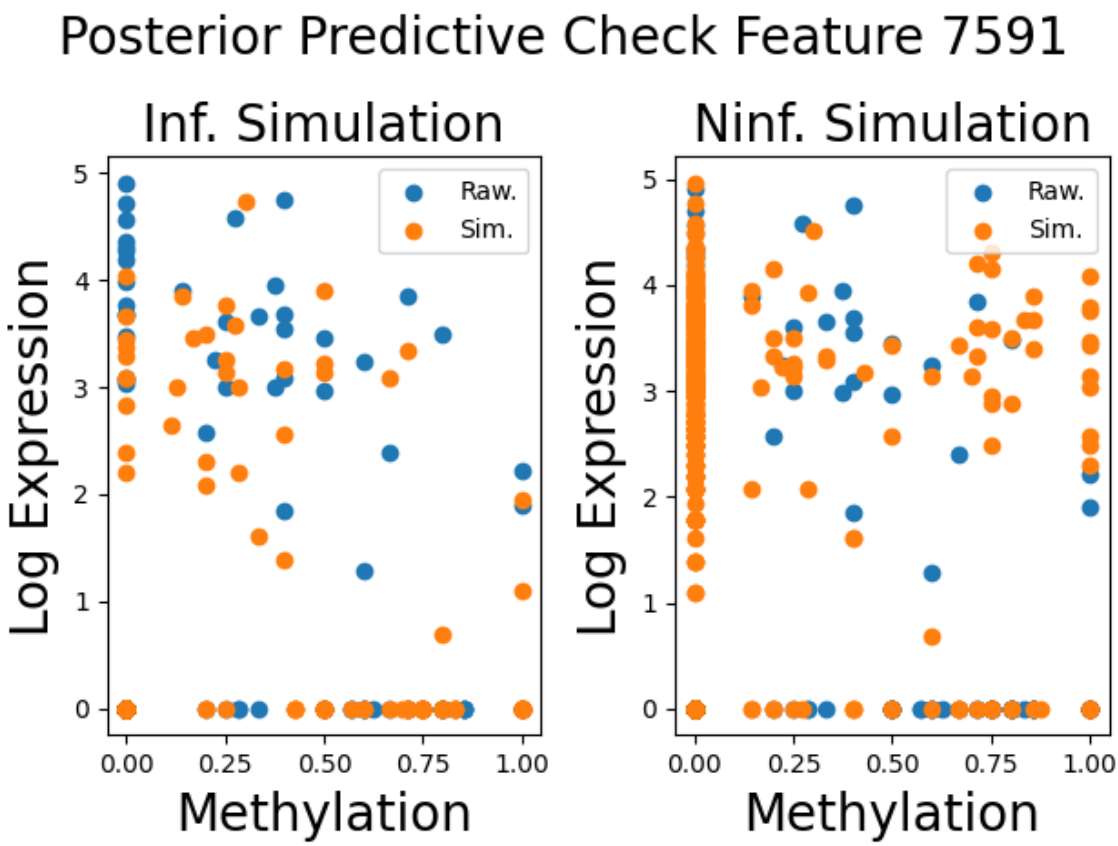


Fig. A.69 **Posterior predictive check for feature** 7591 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 7591 in mESC. Each dot represents a cell.

A.10.2 Posterior predictive checks for mEBC

In this section we present posterior predictive checks for SCRaPL with and without inflation in mEBC data.

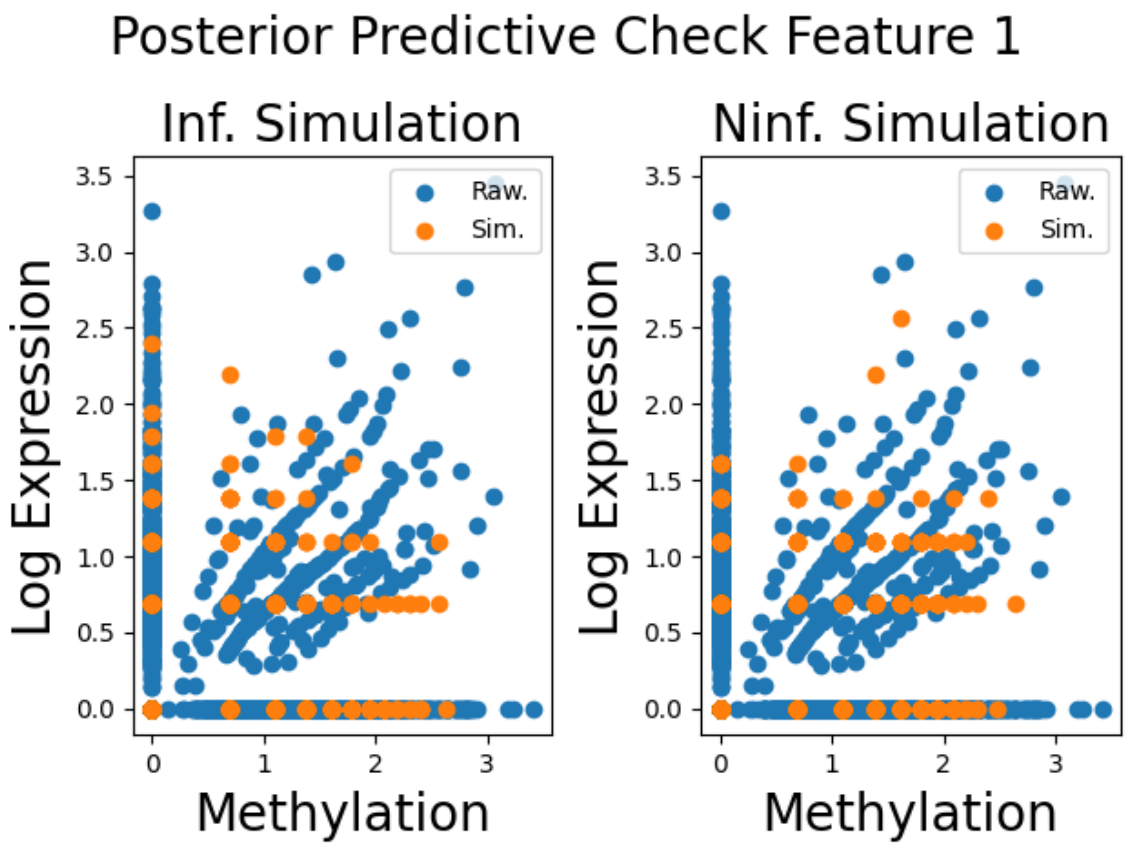


Fig. A.70 **Posterior predictive check for feature** 1 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1 in mESC. Each dot represents a cell.

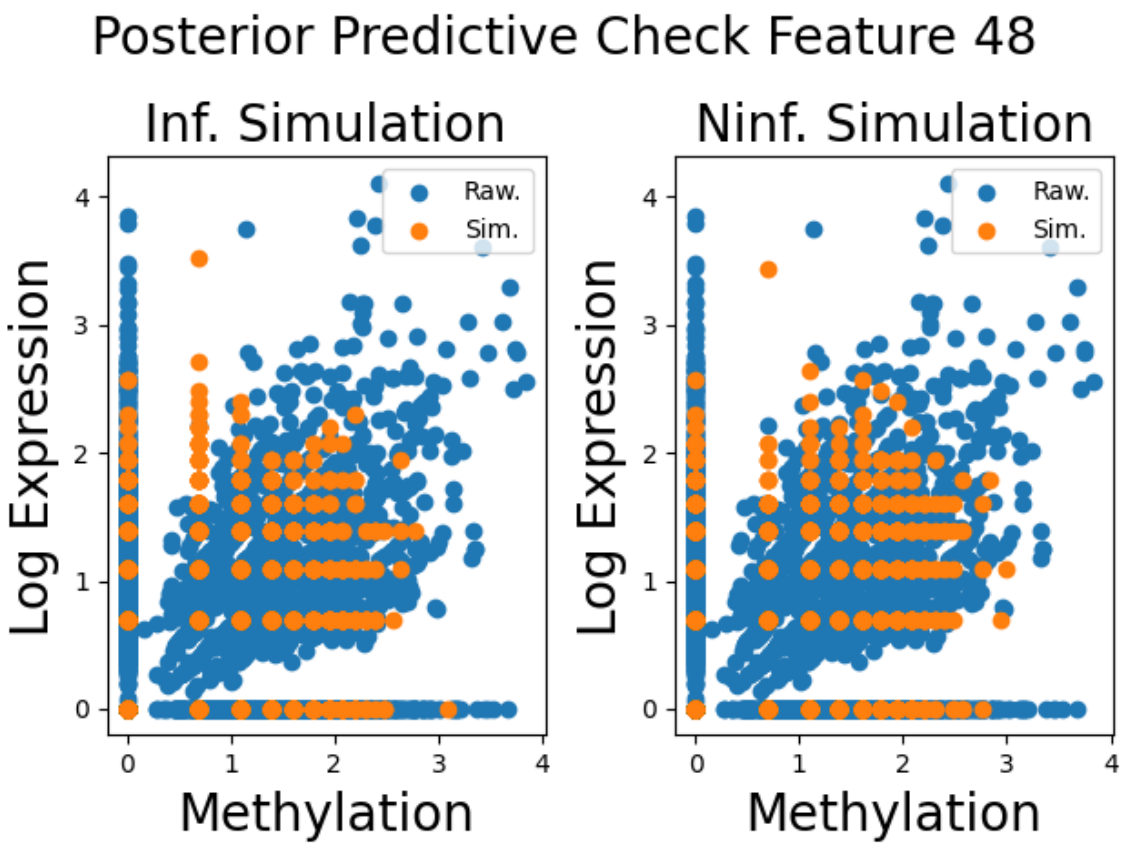


Fig. A.71 **Posterior predictive check for feature** 48 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 48 in mESC. Each dot represents a cell.

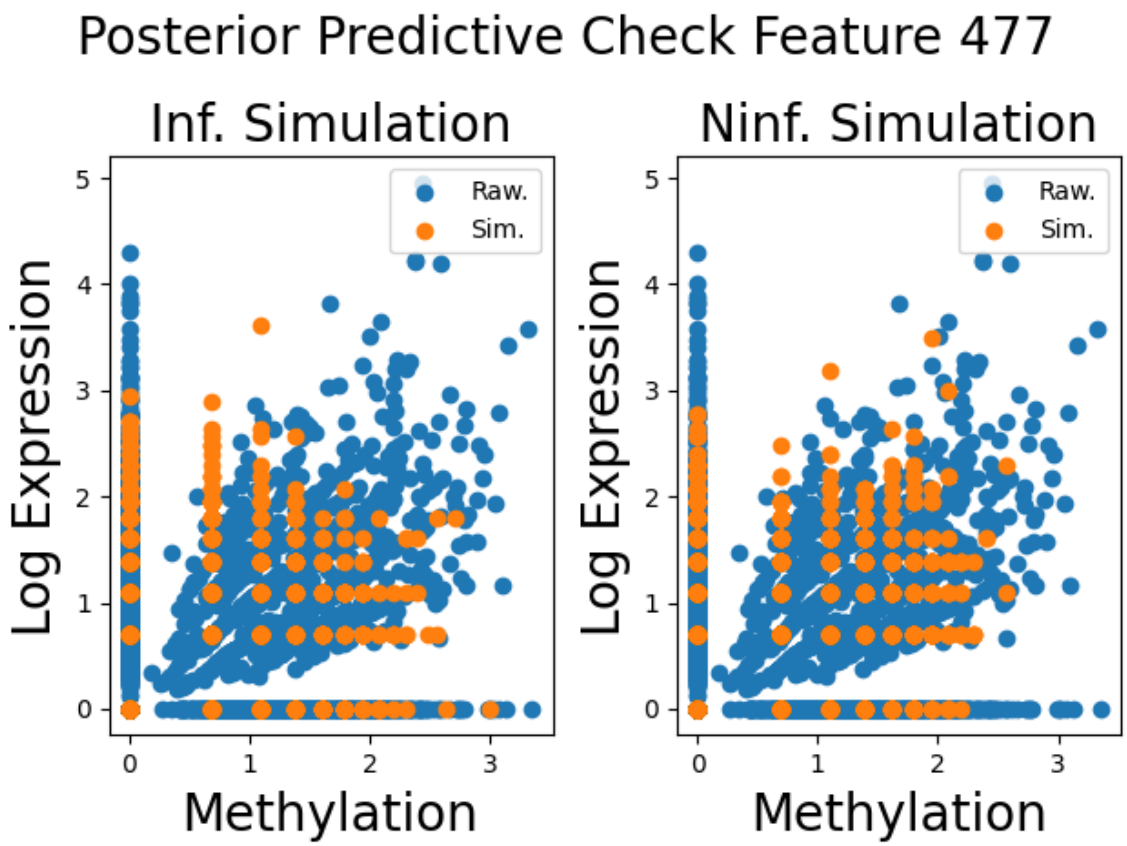


Fig. A.72 **Posterior predictive check for feature** 477 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 477 in mESC. Each dot represents a cell.

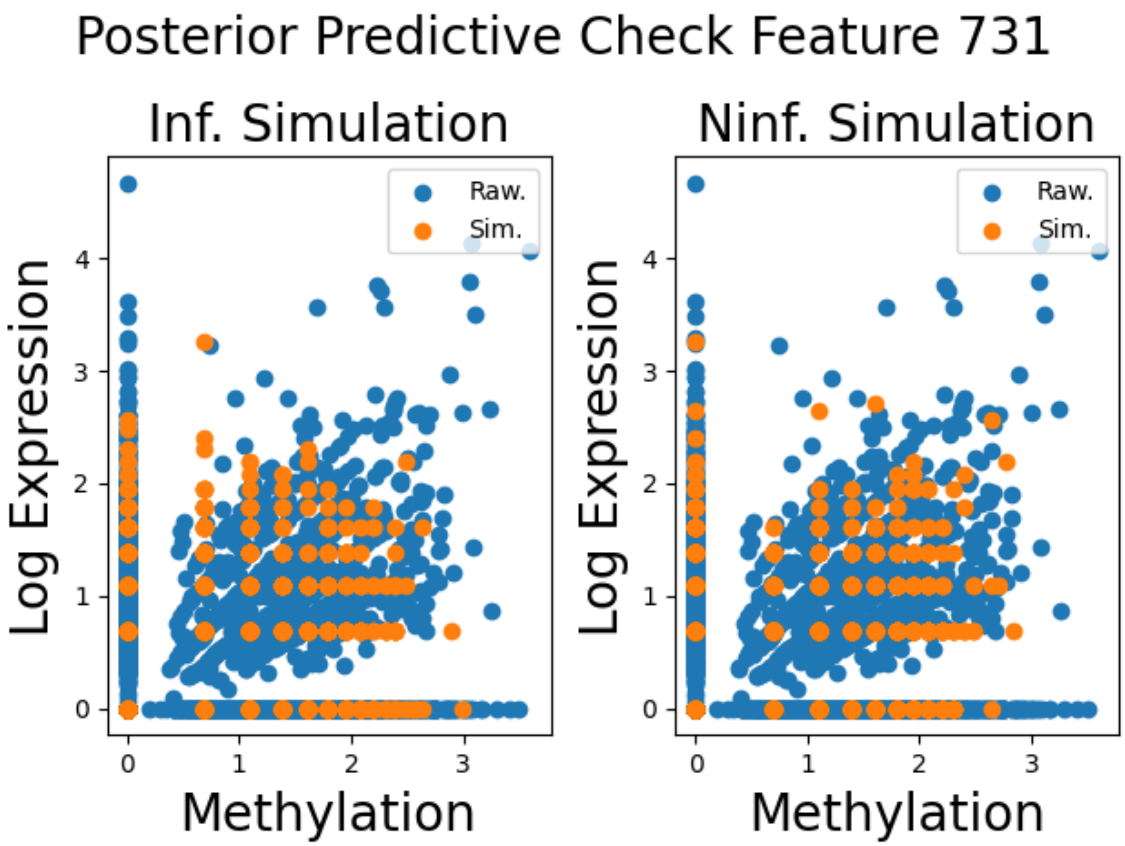


Fig. A.73 **Posterior predictive check for feature** 731 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 731 in mESC. Each dot represents a cell.

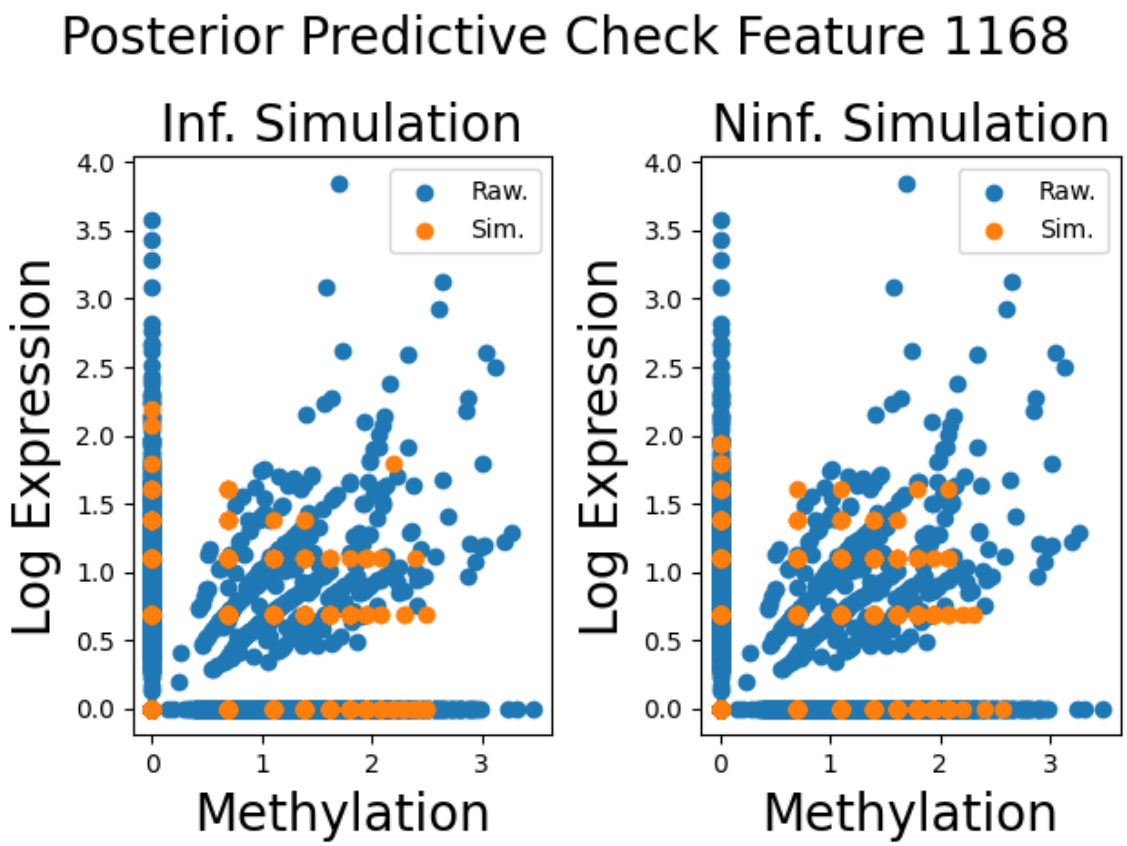


Fig. A.74 **Posterior predictive check for feature** 1168 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1168 in mESC. Each dot represents a cell.

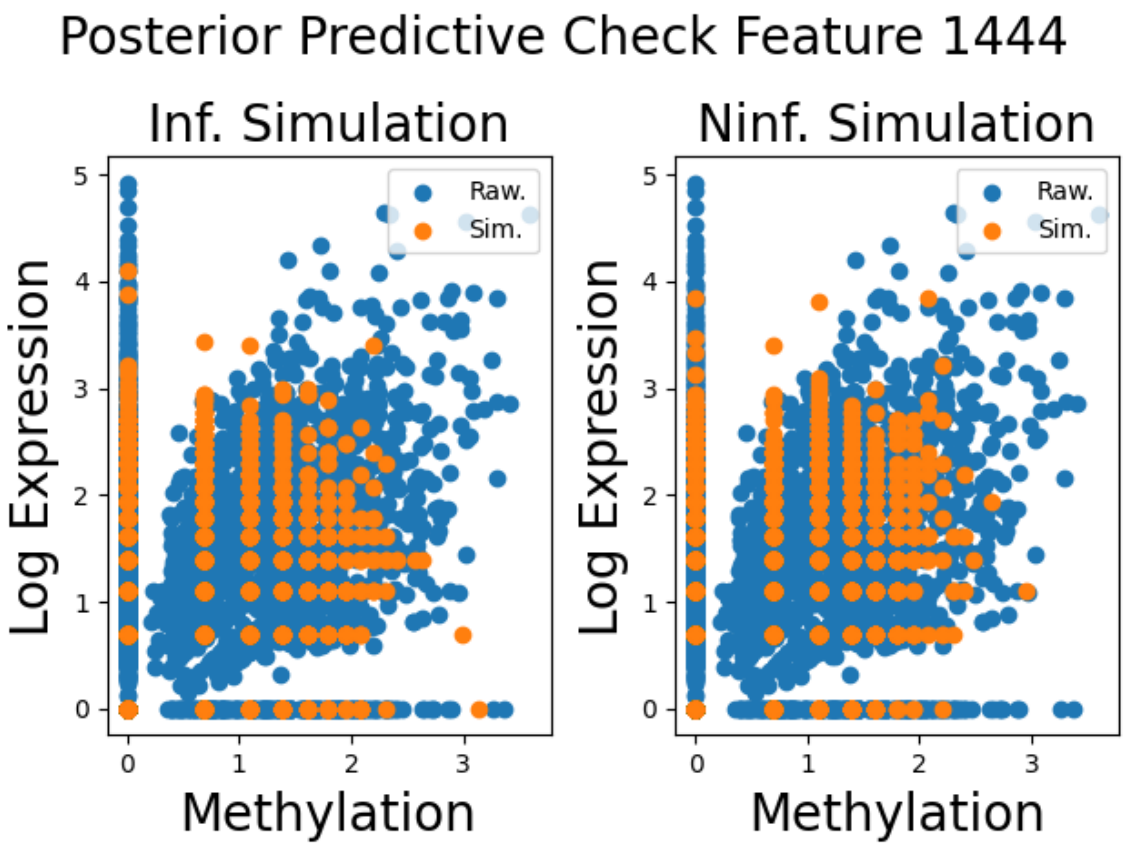


Fig. A.75 **Posterior predictive check for feature** 1444 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 1444 in mESC. Each dot represents a cell.

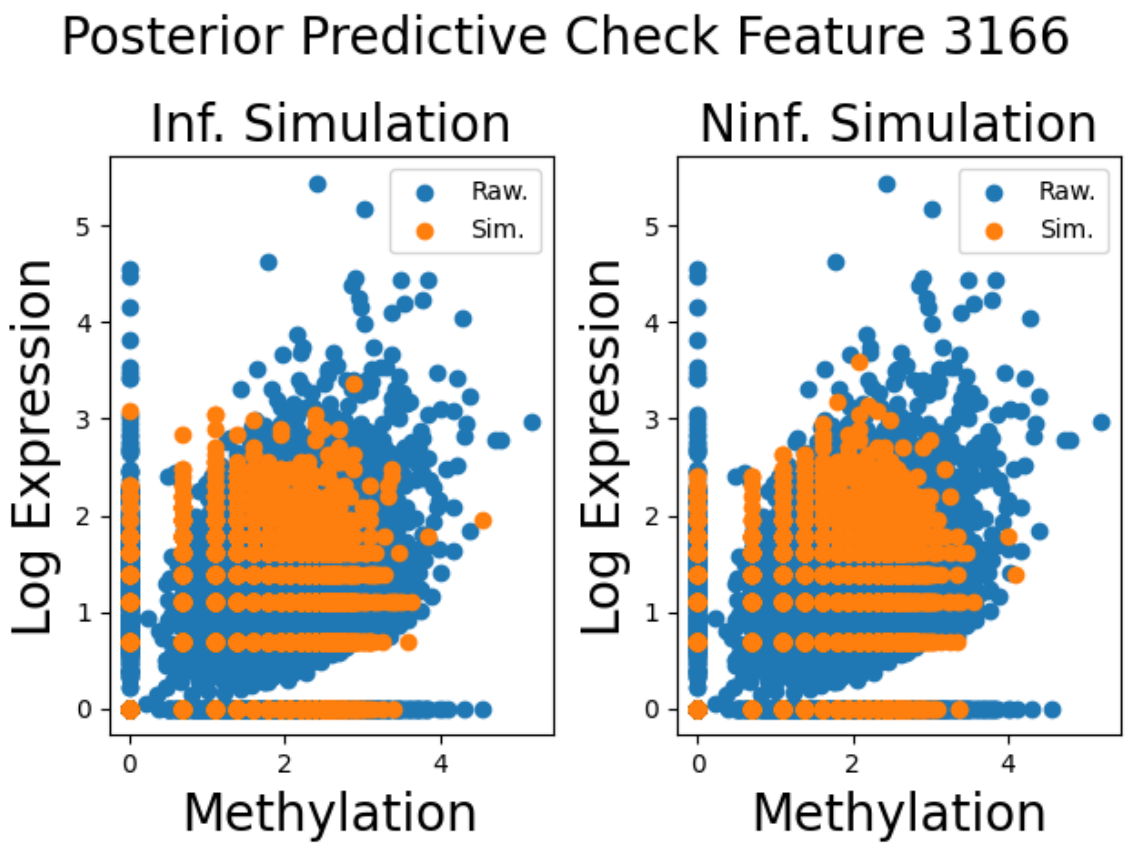


Fig. A.76 **Posterior predictive check for feature** 3166 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 3166 in mESC. Each dot represents a cell.

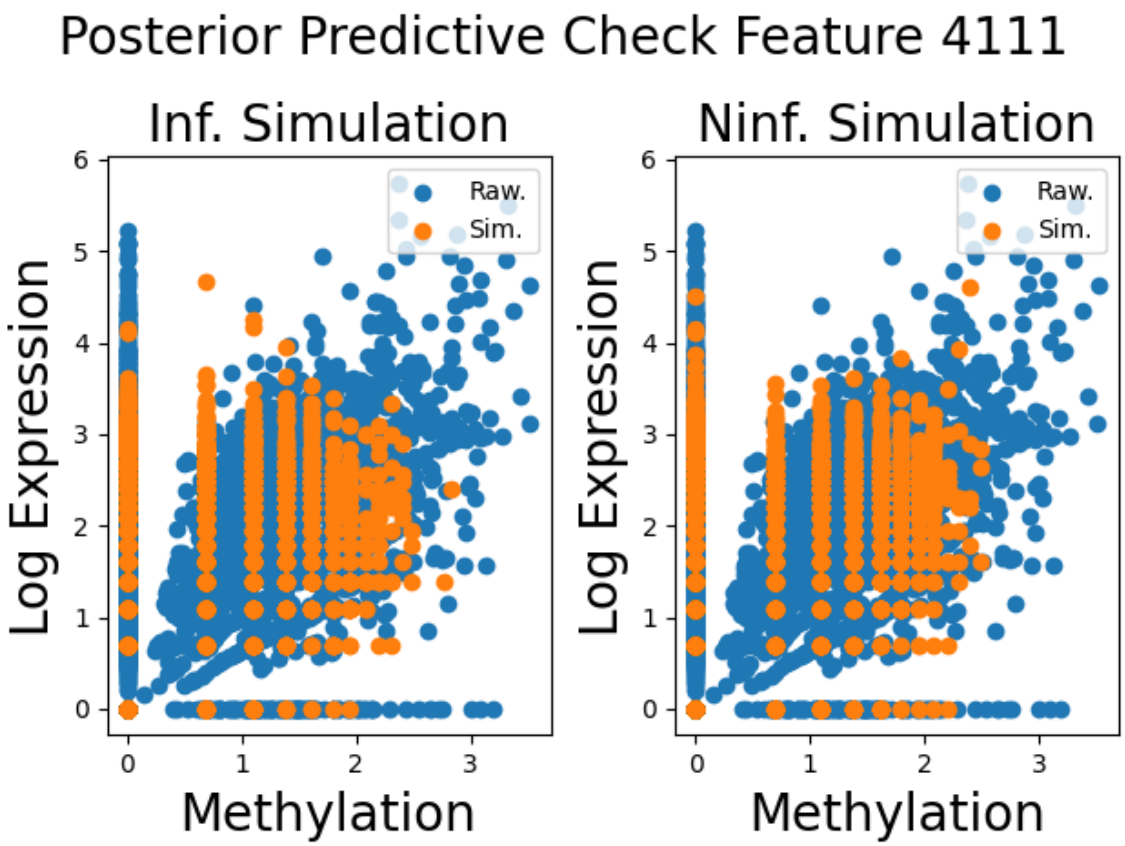


Fig. A.77 **Posterior predictive check for feature** 4111 **of mESC data.** Simulated and raw comparison scatter-plot for SCRaPL with inflate (left) and SCRaPL without inflation (right) for feature 4111 in mESC. Each dot represents a cell.

A.11 SCRaPL without over-dispersion

For each cell i ($\in \{1,\ldots I\}$) and genomic region j ($\in \{1,\ldots J\}$), observed transcriptomic and epigenomic marks $\mathbf{Y}_{ij}$ are modeled using modality specific likelihoods parameterized to depend on latent bivariate Gaussian vector $\mathbf{Z}_{ij}$ with unknown latent mean μ_j and covariance Σ_j.

$$\mathbf{Z}_{ij} = \begin{pmatrix} Z_{ij1} \\ Z_{ij2} \end{pmatrix} \mid \mu_j, \Sigma_j \overset{ind}{\sim} \mathrm{N}\left(\mu_j, \Sigma_j\right), \tag{A.24}$$

where

$$\mu_j = \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \end{pmatrix} \text{ and } \Sigma_j = \begin{pmatrix} \sigma_{j1}^2 & \rho_j \sigma_{j1} \sigma_{j2} \\ \rho_j \sigma_{j1} \sigma_{j2} & \sigma_{j2}^2 \end{pmatrix}. \tag{A.25}$$

Similarly to Fig 5.2, we assume independence across all features, which will be analyzed separately.

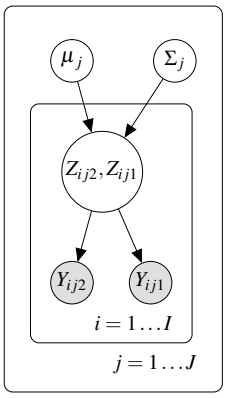


Fig. A.78 Graphical representations of model wihtout over-dispersion fitted to each genomic region $j \in J$ with mESC and mEBC data. Observed genomic data (Y_{ij1} is RNA expression; Y_{ij2} is Chromatin Accessibility), with their associated latent variables (Z_{ij1}, Z_{ij2}) and feature-specific model parameters (μ_j, Σ_j) capturing feature specific epigenomic and transcriptomic marks.

RNA expression noise model. Let Y_{ij1} be a random variable representing the normalized and log-transformed gene expression levels observed for each cell i and feature j. Conditional on the value of the latent variable Z_{ij1}, we assume that

$$\mathbb{P}\left(Y_{ij1} = y_{ij1} \,|\, Z_{ij1} = z_{ij1}\right) = \mathrm{N}\left(z_{ij1}, \sigma_t^2\right) \tag{A.26}$$

DNAm noise model. For each cell i and feature j, let Y_{ij2} be the logit-transformed methylation rate within a pre-specified genomic region for which DNAm reads were obtained. Conditioned on the latent variable Z_{ij2}, we assume

$$\mathbb{P}\left(Y_{ij2} = y_{ij2} \,|\, Z_{ij2} = z_{ij2}\right) = \mathrm{N}\left(z_{ij2}, \sigma_e^2\right) \tag{A.27}$$

Chromatin accessibility noise model. Despite the model being exactly the same, the meaning of the variable significantly changes depending the generating technology. If data are log-transformed normalized counts such as in droplet based protocols, then our modelling approach follows the one described for RNA expression data. If the data consists of logit-transformed rates within genomic regions, then the Normal noise model used for DNAm is applied.

Prior distribution. For the model without dispersion we use the same prior as with SCRaPL, namely

$$\mu_j \overset{iid}{\sim} \mathbf{N}(\mathbf{m},\mathbf{H}), \tag{A.28}$$

$$\sigma_{j1},\sigma_{j2} \overset{iid}{\sim} \text{Inv-Gamma}(c_1,c_2), \tag{A.29}$$

$$\rho_j \overset{iid}{\sim} \text{TB}(d_1,d_2) = \frac{1}{\text{B}\left(d_1,d_2\right)}\frac{1}{2^{d_1+d_2-1}}\left(1+\rho_j\right)^{d_1-1}\left(1-\rho_j\right)^{d_2-1} \tag{A.30}$$

Posterior distribution. For feature i the posterior distribution is given by

$$\begin{aligned}\mathbb{P}\left(\mathbf{Z}_{i1},\ldots,\mathbf{Z}_{iJ},\mu_j,\sigma_{j1},\sigma_{j2},\rho_j\,|\,\mathbf{Y}_{i1},\ldots,\mathbf{Y}_{iJ}\right) = \prod_{i=1}^{I}\text{N}\left(y_{ij2},z_{ij2},\sigma_e^2\right)\\ \text{N}\left(y_{ij1},z_{ij1},\sigma_t^2\right)\text{N}\left(\mathbf{Z}_{ij},\mu_j,\Sigma_j\right)\mathbb{P}\left(\sigma_{j1}\right)\mathbb{P}\left(\sigma_{j2}\right)\mathbb{P}\left(\rho_j\right)\text{N}(\mu_j,\mathbf{m},\mathbf{H})\end{aligned} \tag{A.31}$$

$$\begin{aligned}\mathbb{P}\left(\mu_j,\sigma_{j1},\sigma_{j2},\rho_j\,|\,\mathbf{Y}_{i1},\ldots,\mathbf{Y}_{iJ}\right) = \prod_{i=1}^{I}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\text{N}\left(y_{ij2},z_{ij2},\sigma_e^2\right)\\ \text{N}\left(y_{ij1},z_{ij1},\sigma_t^2\right)\text{N}\left(\mathbf{Z}_{ij},\mu_j,\Sigma_j\right)\mathbb{P}\left(\sigma_{j1}\right)\mathbb{P}\left(\sigma_{j2}\right)\mathbb{P}\left(\rho_j\right)\text{N}(\mu_j,\mathbf{m},\mathbf{H})d\mathbf{Z}_{i1}\ldots d\mathbf{Z}_{iJ}\end{aligned} \tag{A.32}$$

$$\begin{aligned}\mathbb{P}\left(\mu_j,\sigma_{j1},\sigma_{j2},\rho_j\,|\,\mathbf{Y}_{i1},\ldots,\mathbf{Y}_{iJ}\right) =\\ \prod_{i=1}^{I}\text{N}\left(\mathbf{Y}_{ij},\mu_j,\tilde{\Sigma}+\Sigma_j\right)\mathbb{P}\left(\sigma_{j1}\right)\mathbb{P}\left(\sigma_{j2}\right)\mathbb{P}\left(\rho_j\right)\text{N}(\mu_j,\mathbf{m},\mathbf{H})\end{aligned} \tag{A.33}$$

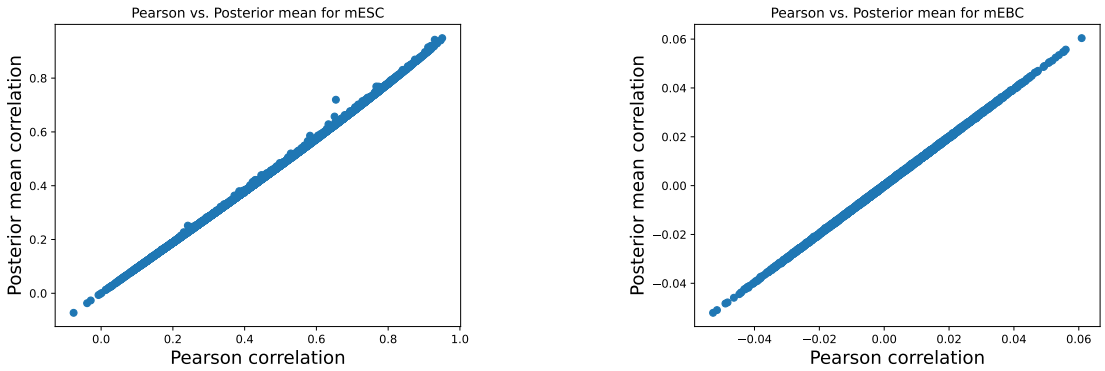
Where

$$\tilde{\Sigma} = \begin{pmatrix}\sigma_t^2 & 0\\ 0 & \sigma_e^2\end{pmatrix}. \tag{A.34}$$

$$\begin{aligned}\mathbb{P}\left(\sigma_{j1},\sigma_{j2},\rho_j\,|\,\mathbf{Y}_{i1},\ldots,\mathbf{Y}_{iJ}\right) =\\ \prod_{i=1}^{I}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\text{N}\left(\mathbf{Y}_{ij},\mu_j,\tilde{\Sigma}+\Sigma_j\right)\mathbb{P}\left(\sigma_{j1}\right)\mathbb{P}\left(\sigma_{j2}\right)\mathbb{P}\left(\rho_j\right)\text{N}(\mu_j,\mathbf{m},\mathbf{H})d\mu_j\end{aligned} \tag{A.35}$$

$$\mathbb{P}\left(\sigma_{j1},\sigma_{j2},\rho_j\,|\,\mathbf{Y}_{i1},\ldots,\mathbf{Y}_{iJ}\right) = \prod_{i=1}^{I}\text{N}\left(\mathbf{Y}_{ij},\mathbf{m},\tilde{\Sigma}+\Sigma_j+\mathbf{H}\right)\mathbb{P}\left(\sigma_{j1}\right)\mathbb{P}\left(\sigma_{j2}\right)\mathbb{P}\left(\rho_j\right) \tag{A.36}$$

In practise $\tilde{\Sigma} = \mathbf{H} = 0.01\mathbb{I}$ which yields a Bayesian version of Pearson correlation as it is indicated from scatter plots of Fig. A.79.



(a) (b)

Fig. A.79 **Pearson correlation and Normal likelihood model correlation comparison.**(A.79a) Pearson correlation and Normal likelihood model correlation comparison in mESC. (A.79b) Pearson correlation and Normal likelihood model correlation comparison in mEBC.

A.12 Prior Sensitivity

In this section we examine posterior sensitivity for different priors in mESC and mEBC data. We test correlation, standard deviation and mean priors with different parameters to asses their impact on posterior covariance. Then we present results for the same features that had been considered in previous analyses.

A.12.1 Prior Sensitivity for mESC data

In this subsection we present the prior sensitivity analysis for mESC data.

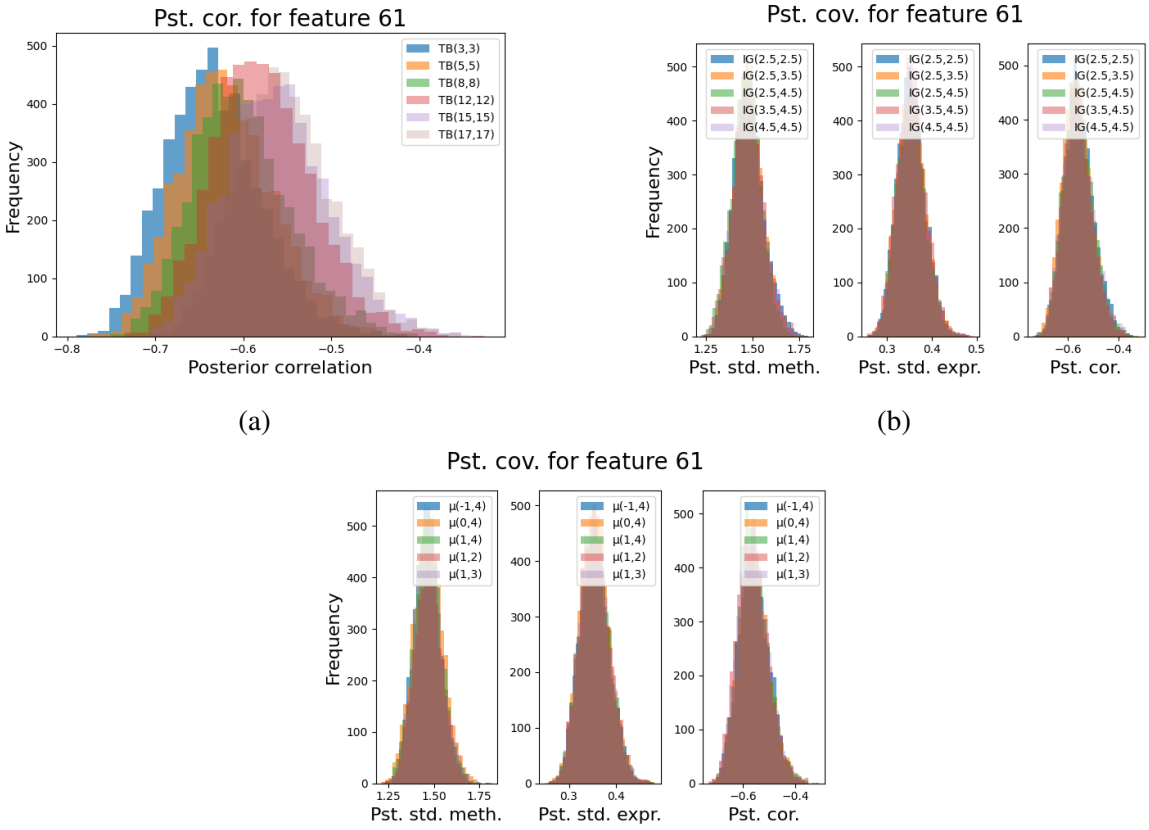
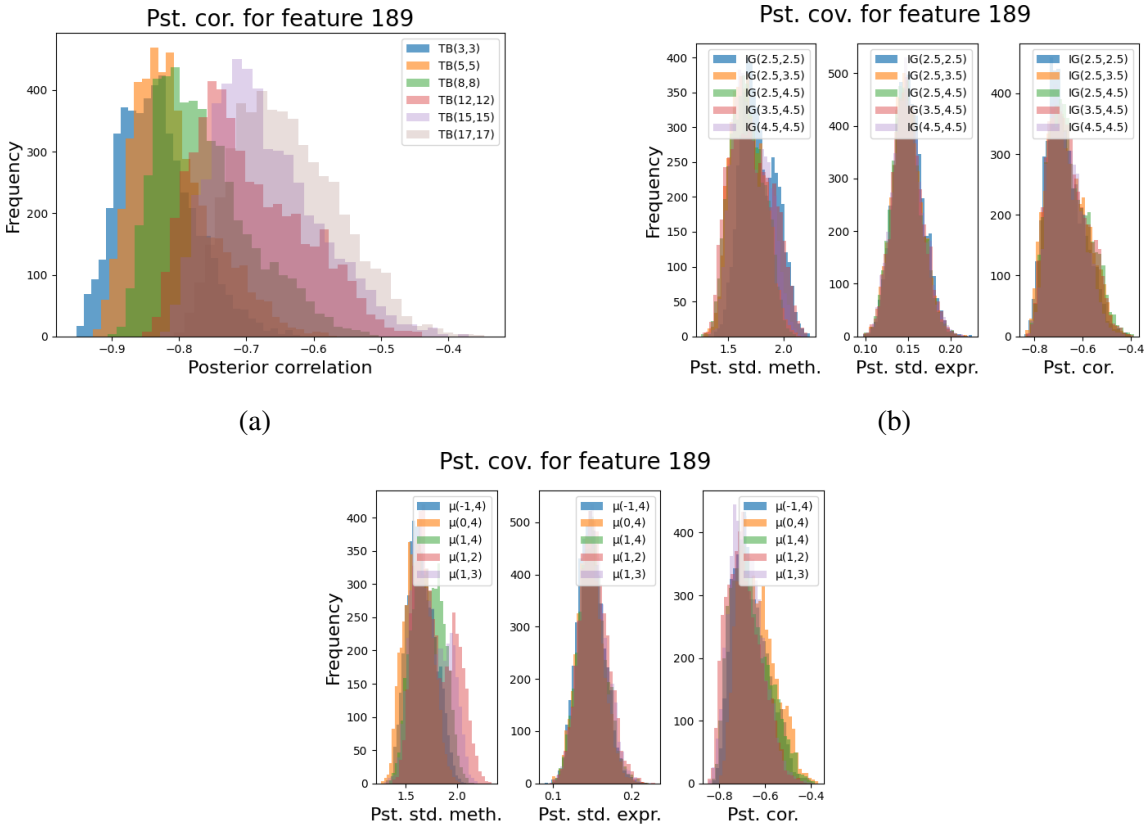


Fig. A.80 **Posterior correlation/covariance for feature** 61 **with different correlation, standard deviation and mean priors in mESC data.**(A.80a) Posterior correlation distribution for feature 61 and different correlation. (A.80b) Posterior correlation distribution for feature 61 and different correlation. (A.80c) Posterior covariance distribution for feature 61 and different mean priors in mESC data.



(a) (b)

(c)

Fig. A.81 **Posterior correlation/covariance for feature** 189 **with different correlation, standard deviation and mean priors in mESC data.**(A.81a) Posterior correlation distribution for feature 189 and different correlation. (A.81b) Posterior correlation distribution for feature 189 and different correlation. (A.81c) Posterior covariance distribution for feature 189 and different mean priors in mESC data.

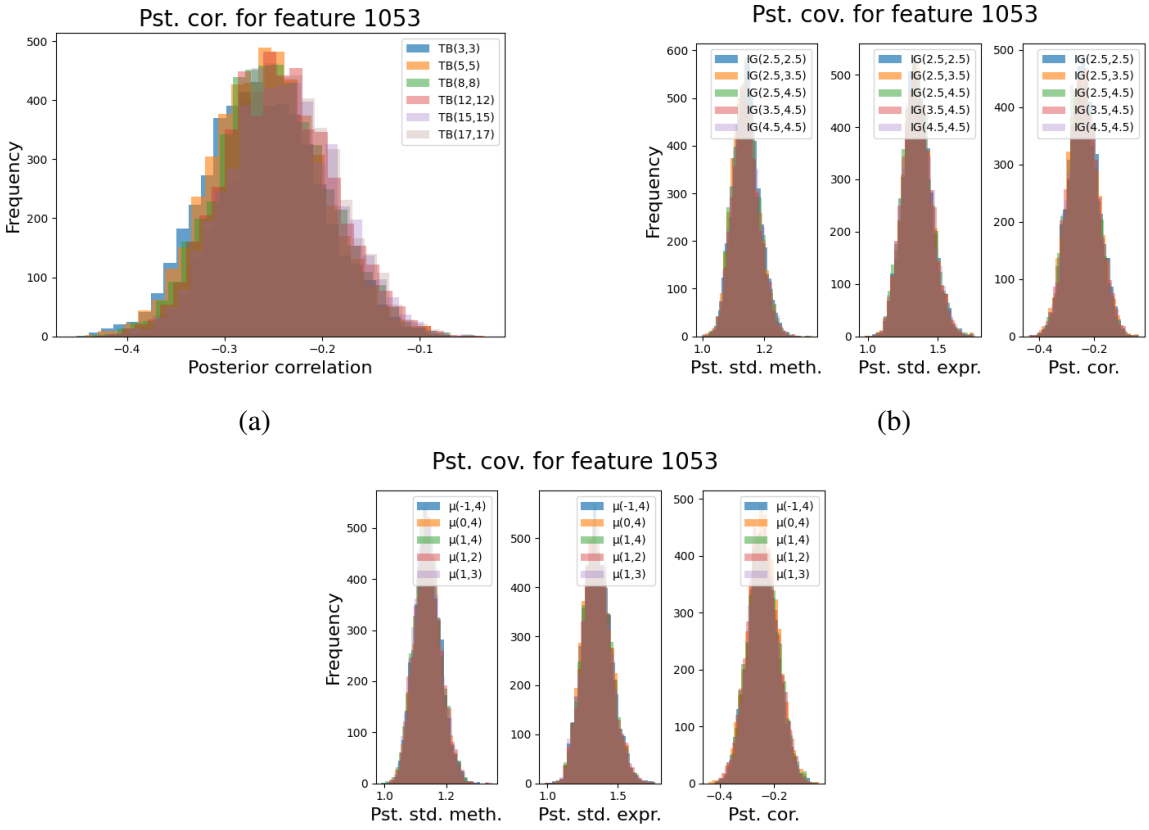


Fig. A.82 **Posterior correlation/covariance for feature** 1053 **with different correlation, standard deviation and mean priors in mESC data.**(A.82a) Posterior correlation distribution for feature 1053 and different correlation. (A.82b) Posterior correlation distribution for feature 1053 and different correlation. (A.82c) Posterior covariance distribution for feature 1053 and different mean priors in mESC data.

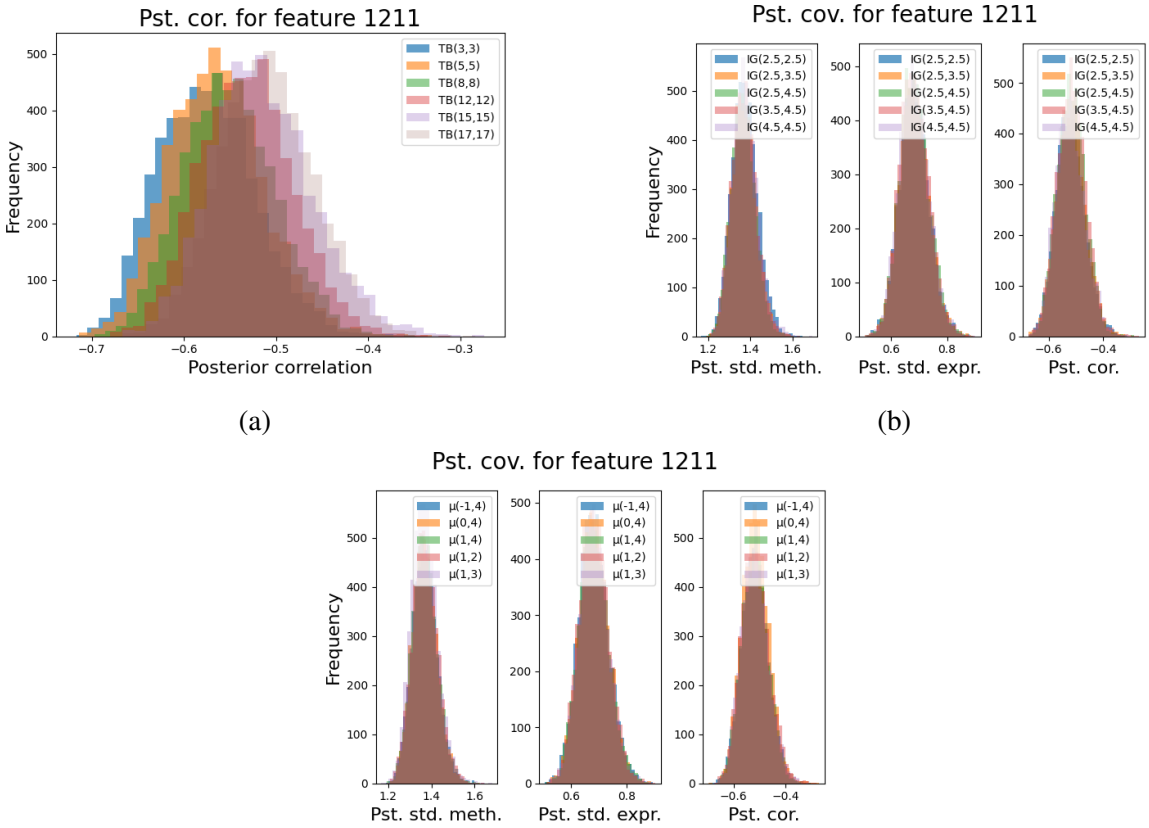
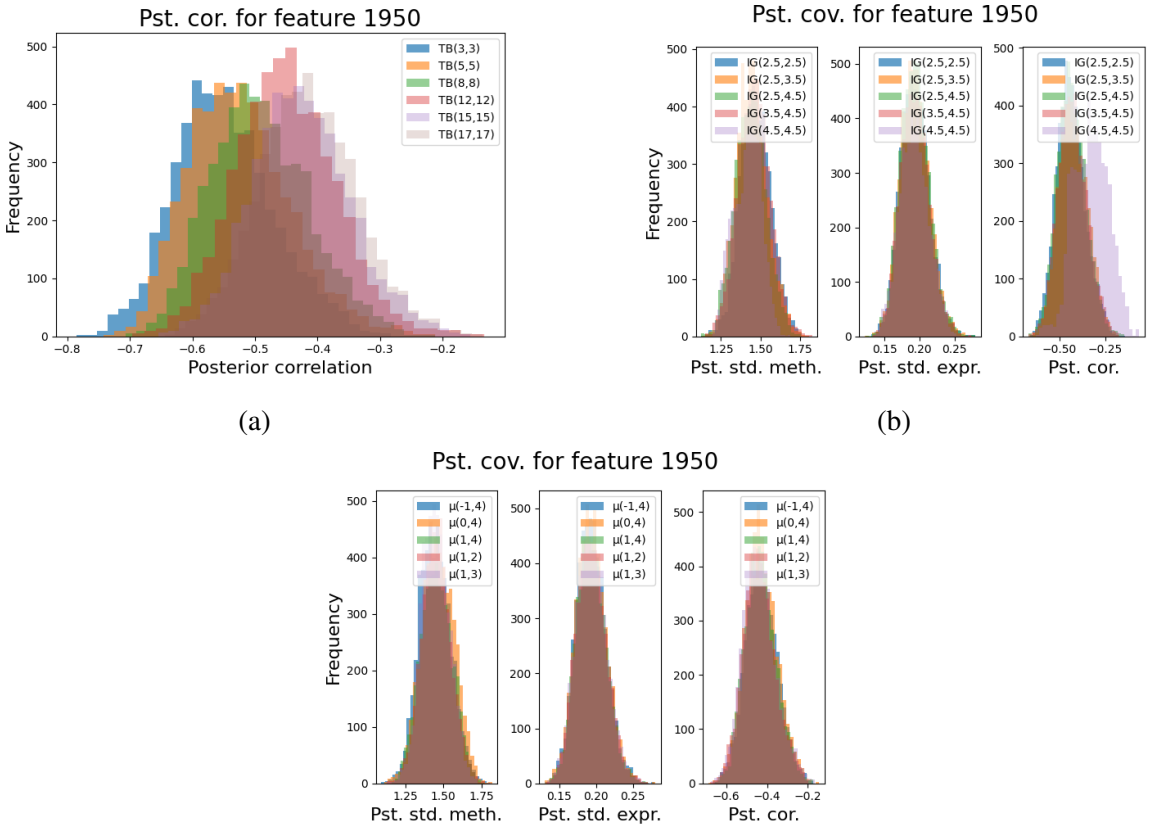


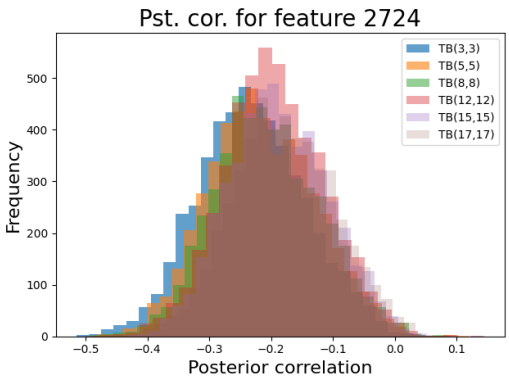
Fig. A.83 **Posterior correlation/covariance for feature** 1211 **with different correlation, standard deviation and mean priors in mESC data.**(A.83a) Posterior correlation distribution for feature 1211 and different correlation. (A.83b) Posterior correlation distribution for feature 1211 and different correlation. (A.83c) Posterior covariance distribution for feature 1211 and different mean priors in mESC data.



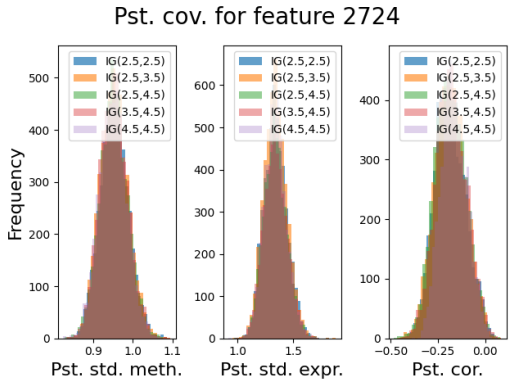
(a) (b)

(c)

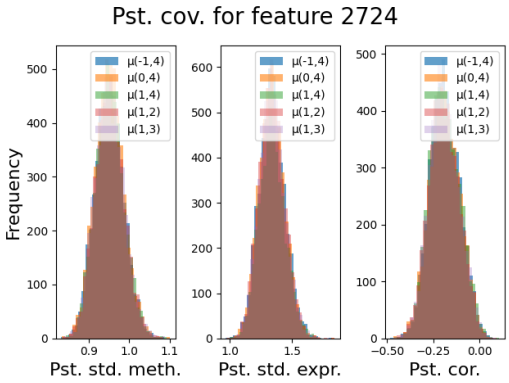
Fig. A.84 **Posterior correlation/covariance for feature** 1950 **with different correlation, standard deviation and mean priors in mESC data.**(A.84a) Posterior correlation distribution for feature 1950 and different correlation. (A.84b) Posterior correlation distribution for feature 1950 and different correlation. (A.84c) Posterior covariance distribution for feature 1950 and different mean priors in mESC data.



(a)



(b)



(c)

Fig. A.85 **Posterior correlation/covariance for feature** 2724 **with different correlation, standard deviation and mean priors in mESC data.**(A.85a) Posterior correlation distribution for feature 2724 and different correlation. (A.85b) Posterior correlation distribution for feature 2724 and different correlation. (A.85c) Posterior covariance distribution for feature 2724 and different mean priors in mESC data.

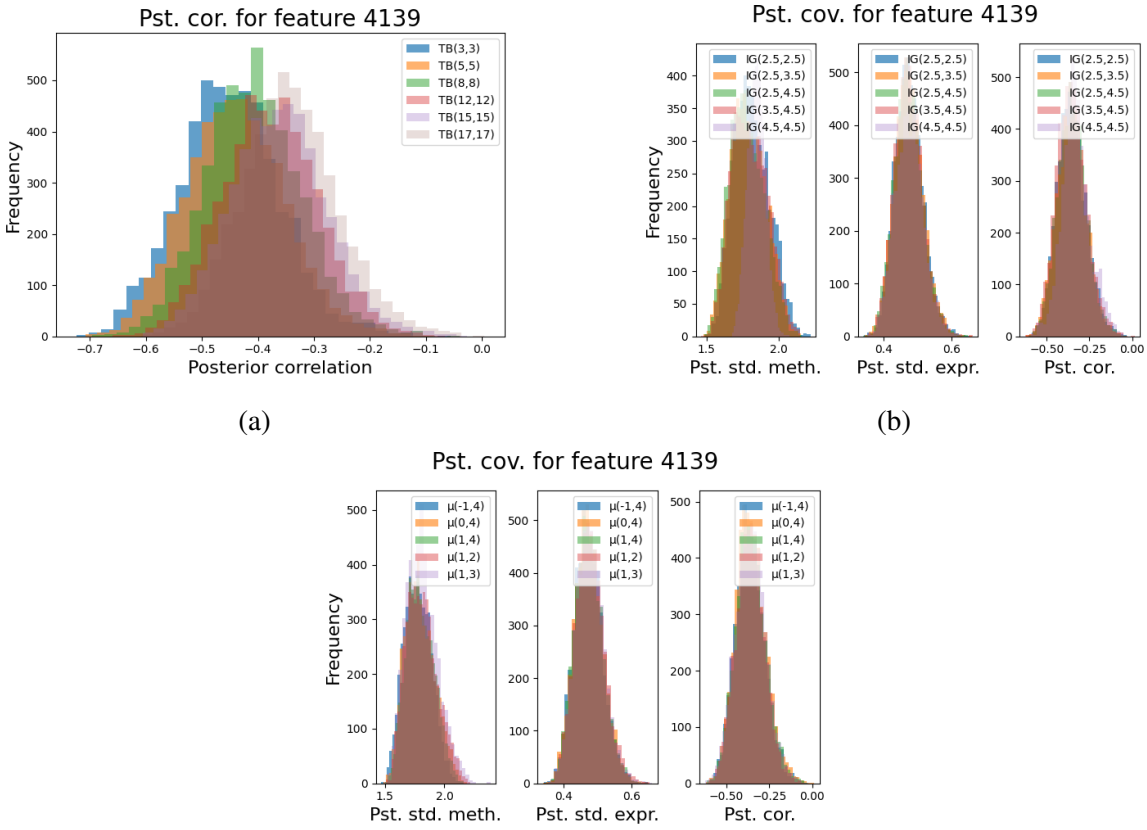
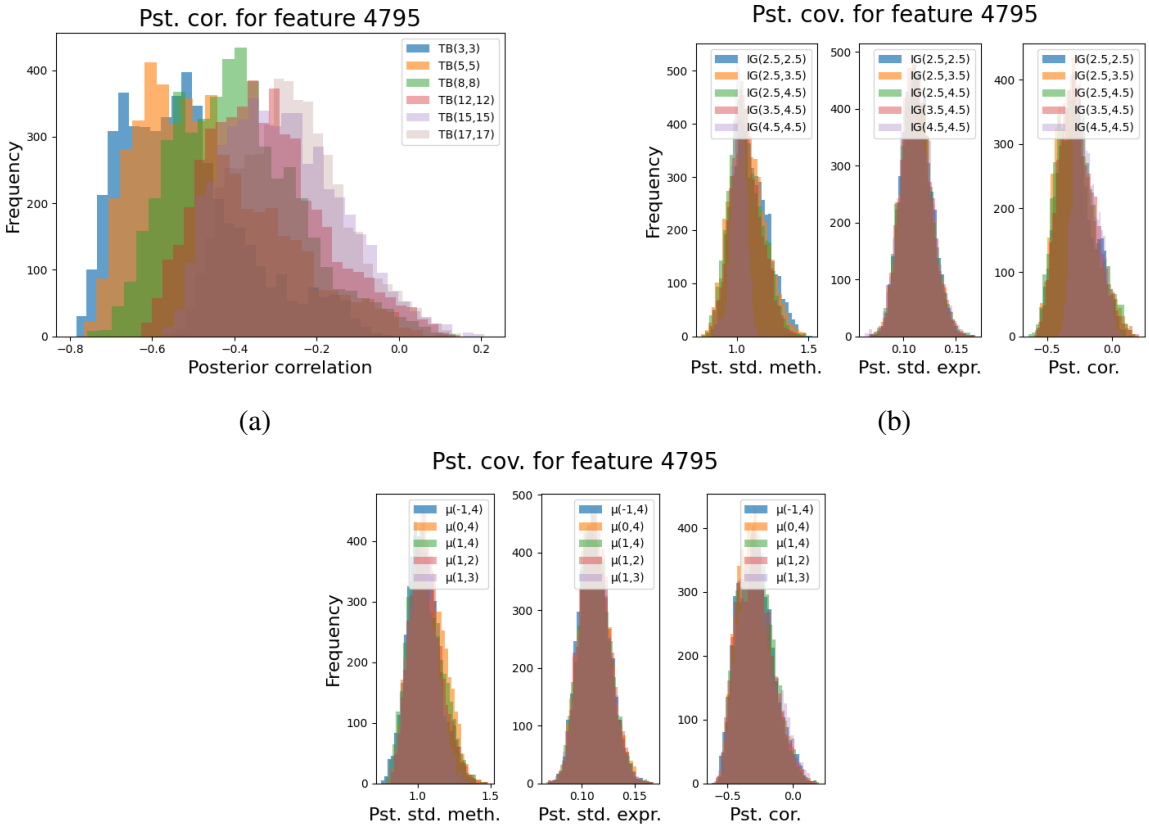


Fig. A.86 **Posterior correlation/covariance for feature** 4139 **with different correlation, standard deviation and mean priors in mESC data.**(A.86a) Posterior correlation distribution for feature 4139 and different correlation. (A.86b) Posterior correlation distribution for feature 4139 and different correlation. (A.86c) Posterior covariance distribution for feature 4139 and different mean priors in mESC data.



(a) (b)

(c)

Fig. A.87 **Posterior correlation/covariance for feature** 4795 **with different correlation, standard deviation and mean priors in mESC data.**(A.87a) Posterior correlation distribution for feature 4795 and different correlation. (A.87b) Posterior correlation distribution for feature 4795 and different correlation. (A.87c) Posterior covariance distribution for feature 4795 and different mean priors in mESC data.

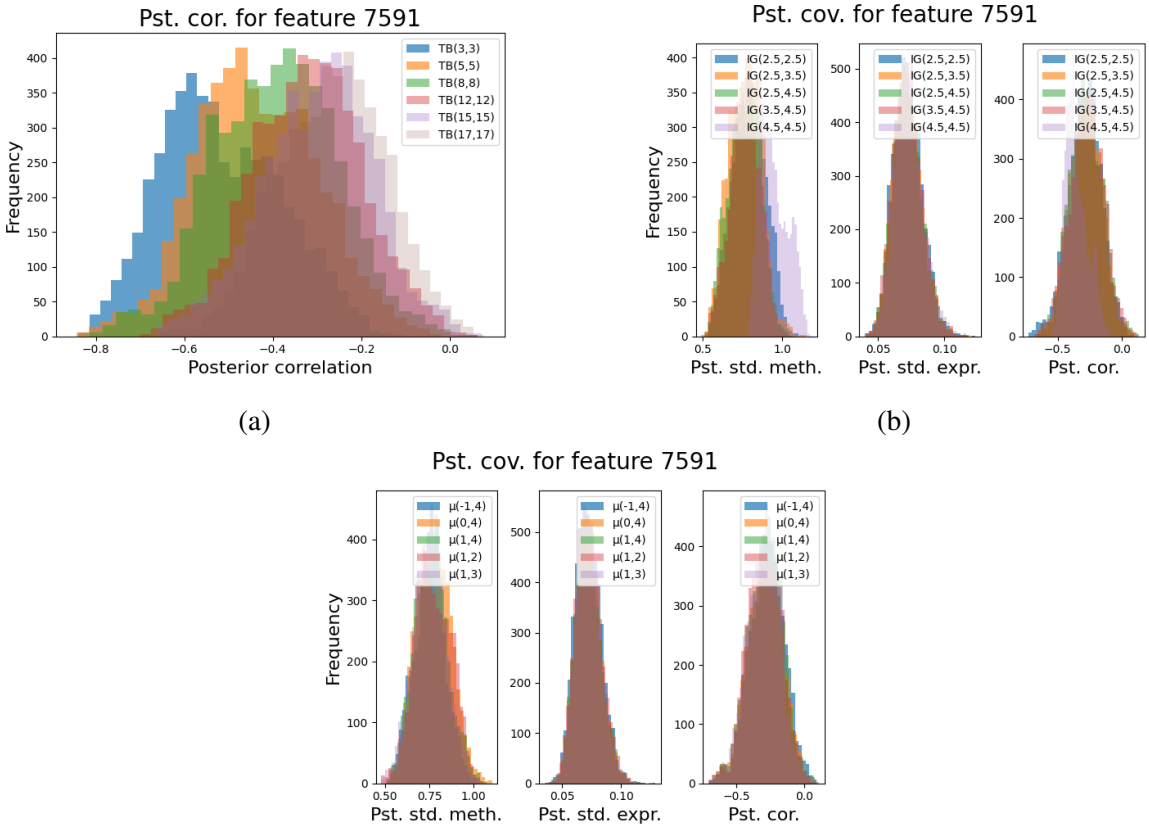
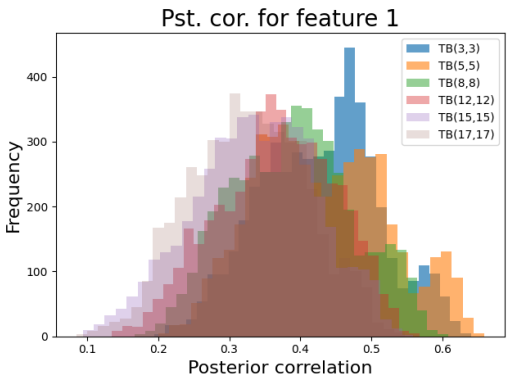


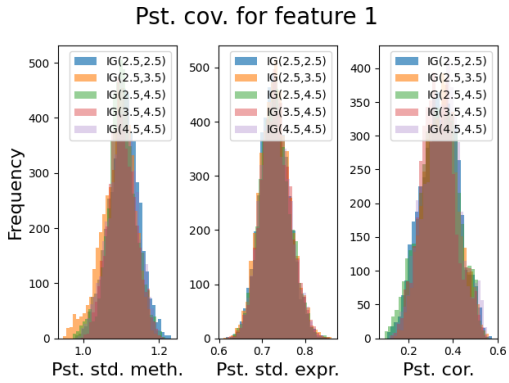
Fig. A.88 **Posterior correlation/covariance for feature** 7591 **with different correlation, standard deviation and mean priors in mESC data.**(A.88a) Posterior correlation distribution for feature 7591 and different correlation. (A.88b) Posterior correlation distribution for feature 7591 and different correlation. (A.88c) Posterior covariance distribution for feature 7591 and different mean priors in mESC data.

A.12.2 Prior Sensitivity for mEBC data

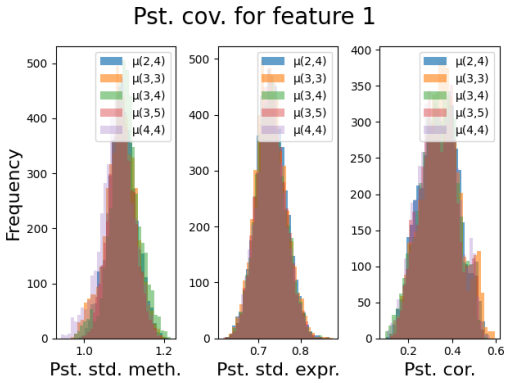
In this subsection we present the prior sensitivity analysis for mEBC data.



(a)

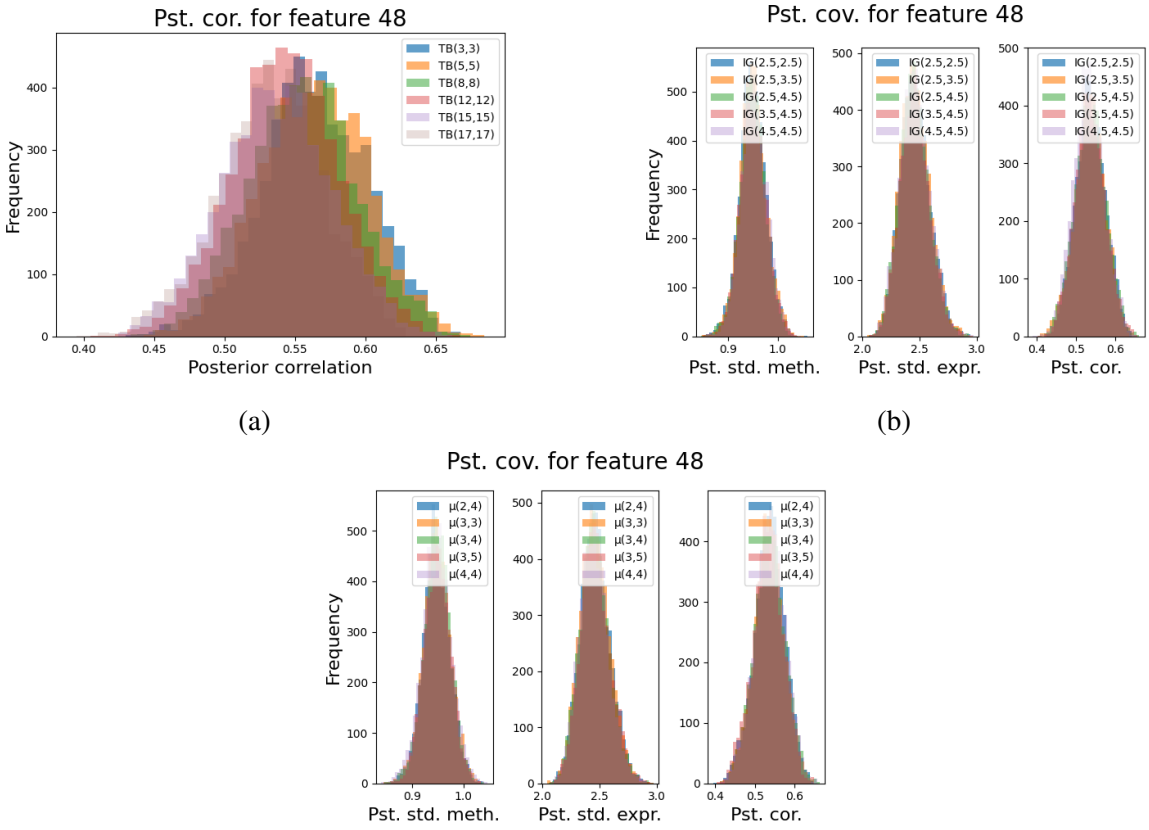


(b)



(c)

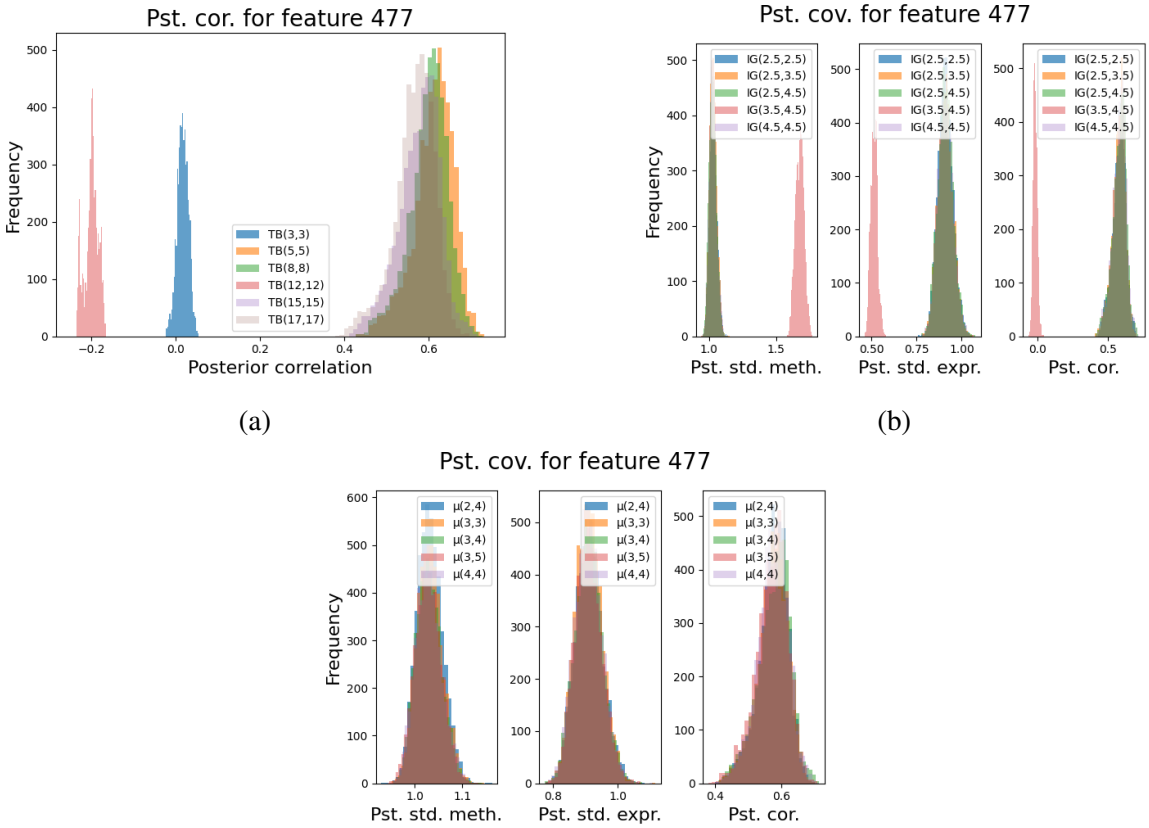
Fig. A.89 **Posterior correlation/covariance for feature** 1 **with different correlation, standard deviation and mean priors in mEBC data.**(A.89a) Posterior correlation distribution for feature 1 and different correlation. (A.89b) Posterior correlation distribution for feature 1 and different correlation. (A.89c) Posterior covariance distribution for feature 1 and different mean priors in mEBC data.



(a) (b)

(c)

Fig. A.90 **Posterior correlation/covariance for feature** 48 **with different correlation, standard deviation and mean priors in mEBC data.**(A.90a) Posterior correlation distribution for feature 48 and different correlation. (A.90b) Posterior correlation distribution for feature 48 and different correlation. (A.90c) Posterior covariance distribution for feature 48 and different mean priors in mEBC data.



(a) (b)

(c)

Fig. A.91 **Posterior correlation/covariance for feature** 477 **with different correlation, standard deviation and mean priors in mEBC data.**(A.91a) Posterior correlation distribution for feature 477 and different correlation. (A.91b) Posterior correlation distribution for feature 477 and different correlation. (A.91c) Posterior covariance distribution for feature 477 and different mean priors in mEBC data.

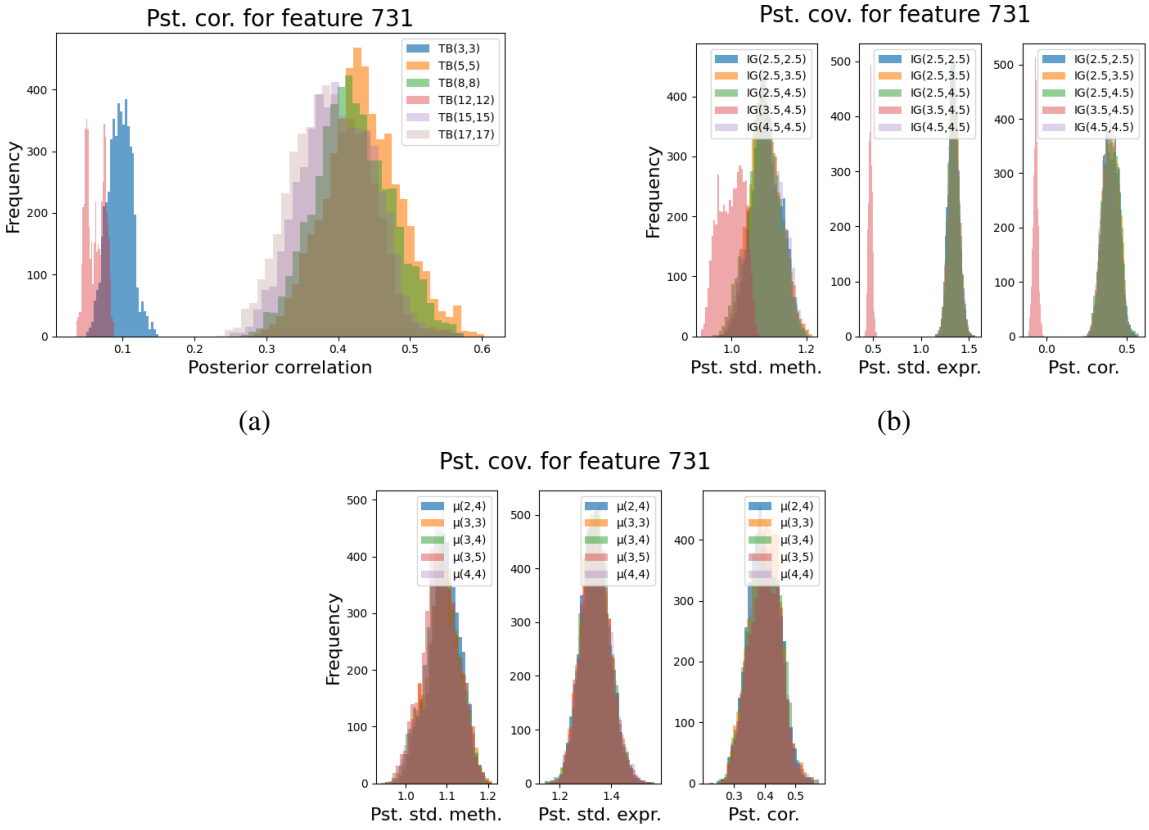
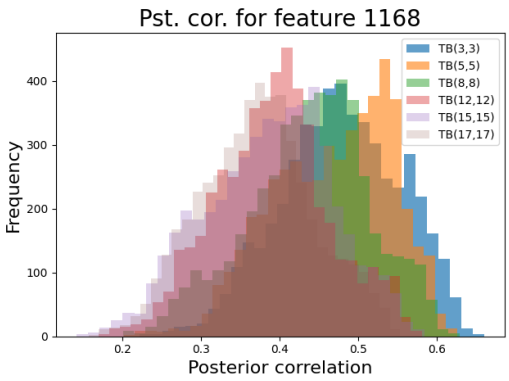
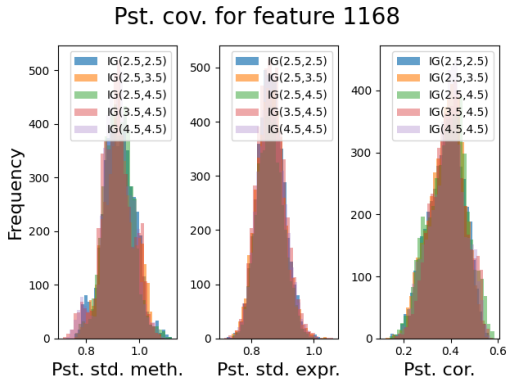


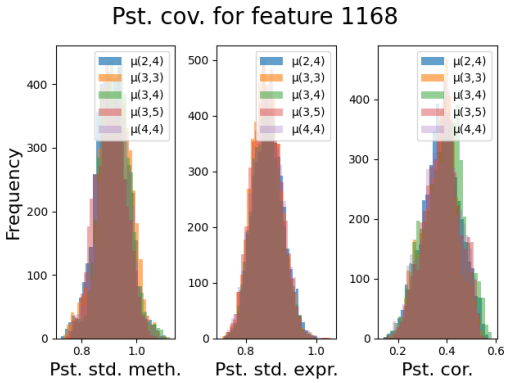
Fig. A.92 **Posterior correlation/covariance for feature** 731 **with different correlation, standard deviation and mean priors in mEBC data.**(A.92a) Posterior correlation distribution for feature 731 and different correlation. (A.92b) Posterior correlation distribution for feature 731 and different correlation. (A.92c) Posterior covariance distribution for feature 731 and different mean priors in mEBC data.



(a)



(b)



(c)

Fig. A.93 **Posterior correlation/covariance for feature** 1168 **with different correlation, standard deviation and mean priors in mEBC data.**(A.93a) Posterior correlation distribution for feature 1168 and different correlation. (A.93b) Posterior correlation distribution for feature 1168 and different correlation. (A.93c) Posterior covariance distribution for feature 1168 and different mean priors in mEBC data.

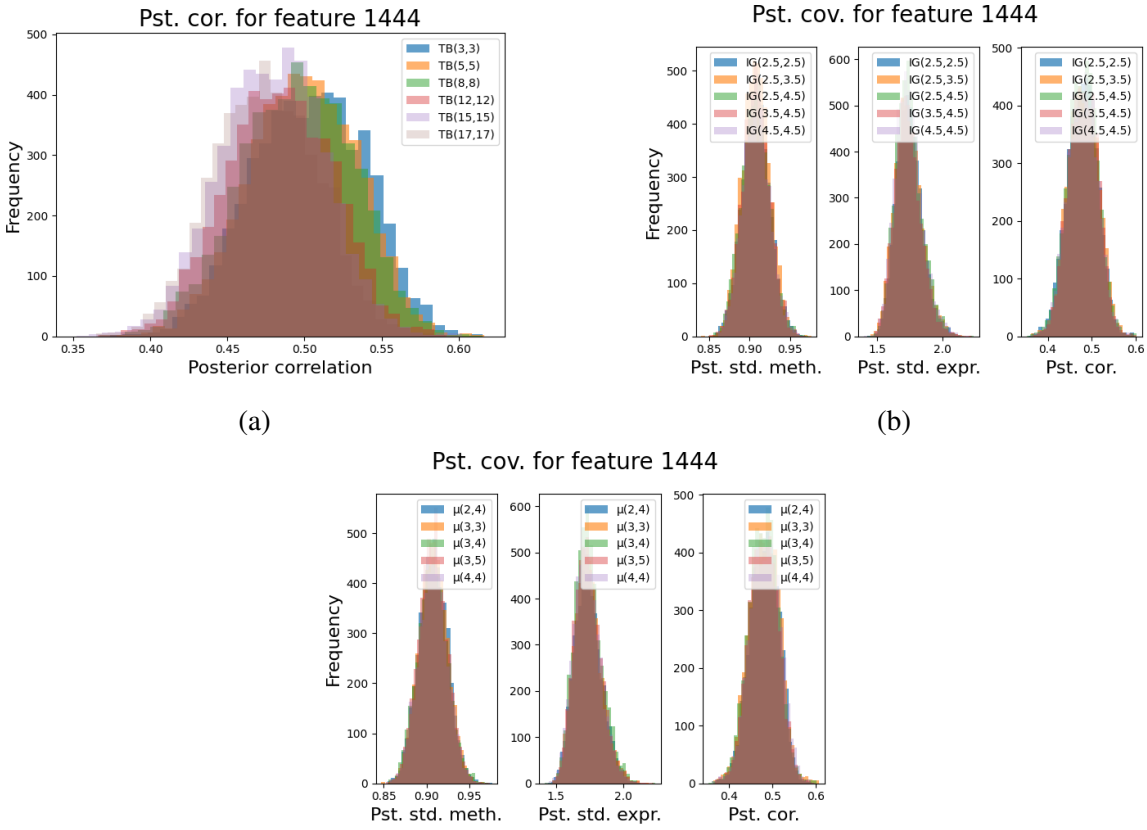
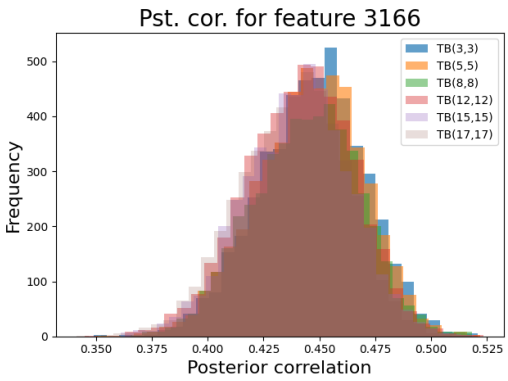
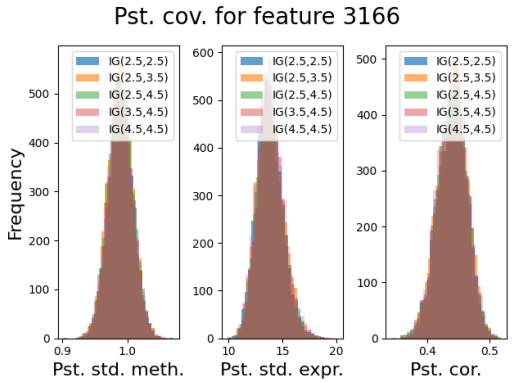


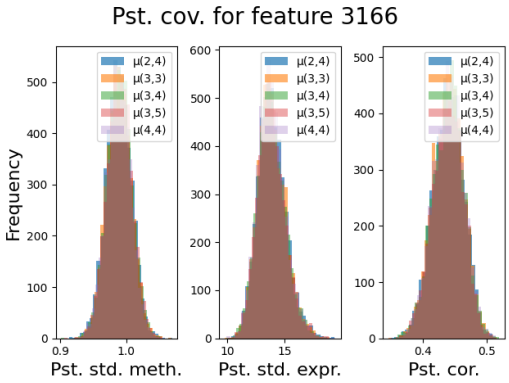
Fig. A.94 **Posterior correlation/covariance for feature** 1444 **with different correlation, standard deviation and mean priors in mEBC data.**(A.94a) Posterior correlation distribution for feature 1444 and different correlation. (A.94b) Posterior correlation distribution for feature 1444 and different correlation. (A.94c) Posterior covariance distribution for feature 1444 and different mean priors in mEBC data.



(a)



(b)



(c)

Fig. A.95 **Posterior correlation/covariance for feature** 3166 **with different correlation, standard deviation and mean priors in mEBC data.**(A.95a) Posterior correlation distribution for feature 3166 and different correlation. (A.95b) Posterior correlation distribution for feature 3166 and different correlation. (A.95c) Posterior covariance distribution for feature 3166 and different mean priors in mEBC data.

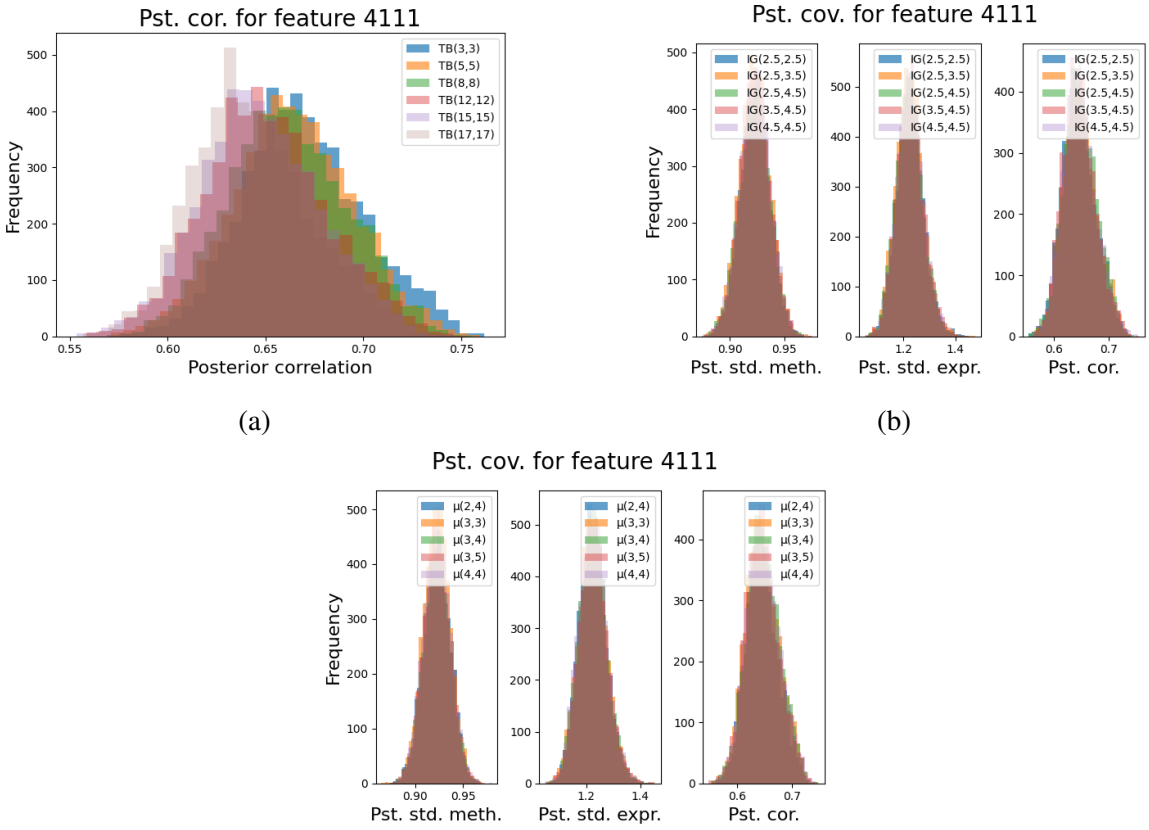
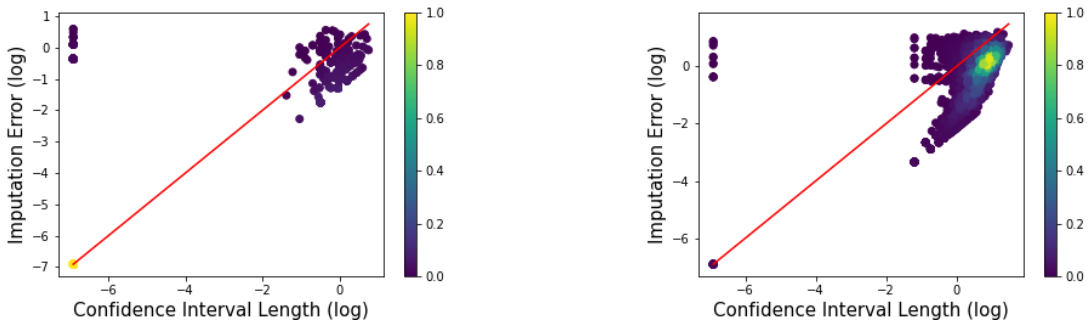


Fig. A.96 **Posterior correlation/covariance for feature** 4111 **with different correlation, standard deviation and mean priors in mEBC data.**(A.96a) Posterior correlation distribution for feature 4111 and different correlation. (A.96b) Posterior correlation distribution for feature 4111 and different correlation. (A.96c) Posterior covariance distribution for feature 4111 and different mean priors in mEBC data.

Appendix B

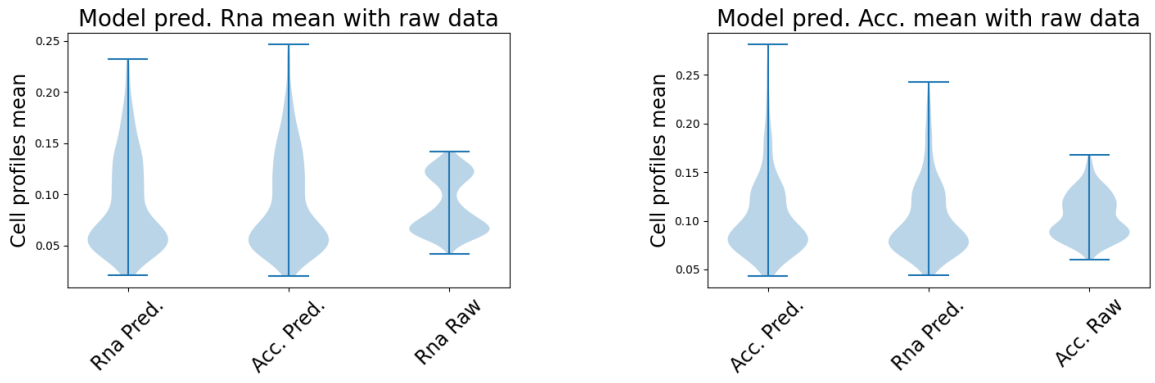
Supplementary information for scMVI

B.1 Imputing missing data



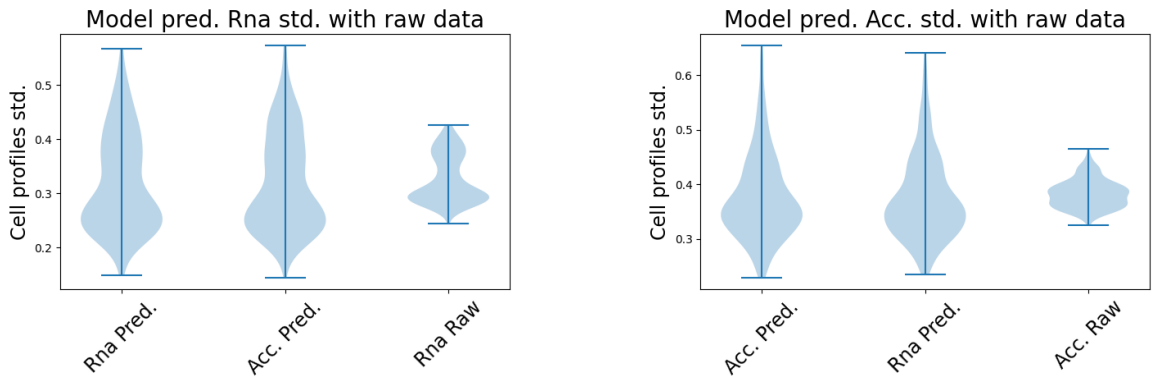
(a) (b)

Fig. B.1 **Imputation statistics for droplet and plate based protocols. Figures used to asses scMVI's imputation quality. Useful comparison between statistics of imputed and ground truth values.** Scatter plots summarizing average euclidean distance between imputed and observed value as a function of 95% confidence interval length for droplet (i.e. mESC_Pj and PBMC) (B.1a) and plate based (i.e. mESC_Ag)(B.1b) protocols.



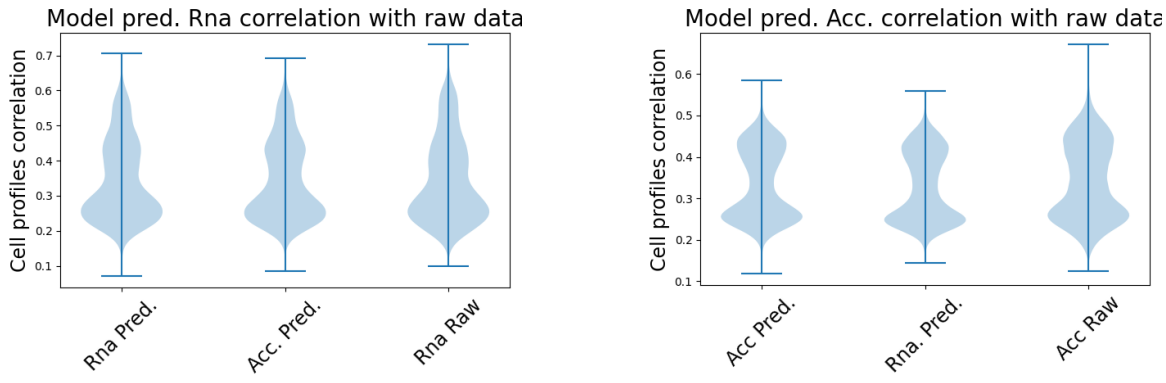
(a) (b)

Fig. B.2 **Violin plots summarizing distributions of imputed and raw profile means for gene expression and chromatin accessibility in mEBC. Useful comparison between statistics of imputed and ground truth values.** (B.2a) Violin plots summarizing distributions of imputed and raw profile means for gene expression in mEBC test cell data. (B.2b) Violin plots summarizing distributions of imputed and raw profile means for chromatin accessibility in mEBC test cell data.



(a) (b)

Fig. B.3 **Violin plots summarizing distributions of imputed and raw profile standard deviations for gene expression and chromatin accessibility in mEBC. Useful comparison between statistics of imputed and ground truth values.** (B.3a) Violin plots summarizing distributions of imputed and raw profile standard deviations for gene expression in mEBC test cell data. (B.3b) Violin plots summarizing distributions of imputed and raw profile standard deviations for chromatin accessibility in mEBC test cell data.



(a) (b)

Fig. B.4 **Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression and chromatin accessibility in mEBC. Useful comparison between statistics of imputed and ground truth values.** (B.4a) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for gene expression in mEBC test cell data. (B.4b) Violin plots summarizing Pearson correlation distributions between imputed and raw profiles for chromatin accessibility in mEBC test cell data.

Imputing Chromatin Accessibility from Gene Expression (PBMC)

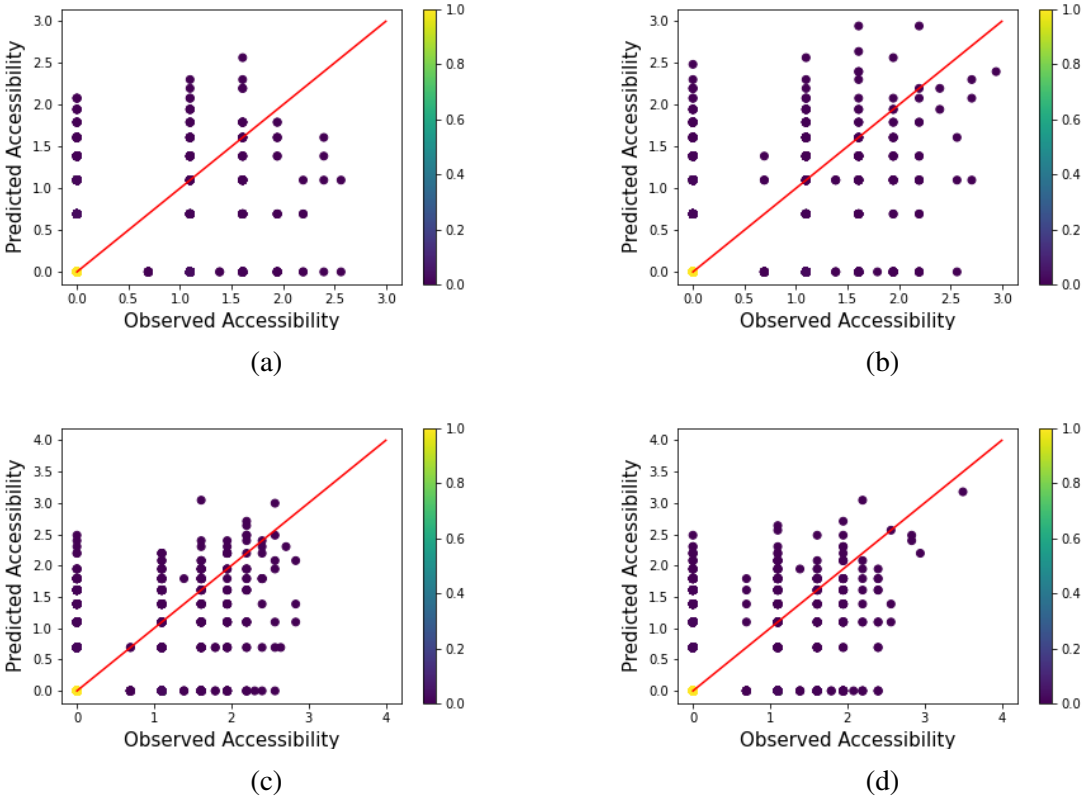


Fig. B.5 **Imputing chromatin accessibility with gene expression for PBMC data. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed chromatin accessibility from gene expression as a function of observed chromatin accessibility for cells from PBMC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match.

Imputing Gene Expression from Chromatin Accessibility (PBMC)

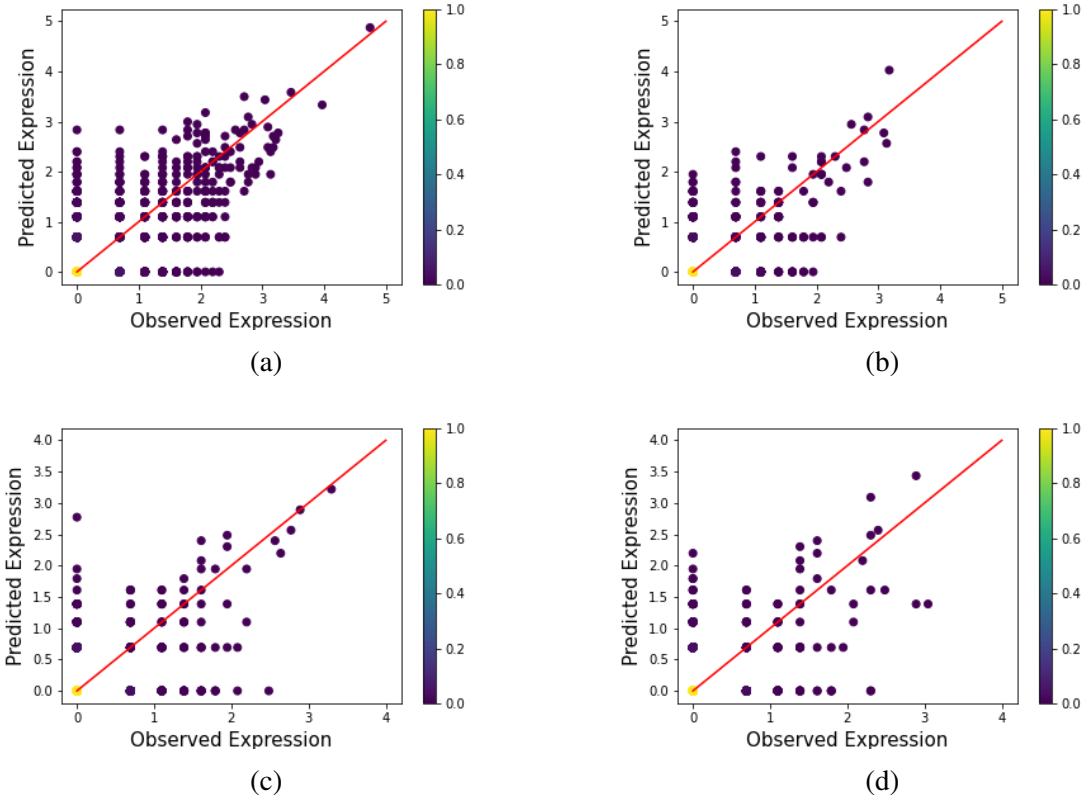


Fig. B.6 **Imputing expression with chromatin accessibility for PBMC. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed expression from chromatin accessibility as a function of observed expression for cells from PBMC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match.

Imputing Gene Expression from Chromatin Accessibility (mESC_Ag)

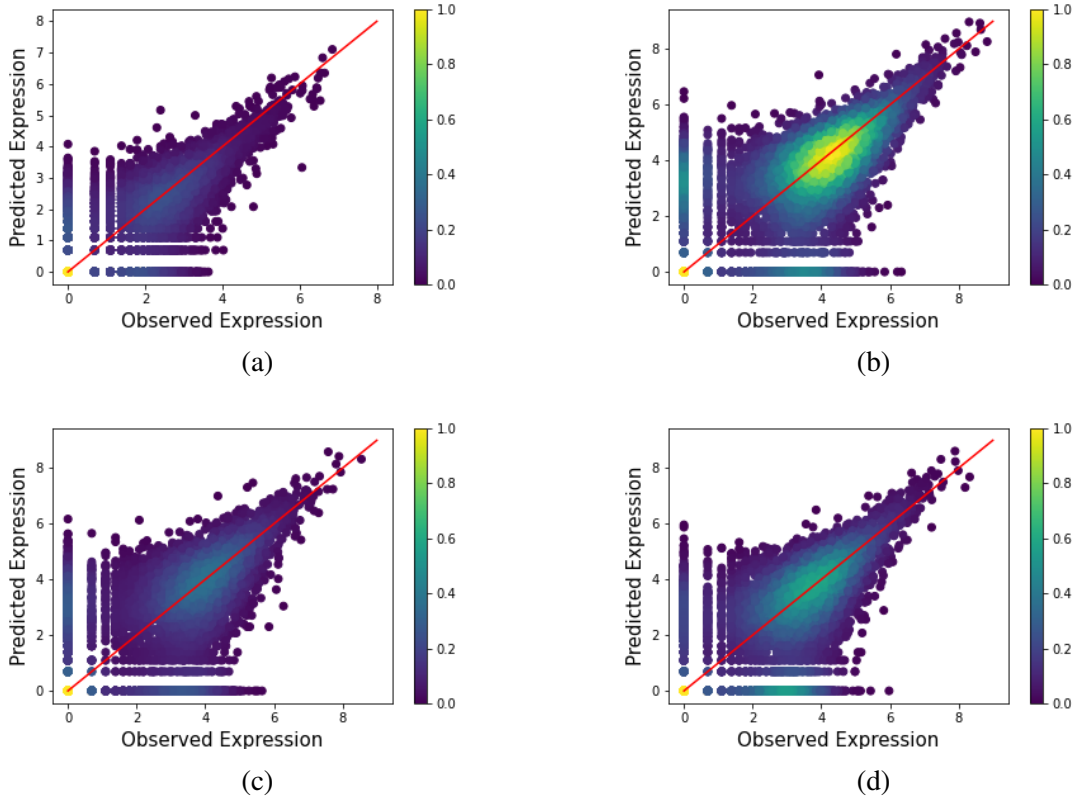


Fig. B.7 **Imputing expression with chromatin accessibility for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed expression from chromatin accessibility as a function of observed gene expression for cells from mESC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match.

Imputing Gene Expression from DNA Methylation (mESC_Ag)

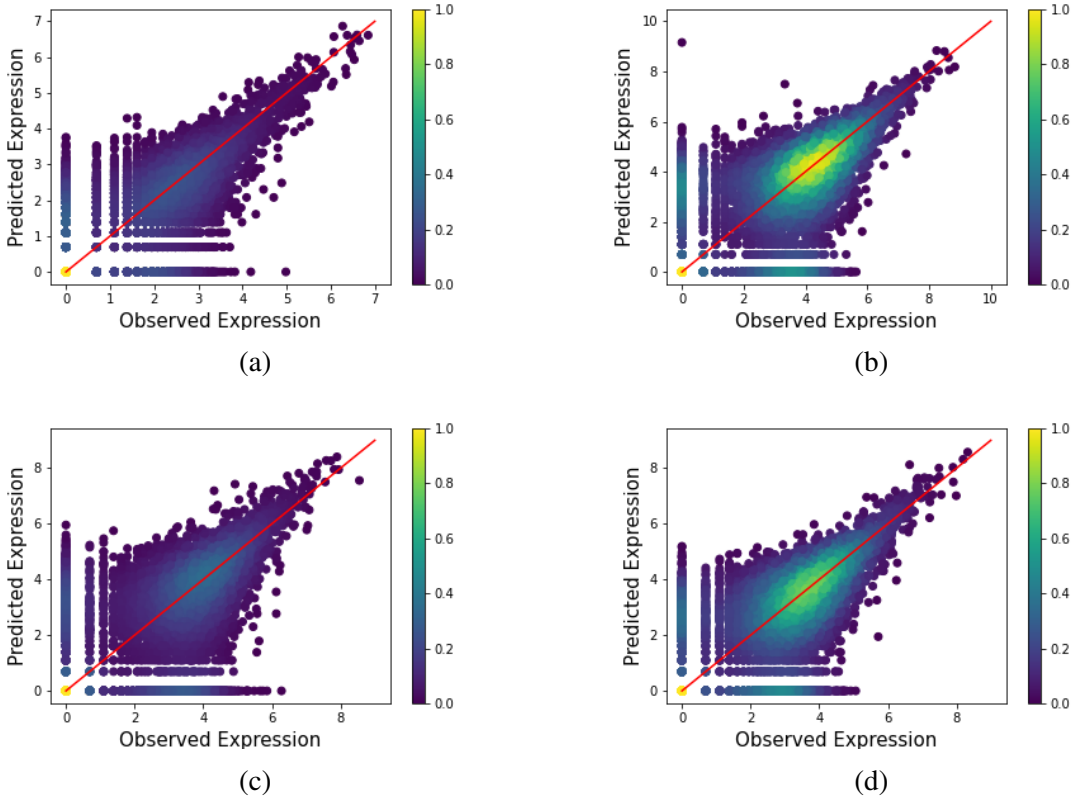


Fig. B.8 **Imputing expression with DNA methylation for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed expression from DNA methylation as a function of observed expression for cells from mESC test set. Each dot represents a gene and it is color-coded by frequency of point occurrence. The red line is $y = x$ and corresponds to perfect match.

Imputing Chromatin Accessibility from Gene Expression (mESC_Ag)

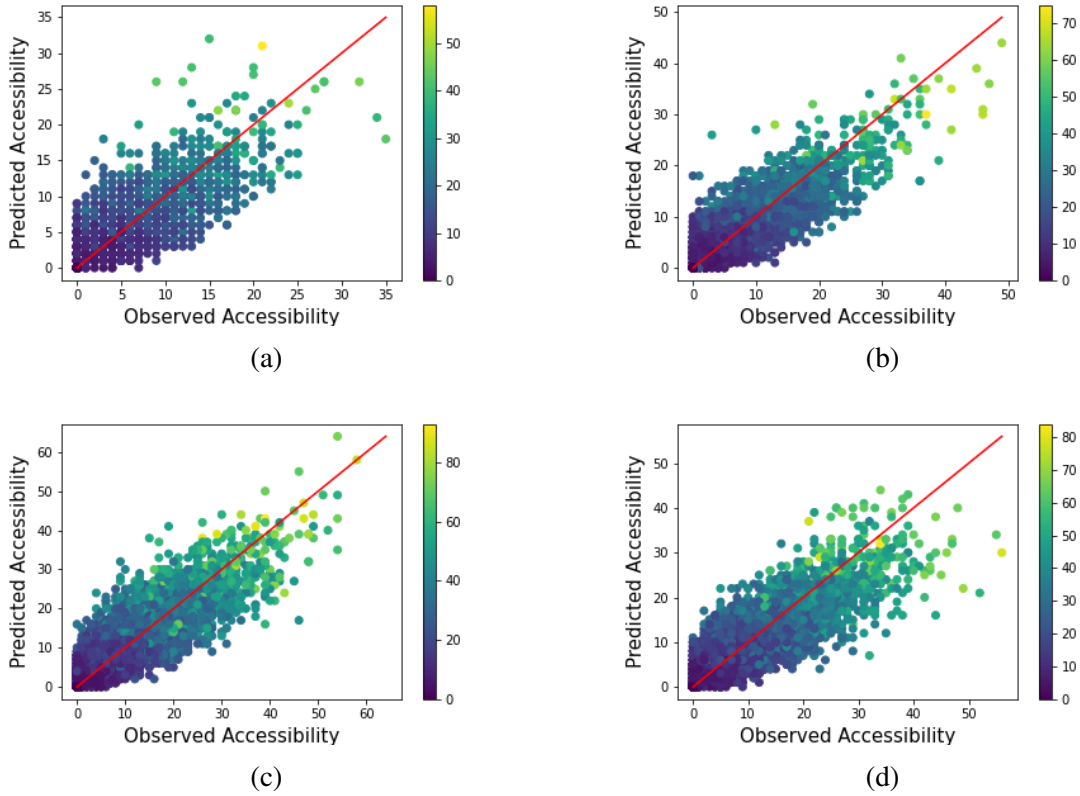


Fig. B.9 **Imputing chromatin accessibility with gene expression for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed chromatin accessibility from gene expression as a function of observed chromatin accessibility for cells from mESC test set. Each dot represents a gene and it is color-coded by GpC coverage. The red line is $y = x$ and corresponds to perfect match.

Imputing Chromatin Accessibility from DNA Methylation (mESC_Ag)

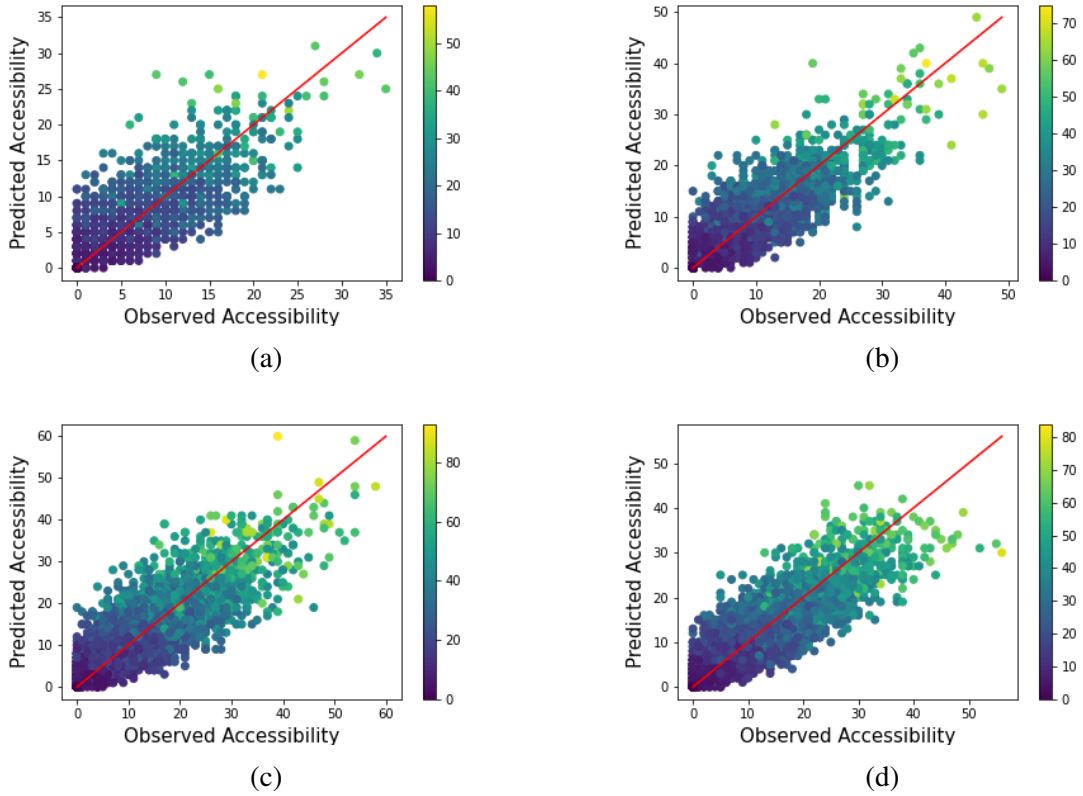


Fig. B.10 **Imputing chromatin accessibility with DNA methylation for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed chromatin accessibility from DNA methylation as a function of observed chromatin accessibility for cells from mESC test set. Each dot represents a gene and it is color-coded by GpC coverage. The red line is $y = x$ and corresponds to perfect match.

Imputing DNA Methylation from Gene Expression (mESC_Ag)

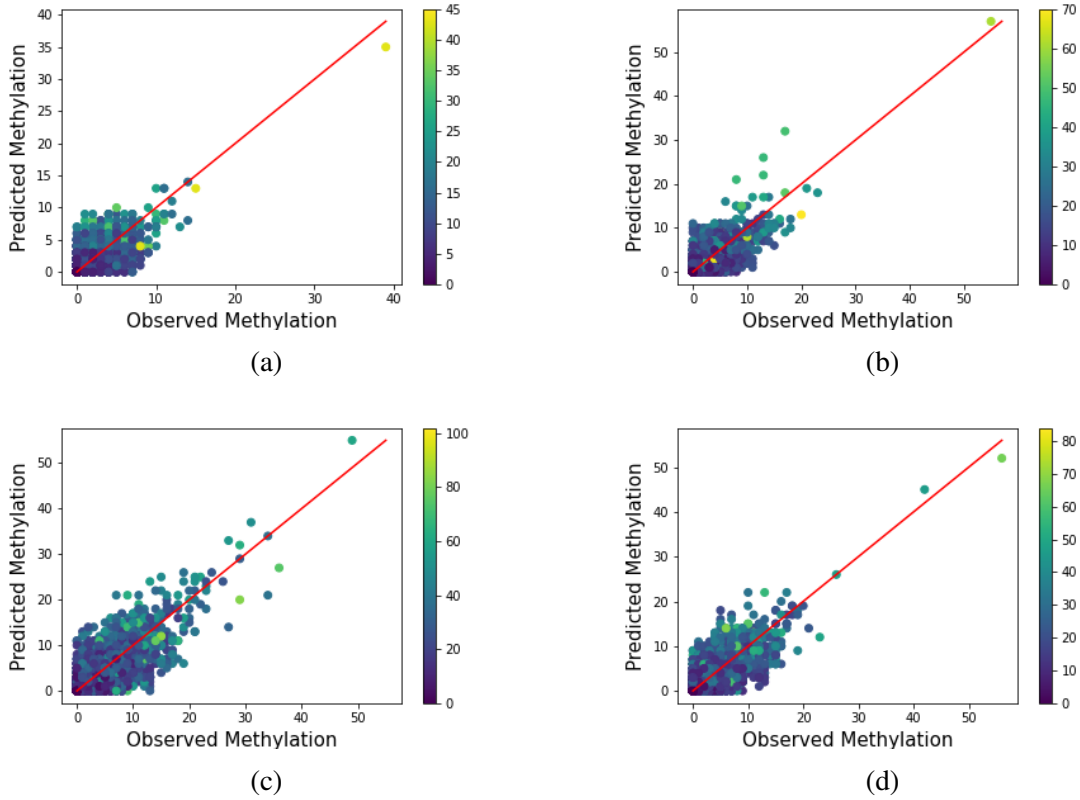


Fig. B.11 **Imputing DNA methylation with gene expression for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed DNA methylation from gene expression as a function of observed DNA methylation for cells from mESC test set. Each dot represents a gene and it is color-coded by CpG coverage. The red line is $y = x$ and corresponds to perfect match.

Imputing DNA Methylation from Accessibility (mESC_Ag)

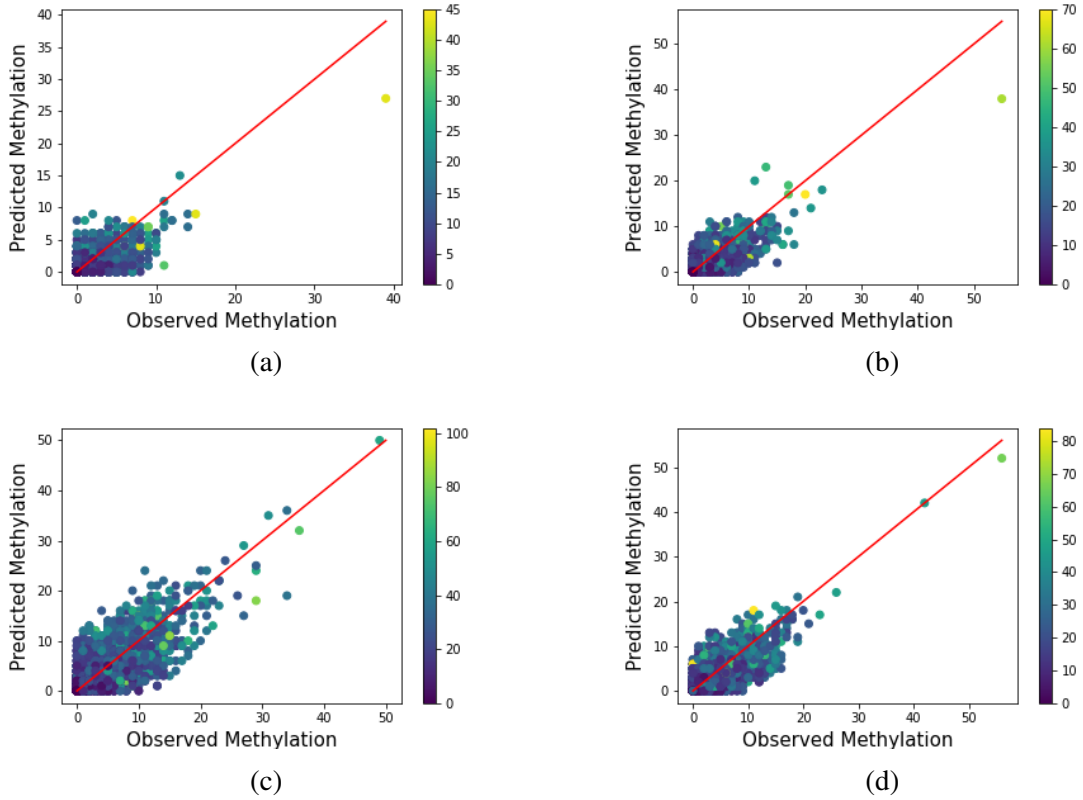
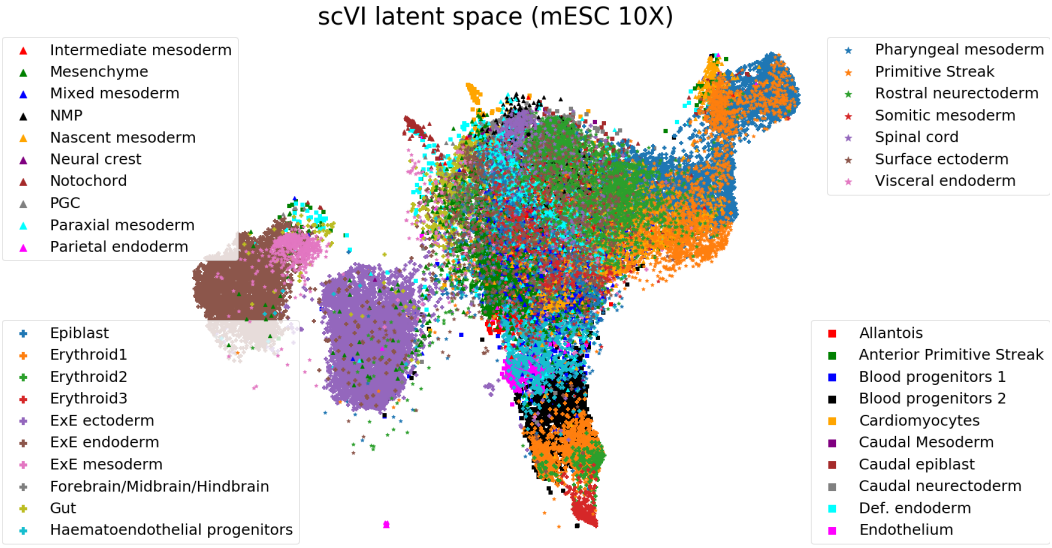
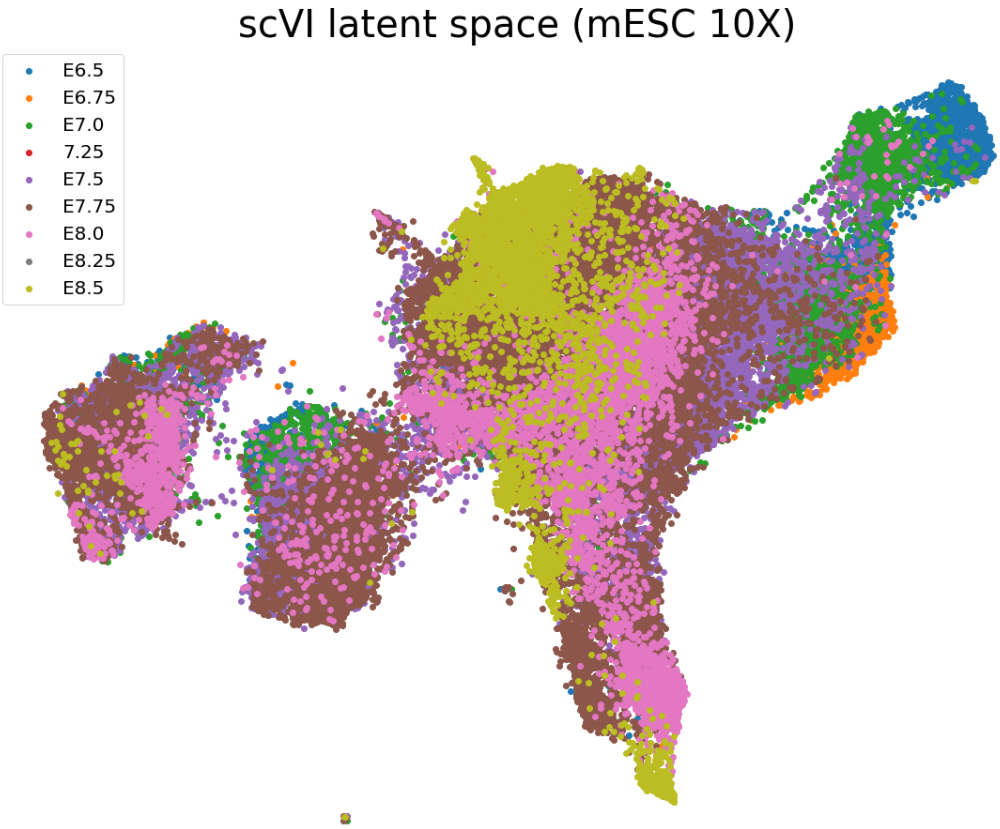


Fig. B.12 **Imputing DNA methylation with chromatin accessibility for mESC data sequenced with scNMT protocol. Figures used to compare scMVI's imputed profiles with ground truth for several datasets.** Scatter plots depicting imputed DNA methylation from chromatin accessibility as a function of observed DNA methylation for cells from mESC test set. Each dot represents a gene and it is color-coded by CpG coverage. The red line is $y = x$ and corresponds to perfect match.

B.2 Comparison with related methods

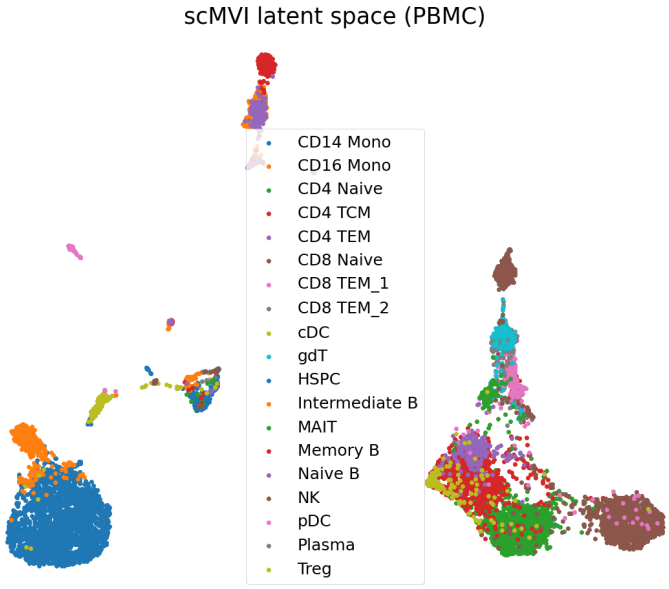


(a)

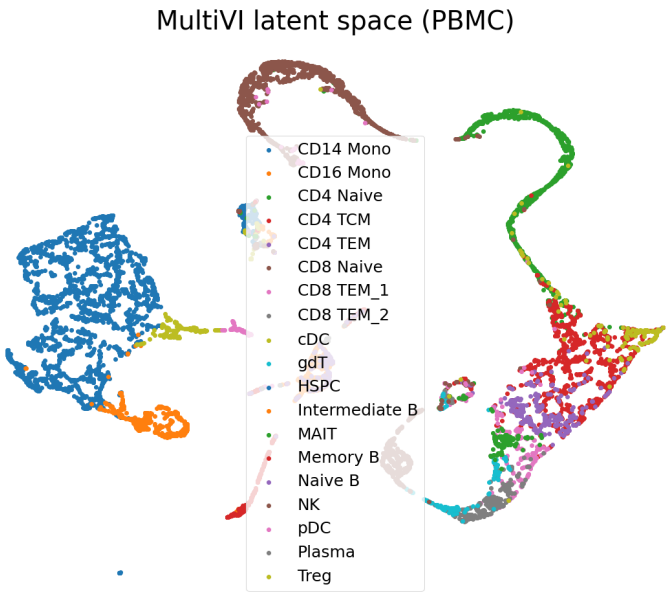


(b)

Fig. B.13 **Summary of scVI's latent space on mESC_Pj color-coded by lineage and developmental stage. Figures demonstrating that scMVI learns latent representations correlated with cell phenotype or developmental state.** 2D UMAP projection of scVI's latent space, trained on mESC_Pj. Each dot represents a cell and is color-coded by lineage (B.13a) and developmental stage (B.13b).

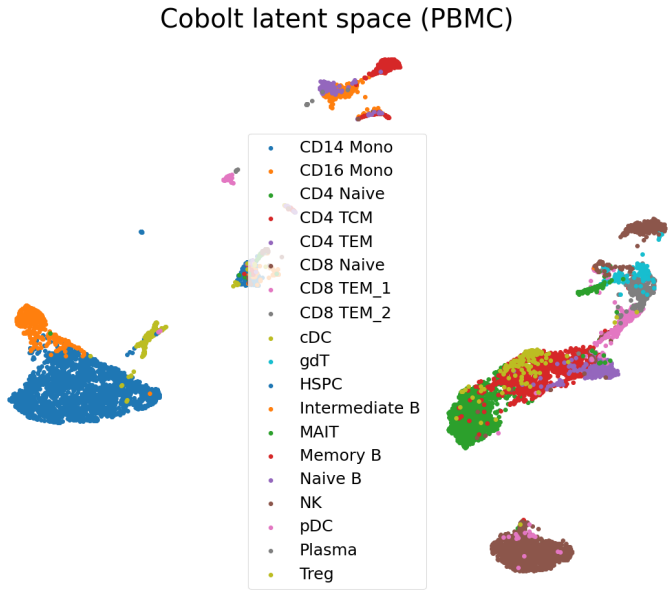


(a)

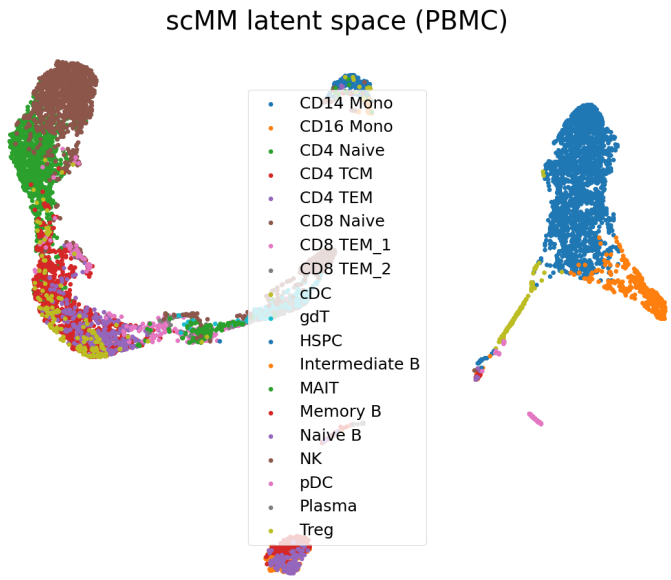


(b)

Fig. B.14 **UMAP projection of scMVI and MultiVI latent manifolds on paired PBMC data. Figures used to qualitatively compare latent representations from related models to scMVI.** Two dimensional depiction of scMVI's (B.14a) and MultiVI's (B.14b) latent space trained on paired PBMC data. Each dot represents a cell and is color-coded by cell type. Since scMVI estimates one latent representation per molecular layer, Figure B.14a is generated by averaging across them before visualising with UMAP.

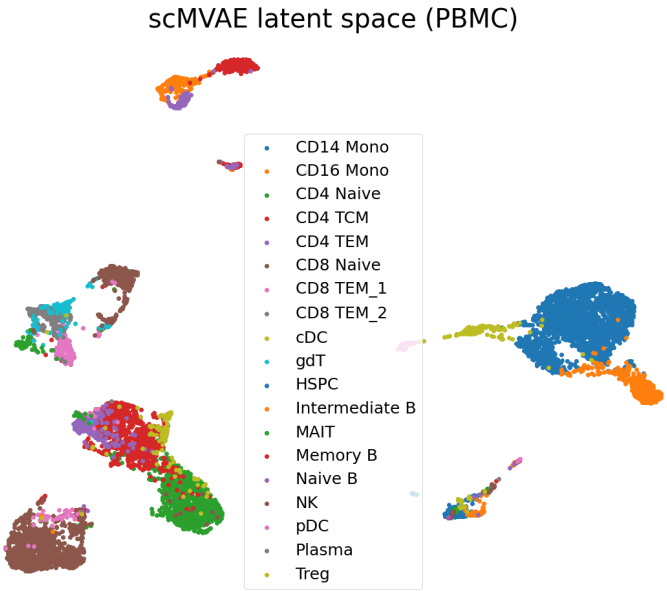


(a)



(b)

Fig. B.15 **UMAP projection of Cobolt and scMM latent manifolds on paired PBMC data. Figures used to qualitatively compare latent representations from related models to scMVI.** Two dimensional depiction of Cobolt's (B.15a) and scMM's (B.15b) latent space trained on paired PBMC data. Each dot represents a cell and is color-coded by cell type.



(a)

Fig. B.16 **UMAP projection of scMVAE and scVI latent manifolds on paired PBMC data. Figures used to qualitatively compare latent representations from related models to scMVI.** Two dimensional depiction of scMVAE's (B.16a) latent space trained on paired PBMC data. Each dot represents a cell and is color-coded by cell type.

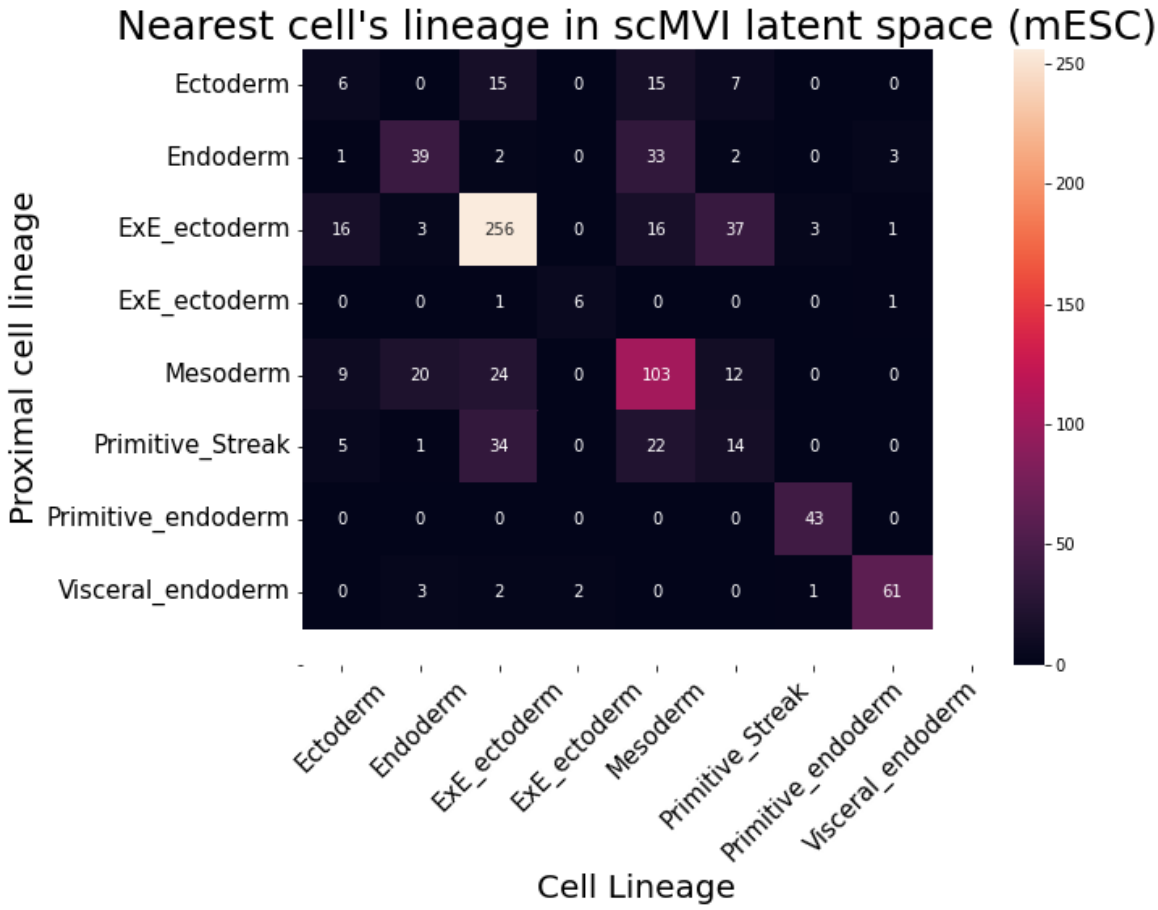
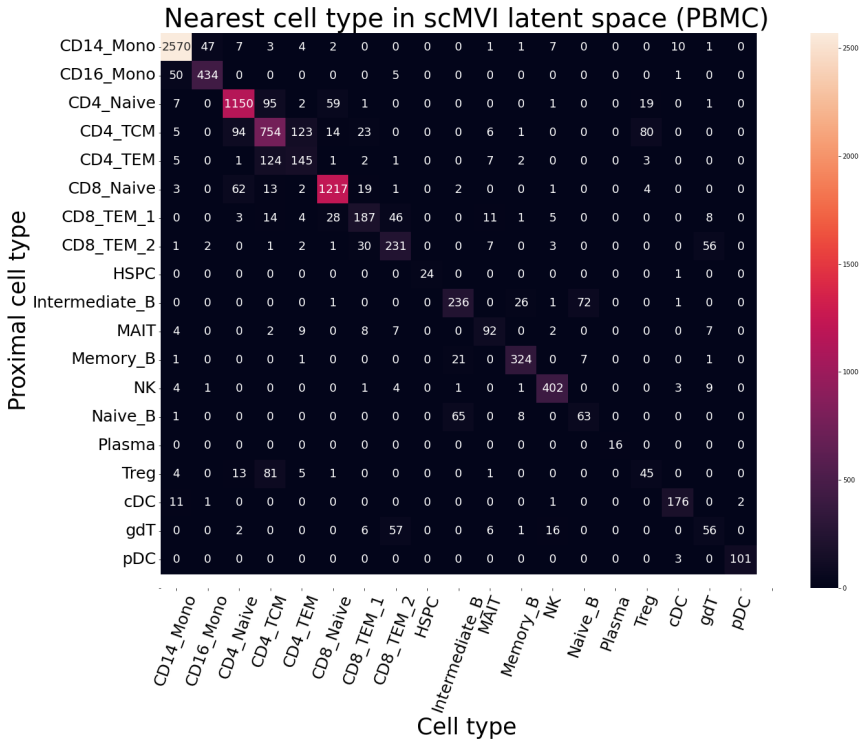
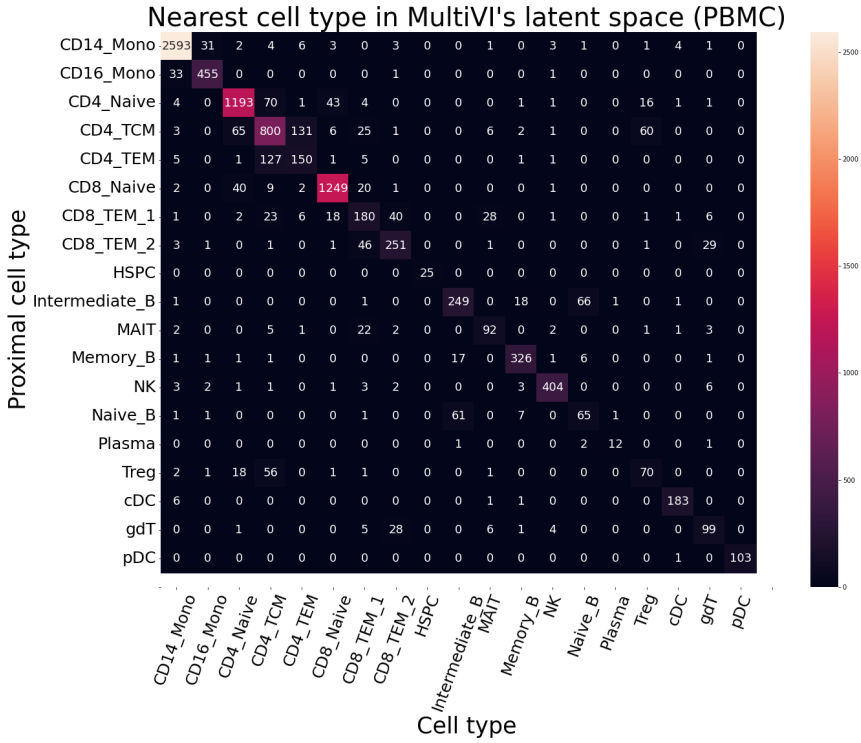


Fig. B.17 **Heatmap summarizing proximal cell lineage in scMVI's latent space for each cell on paired mESC_Ag. Heatmap used to demonstrate which cell type would be picked in dataset integration with scMVI.**. For each latent representation in mESC_Ag, estimated with paired data, we track proximal cell lineage to understand how scMVI's latent space is structured. Since scMVI estimates one latent representation per molecular layer, the heatmap is generated after averaging across them.

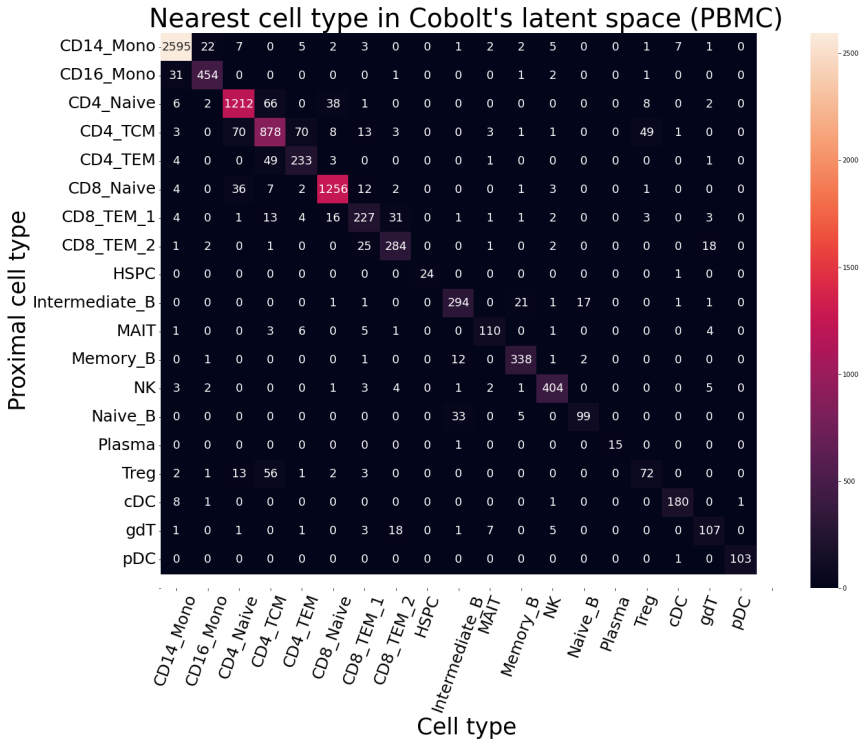


(a)

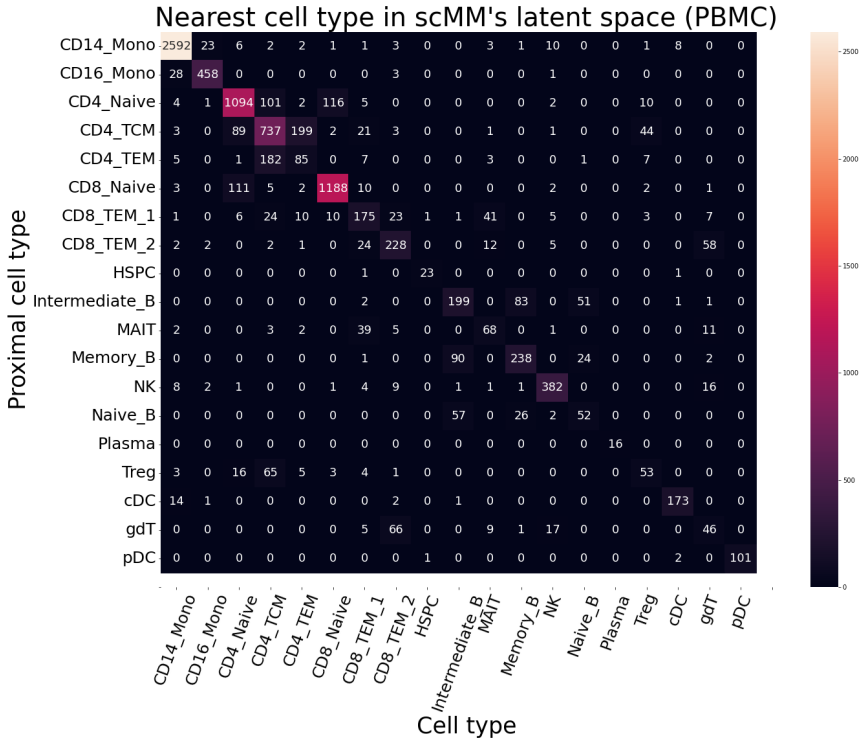


(b)

Fig. B.18 **Heatmap summarizing proximal cell type in scMVI's and MultiVI's latent space for each cell on paired PBMC data. Heatmaps used to compare which cell type would be picked in dataset integration with alternative models.** For each PBMC latent representation we track proximal cell types to understand how scMVI's (B.14a) and MultiVI's (B.18b) latent space is structured. Since scMVI estimates one latent representation per molecular layer, Heatmap B.18a is generated after averaging across them.

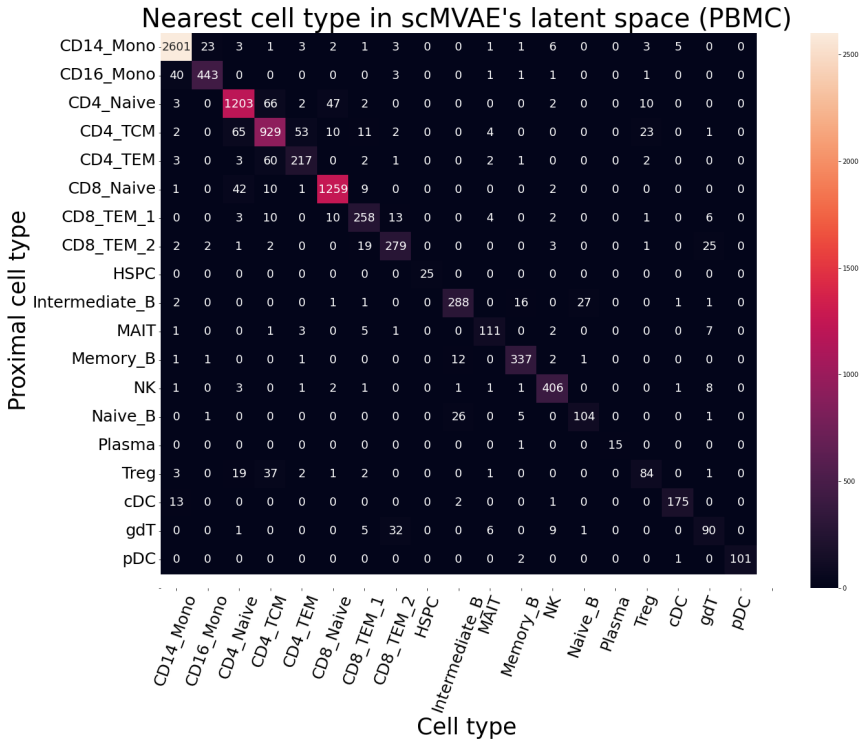


(a)



(b)

Fig. B.19 **Heatmap summarizing proximal cell type in Cobolt's and scMM's latent space for each cell on paired PBMC data. Heatmaps used to compare which cell type would be picked in dataset integration with alternative models.** For each PBMC latent representation we track proximal cell types to understand how Cobolt's(B.19a) and scMM's(B.19b) latent space is structured.



(a)

Fig. B.20 **Heatmap summarizing proximal cell type in scMVAE's latent space for each cell on paired PBMC data. Heatmap used to compare which cell type would be picked in dataset integration with alternative models.** For each PBMC latent representation we track proximal cell types to understand how scMVAE's(B.20a) latent space is structured.

B.3 Exploratory Analysis

Differential directions analysis. Assuming an average cell representation $\vec{s}_t$ for cells with label t, generated by averaging across molecular latent representation of cells in set I_t. Depending data nature, label t denotes time, subtype or lineage. So:

$$\vec{s}_t = \frac{1}{|J_t|} \sum_{j \in J_t} \frac{z_{ai} + z_{mi} + z_{ri}}{3} \tag{B.1}$$

Between different subtypes or consecutive developmental stages E_t-E_{t+1} the differential direction $\Delta\vec{s}_t$ is estimated as follows:

$$\Delta\vec{s}_t = \vec{s}_{t+1} - \vec{s}_t \tag{B.2}$$

To understand how differential direction impacts each modality, we create DDLFC by projecting $\Delta\vec{s}_t$ to the observed space. DDLFC can be understood as a heuristic measuring how latent space perturbations materialize on average in observed space. Hence for DNA methylation, chromatin accessibility and gene expression we get the following expression for DDLFC.

$$D_m^t = f_m\left(\Delta\vec{s}_t\right) - f_m\left(\vec{0}\right) \tag{B.3}$$

$$D_a^t = f_a\left(\Delta\vec{s}_t\right) - f_a\left(\vec{0}\right) \tag{B.4}$$

For gene expression the only meaningful parameter to consider is mean normalized gene expression.

$$D_r^t = f_r^m\left(\Delta\vec{s}_t\right) - f_r^m\left(\vec{0}\right) \tag{B.5}$$

DDLFC vs LFC_NN comparison

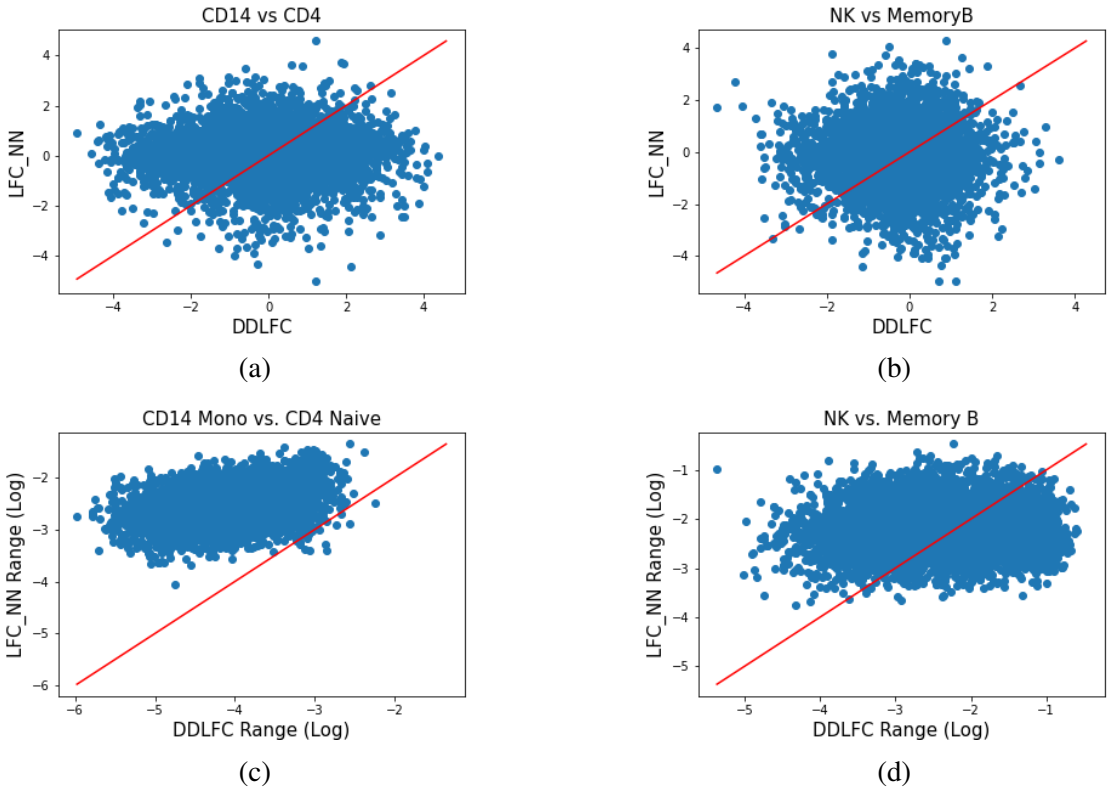


Fig. B.21 **Median DDLFC and LFC_NN comparison for PBMC. Scatter plots used to asses how robust is DDLFC compared to LFC under various scenarios for specific cell types.** Scatter plots summarizing average and range of median DDLFC against average and range of median LFC_NN in pairs where it is less (CD14 vs CD4)(B.21a,B.21c) and most (Natural kille vs Memory B)(B.21b,B.21d) impacted from subsampling and label imbalance, for PBMC. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range.

DDLFC vs DT comparison

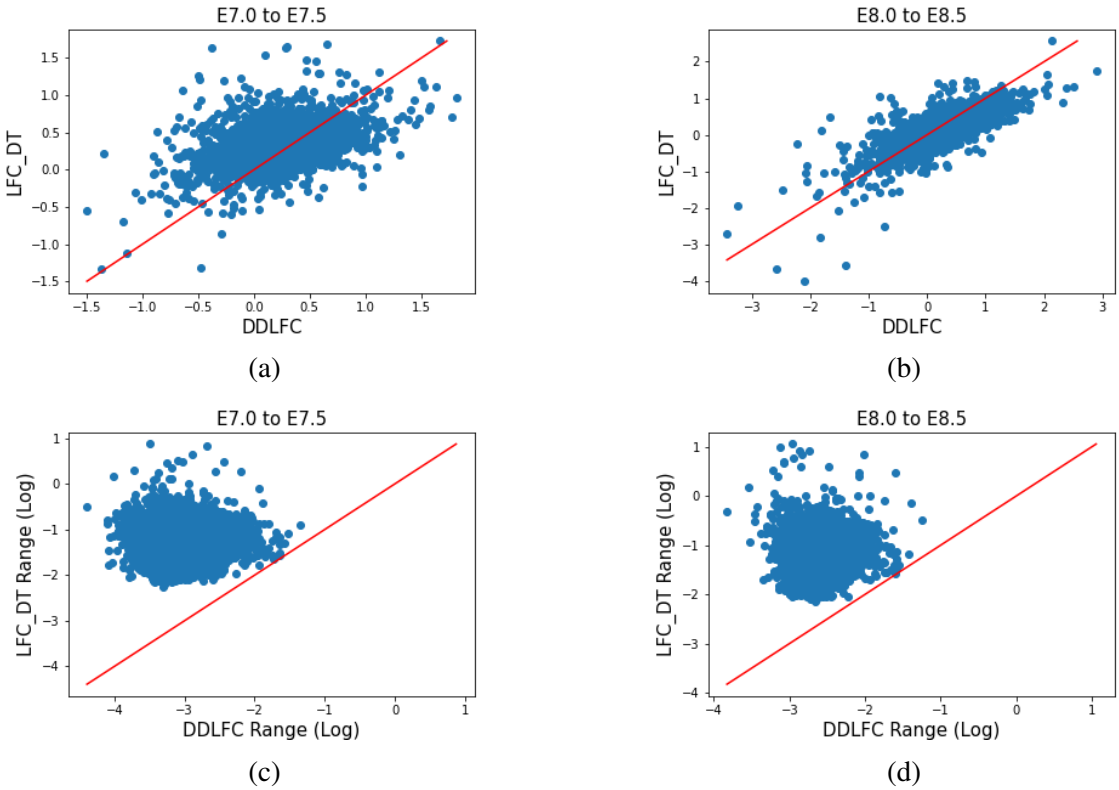


Fig. B.22 **Median DDLFC and LFC_DT comparison for mESC_Pj. Scatter plots used to asses how robust is DDLFC compared to differential expression analysis under various scenarios for specific developmental steps.** Scatter plots summarizing average and range of median DDLFC against LFC_DT(and its range) in label pairs where it is less (Embryonic days $E7.0$ vs $E7.5$)(B.22a,B.22c) and most (Embryonic days $E8.0$ vs $E8.5$)(B.22b,B.22d) impacted from subsampling and label imbalance, for mESC_Pj. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range.

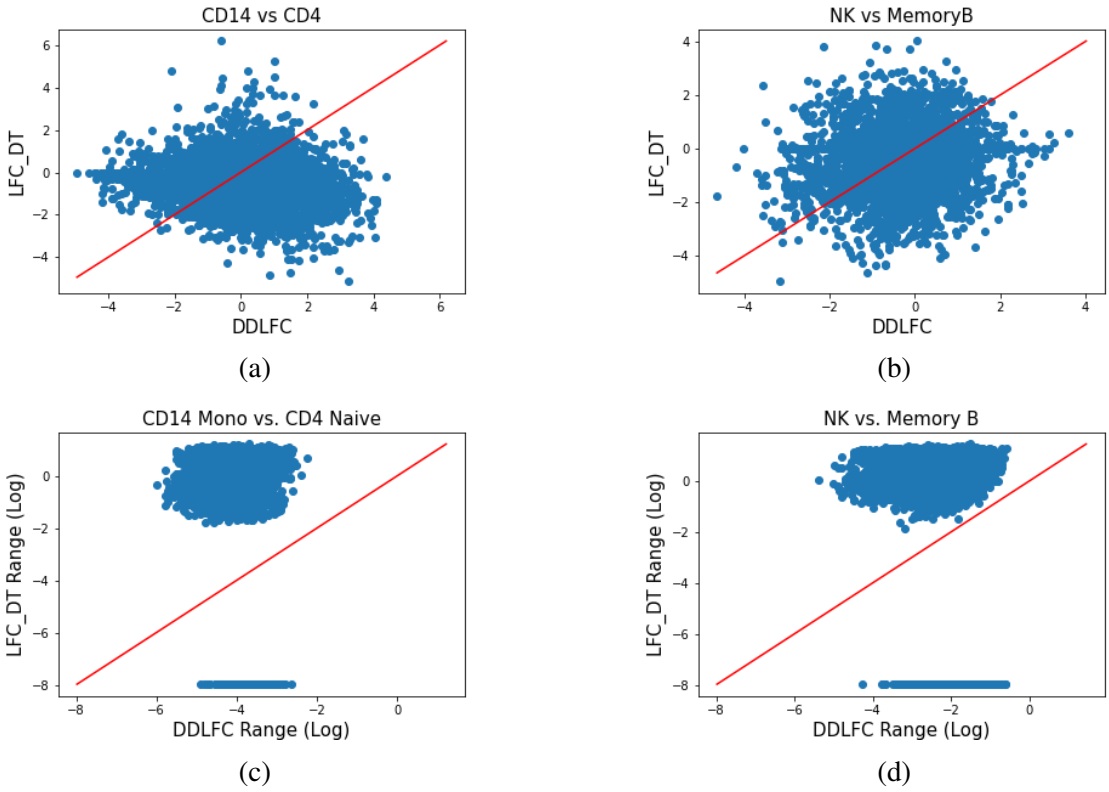


Fig. B.23 **Median DDLFC and LFC_DT comparison for PBMC. Scatter plots used to asses how robust is DDLFC compared to differential expression analysis under various scenarios for specific cell types.** Scatter plots summarizing average and range of median DDLFC against LFC_DT(and its range) in labels where it is less (CD14 vs CD4)(B.23a,B.23c) and most (Naturall Killer vs Memory B)(B.23b,B.23d) impacted from subsampling and label imbalance, for PBMC. Each dot represents a genomic region. The red line depicts perfect match between two quantities' range.

Gene Set Enrichment Analysis

Current Gene List: scMVI_lw_0
Current Background: allscMVI
501 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

12 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	positive regulation of neuron projection development	RT		12	2.4	2.3E-2	2.1
☐	GOTERM_BP_DIRECT	immune system process	RT		17	3.4	1.6E-2	1.9
☐	GOTERM_BP_DIRECT	protein import into nucleus	RT		10	2.0	1.3E-1	1.7
☐	GOTERM_BP_DIRECT	brain development	RT		15	3.0	7.5E-2	1.6
☐	GOTERM_BP_DIRECT	protein autophosphorylation	RT		11	2.2	1.7E-1	1.5
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		21	4.2	7.4E-2	1.5
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		23	4.6	6.4E-2	1.5
☐	GOTERM_BP_DIRECT	regulation of apoptotic process	RT		11	2.2	2.3E-1	1.4
☐	GOTERM_BP_DIRECT	innate immune response	RT		12	2.4	2.3E-1	1.4
☐	GOTERM_BP_DIRECT	heart development	RT		12	2.4	2.5E-1	1.4
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		26	5.2	1.1E-1	1.3
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		30	6.0	2.7E-1	1.2

Fig. B.24 **GSEA analysis on** 500 **genomic regions with the strongest negative DDLFC between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_up_0
Current Background: allscMVI
500 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

10 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	transcription, DNA-templated	RT		10	2.0	8.3E-2	1.9
☐	GOTERM_BP_DIRECT	mitochondrial translation	RT		10	2.0	1.2E-1	1.7
☐	GOTERM_BP_DIRECT	cell differentiation	RT		34	6.8	1.1E-2	1.5
☐	GOTERM_BP_DIRECT	cell adhesion	RT		13	2.6	1.4E-1	1.5
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		19	3.8	7.6E-2	1.5
☐	GOTERM_BP_DIRECT	mRNA splicing, via spliceosome	RT		12	2.4	2.4E-1	1.4
☐	GOTERM_BP_DIRECT	protein polyubiquitination	RT		11	2.2	2.7E-1	1.4
☐	GOTERM_BP_DIRECT	RNA splicing	RT		19	3.8	2.4E-1	1.3
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		27	5.4	1.7E-1	1.3
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		44	8.8	2.1E-1	1.2

Fig. B.25 **GSEA analysis on** 500 **genomic regions with the strongest positive DDLFC between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scVI_lw_5
Current Background: allscMVI
500 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

22 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	neural tube closure	RT		12	2.4	1.7E-2	2.2
☐	GOTERM_BP_DIRECT	negative regulation of cell migration	RT		10	2.0	4.3E-2	2.1
☐	GOTERM_BP_DIRECT	positive regulation of neuron projection development	RT		10	2.0	1.1E-1	1.8
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		22	4.4	1.5E-2	1.7
☐	GOTERM_BP_DIRECT	chromatin organization	RT		23	4.6	3.1E-2	1.6
☐	GOTERM_BP_DIRECT	chromatin remodeling	RT		12	2.4	1.6E-1	1.5
☐	GOTERM_BP_DIRECT	DNA replication	RT		10	2.0	2.2E-1	1.5
☐	GOTERM_BP_DIRECT	immune system process	RT		13	2.6	1.8E-1	1.4
☐	GOTERM_BP_DIRECT	vesicle-mediated transport	RT		18	3.6	1.3E-1	1.4
☐	GOTERM_BP_DIRECT	response to drug	RT		13	2.6	2.2E-1	1.4
☐	GOTERM_BP_DIRECT	methylation	RT		11	2.2	2.8E-1	1.4
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		21	4.2	1.5E-1	1.3
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		26	5.2	1.1E-1	1.3
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		34	6.8	7.7E-2	1.3
☐	GOTERM_BP_DIRECT	DNA repair	RT		25	5.0	1.5E-1	1.3
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		25	5.0	1.6E-1	1.3
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		16	3.2	2.6E-1	1.3
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		18	3.6	2.5E-1	1.3
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		27	5.4	2.0E-1	1.2
☐	GOTERM_BP_DIRECT	cellular response to DNA damage stimulus	RT		29	5.8	2.6E-1	1.2
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		41	8.2	2.0E-1	1.2
☐	GOTERM_BP_DIRECT	protein transport	RT		36	7.2	2.6E-1	1.1

Fig. B.26 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_NN between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.**

Current Gene List: scVI_up_5
Current Background: allscMVI
502 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

11 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	mitochondrial translation	RT		12	2.4	2.7E-2	2.1
☐	GOTERM_BP_DIRECT	cell adhesion	RT		15	3.0	4.5E-2	1.7
☐	GOTERM_BP_DIRECT	positive regulation of protein phosphorylation	RT		12	2.4	1.2E-1	1.6
☐	GOTERM_BP_DIRECT	fatty acid metabolic process	RT		10	2.0	1.7E-1	1.6
☐	GOTERM_BP_DIRECT	angiogenesis	RT		10	2.0	2.1E-1	1.5
☐	GOTERM_BP_DIRECT	mRNA splicing, via spliceosome	RT		13	2.6	1.5E-1	1.5
☐	GOTERM_BP_DIRECT	mitotic cell cycle	RT		11	2.2	2.4E-1	1.4
☐	GOTERM_BP_DIRECT	cell differentiation	RT		30	6.0	7.3E-2	1.4
☐	GOTERM_BP_DIRECT	intracellular signal transduction	RT		16	3.2	2.1E-1	1.3
☐	GOTERM_BP_DIRECT	cell migration	RT		12	2.4	2.9E-1	1.3
☐	GOTERM_BP_DIRECT	RNA splicing	RT		20	4.0	1.7E-1	1.3

Fig. B.27 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_NN between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.**

Current Gene List: scMVI_lw_4
Current Background: allscMVI
500 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

22 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	cellular response to leukemia inhibitory factor	RT		10	2.0	6.0E-2	2.0
☐	GOTERM_BP_DIRECT	methylation	RT		16	3.2	1.4E-2	2.0
☐	GOTERM_BP_DIRECT	rhythmic process	RT		11	2.2	6.4E-2	1.9
☐	GOTERM_BP_DIRECT	ribosome biogenesis	RT		11	2.2	9.9E-2	1.7
☐	GOTERM_BP_DIRECT	response to oxidative stress	RT		10	2.0	1.2E-1	1.7
☐	GOTERM_BP_DIRECT	regulation of apoptotic process	RT		13	2.6	7.8E-2	1.7
☐	GOTERM_BP_DIRECT	chromatin remodeling	RT		13	2.6	9.4E-2	1.6
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		17	3.4	5.6E-2	1.6
☐	GOTERM_BP_DIRECT	brain development	RT		14	2.8	1.4E-1	1.5
☐	GOTERM_BP_DIRECT	transcription from RNA polymerase II promoter	RT		12	2.4	1.8E-1	1.5
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		17	3.4	1.3E-1	1.4
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		20	4.0	1.2E-1	1.4
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		17	3.4	1.9E-1	1.3
☐	GOTERM_BP_DIRECT	chromatin organization	RT		19	3.8	2.1E-1	1.3
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		20	4.0	2.3E-1	1.3
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		24	4.8	2.3E-1	1.2
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		26	5.2	2.4E-1	1.2
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		46	9.2	1.3E-1	1.2
☐	GOTERM_BP_DIRECT	apoptotic process	RT		29	5.8	2.3E-1	1.2
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		31	6.2	2.2E-1	1.2
☐	GOTERM_BP_DIRECT	DNA repair	RT		23	4.6	2.9E-1	1.2
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		40	8.0	2.8E-1	1.1

Fig. B.28 **GSEA analysis on** 500 **genomic regions with the strongest negative DDLFC between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_up_4
Current Background: allscMVI
501 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

13 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	defense response to virus	RT		13	2.6	5.0E-3	2.5
☐	GOTERM_BP_DIRECT	aging	RT		13	2.6	7.3E-3	2.4
☐	GOTERM_BP_DIRECT	cell projection organization	RT		13	2.6	3.6E-2	1.9
☐	GOTERM_BP_DIRECT	cilium assembly	RT		14	2.8	8.5E-2	1.6
☐	GOTERM_BP_DIRECT	immune system process	RT		14	2.8	1.1E-1	1.6
☐	GOTERM_BP_DIRECT	endocytosis	RT		12	2.4	1.6E-1	1.5
☐	GOTERM_BP_DIRECT	ubiquitin-dependent protein catabolic process	RT		19	3.8	1.1E-1	1.4
☐	GOTERM_BP_DIRECT	signal transduction	RT		24	4.8	7.5E-2	1.4
☐	GOTERM_BP_DIRECT	cell proliferation	RT		15	3.0	1.9E-1	1.4
☐	GOTERM_BP_DIRECT	ion transport	RT		13	2.6	2.5E-1	1.3
☐	GOTERM_BP_DIRECT	chromatin organization	RT		19	3.8	2.0E-1	1.3
☐	GOTERM_BP_DIRECT	cellular response to DNA damage stimulus	RT		32	6.4	1.1E-1	1.3
☐	GOTERM_BP_DIRECT	mRNA processing	RT		23	4.6	2.5E-1	1.2

Fig. B.29 **GSEA analysis on** 500 **genomic regions with the strongest positive DDLFC between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scVI_lw_3
Current Background: allscMVI
502 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

13 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	response to insulin	RT		10	2.0	1.6E-3	3.5
☐	GOTERM_BP_DIRECT	cellular response to starvation	RT		11	2.2	1.7E-3	3.2
☐	GOTERM_BP_DIRECT	defense response to virus	RT		12	2.4	1.4E-2	2.3
☐	GOTERM_BP_DIRECT	positive regulation of protein phosphorylation	RT		16	3.2	6.2E-3	2.2
☐	GOTERM_BP_DIRECT	transcription, DNA-templated	RT		10	2.0	8.4E-2	1.9
☐	GOTERM_BP_DIRECT	positive regulation of neuron projection development	RT		10	2.0	1.1E-1	1.8
☐	GOTERM_BP_DIRECT	protein autophosphorylation	RT		11	2.2	1.8E-1	1.5
☐	GOTERM_BP_DIRECT	autophagy	RT		11	2.2	1.9E-1	1.5
☐	GOTERM_BP_DIRECT	immune system process	RT		13	2.6	1.9E-1	1.4
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		26	5.2	1.2E-1	1.3
☐	GOTERM_BP_DIRECT	proteasome-mediated ubiquitin-dependent protein catabolic process	RT		14	2.8	2.5E-1	1.3
☐	GOTERM_BP_DIRECT	signal transduction	RT		22	4.4	1.8E-1	1.3
☐	GOTERM_BP_DIRECT	protein phosphorylation	RT		27	5.4	1.9E-1	1.2

Fig. B.30 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_NN between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.**

Current Gene List: scVI_up_3
Current Background: allscMVI
502 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

11 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	cell-cell adhesion	RT		10	2.0	2.3E-2	2.3
☐	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		21	4.2	5.5E-2	1.5
☐	GOTERM_BP_DIRECT	transmembrane transport	RT		10	2.0	3.0E-1	1.4
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		52	10.4	2.1E-2	1.3
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		17	3.4	2.0E-1	1.3
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		14	2.8	2.6E-1	1.3
☐	GOTERM_BP_DIRECT	protein stabilization	RT		13	2.6	2.9E-1	1.3
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		38	7.6	7.7E-2	1.3
☐	GOTERM_BP_DIRECT	mRNA processing	RT		23	4.6	2.8E-1	1.2
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		26	5.2	2.6E-1	1.2
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		42	8.4	2.0E-1	1.2

Fig. B.31 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_NN between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by NNLFC.**

Current Gene List: scMVI_mESC_Pj_DT_lw_0
Current Background: all_scMVI_mESC_Pj_DT
500 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

14 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	neural tube closure	RT		13	2.6	6.7E-3	2.4
☐	GOTERM_BP_DIRECT	protein autophosphorylation	RT		13	2.6	5.1E-2	1.8
☐	GOTERM_BP_DIRECT	immune system process	RT		16	3.2	3.4E-2	1.8
☐	GOTERM_BP_DIRECT	positive regulation of neuron projection development	RT		10	2.0	1.1E-1	1.8
☐	GOTERM_BP_DIRECT	cell projection organization	RT		11	2.2	1.4E-1	1.6
☐	GOTERM_BP_DIRECT	fatty acid metabolic process	RT		10	2.0	1.7E-1	1.6
☐	GOTERM_BP_DIRECT	brain development	RT		14	2.8	1.3E-1	1.5
☐	GOTERM_BP_DIRECT	innate immune response	RT		12	2.4	2.3E-1	1.4
☐	GOTERM_BP_DIRECT	mRNA splicing, via spliceosome	RT		12	2.4	2.4E-1	1.4
☐	GOTERM_BP_DIRECT	signal transduction	RT		23	4.6	1.2E-1	1.4
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		26	5.2	1.1E-1	1.3
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		19	3.8	1.8E-1	1.3
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		17	3.4	2.1E-1	1.3
☐	GOTERM_BP_DIRECT	translation	RT		20	4.0	2.3E-1	1.3

Fig. B.32 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_DT between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.**

Current Gene List: scMVI_mESC_Pj_DT_up_0
Current Background: all_scMVI_mESC_Pj_DT
501 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

16 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	mitochondrial translation	RT		13	2.6	1.1E-2	2.2
☐	GOTERM_BP_DIRECT	mRNA transport	RT		11	2.2	7.5E-2	1.8
☐	GOTERM_BP_DIRECT	cell adhesion	RT		15	3.0	4.5E-2	1.7
☐	GOTERM_BP_DIRECT	protein localization	RT		10	2.0	1.4E-1	1.7
☐	GOTERM_BP_DIRECT	fatty acid metabolic process	RT		10	2.0	1.7E-1	1.6
☐	GOTERM_BP_DIRECT	angiogenesis	RT		10	2.0	2.1E-1	1.5
☐	GOTERM_BP_DIRECT	cell differentiation	RT		33	6.6	1.9E-2	1.5
☐	GOTERM_BP_DIRECT	transcription from RNA polymerase II promoter	RT		12	2.4	1.8E-1	1.5
☐	GOTERM_BP_DIRECT	protein autophosphorylation	RT		10	2.0	2.9E-1	1.4
☐	GOTERM_BP_DIRECT	chromatin remodeling	RT		11	2.2	2.7E-1	1.4
☐	GOTERM_BP_DIRECT	cell cycle	RT		44	8.8	2.7E-2	1.4
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		14	2.8	2.5E-1	1.3
☐	GOTERM_BP_DIRECT	cell division	RT		28	5.6	1.0E-1	1.3
☐	GOTERM_BP_DIRECT	protein stabilization	RT		13	2.6	2.7E-1	1.3
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		16	3.2	2.8E-1	1.3
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		27	5.4	2.1E-1	1.2

Fig. B.33 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_DT between embryonic days** $E8.0$ **and** $E8.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.**

Current Gene List: scMVI_mESC_Pj_DT_lw_4
Current Background: all_scMVI_mESC_Pj_DT
501 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

23 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	anterior/posterior pattern specification	RT		10	2.0	2.9E-3	3.2
☐	GOTERM_BP_DIRECT	positive regulation of translation	RT		11	2.2	5.4E-3	2.7
☐	GOTERM_BP_DIRECT	positive regulation of NF-kappaB transcription factor activity	RT		10	2.0	1.2E-2	2.6
☐	GOTERM_BP_DIRECT	tRNA processing	RT		12	2.4	5.6E-3	2.6
☐	GOTERM_BP_DIRECT	negative regulation of canonical Wnt signaling pathway	RT		10	2.0	4.4E-2	2.1
☐	GOTERM_BP_DIRECT	cytoplasmic translation	RT		10	2.0	8.0E-2	1.9
☐	GOTERM_BP_DIRECT	ER to Golgi vesicle-mediated transport	RT		10	2.0	1.3E-1	1.7
☐	GOTERM_BP_DIRECT	transmembrane transport	RT		12	2.4	9.5E-2	1.7
☐	GOTERM_BP_DIRECT	chromatin remodeling	RT		13	2.6	9.0E-2	1.7
☐	GOTERM_BP_DIRECT	ribosome biogenesis	RT		10	2.0	1.7E-1	1.6
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		24	4.8	3.9E-2	1.5
☐	GOTERM_BP_DIRECT	methylation	RT		12	2.4	1.7E-1	1.5
☐	GOTERM_BP_DIRECT	cell proliferation	RT		16	3.2	1.2E-1	1.5
☐	GOTERM_BP_DIRECT	cell division	RT		30	6.0	4.1E-2	1.4
☐	GOTERM_BP_DIRECT	mitotic cell cycle	RT		11	2.2	2.4E-1	1.4
☐	GOTERM_BP_DIRECT	signal transduction	RT		24	4.8	7.5E-2	1.4
☐	GOTERM_BP_DIRECT	innate immune response	RT		12	2.4	2.3E-1	1.4
☐	GOTERM_BP_DIRECT	negative regulation of cell proliferation	RT		15	3.0	2.2E-1	1.3
☐	GOTERM_BP_DIRECT	vesicle-mediated transport	RT		17	3.4	2.0E-1	1.3
☐	GOTERM_BP_DIRECT	translation	RT		21	4.2	1.5E-1	1.3
☐	GOTERM_BP_DIRECT	cell cycle	RT		40	8.0	1.0E-1	1.3
☐	GOTERM_BP_DIRECT	cell differentiation	RT		27	5.4	2.0E-1	1.2
☐	GOTERM_BP_DIRECT	phosphorylation	RT		26	5.2	2.8E-1	1.2

Fig. B.34 **GSEA analysis on** 500 **genomic regions with the strongest negative LFC_DT between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.**

Current Gene List: scMVI_mESC_Pj_DT_up_4
Current Background: all_scMVI_mESC_Pj_DT
498 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

15 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	DNA replication	RT		14	2.8	1.7E-2	2.0
☐	GOTERM_BP_DIRECT	protein polyubiquitination	RT		14	2.8	5.8E-2	1.7
☐	GOTERM_BP_DIRECT	regulation of apoptotic process	RT		13	2.6	8.4E-2	1.7
☐	GOTERM_BP_DIRECT	protein localization to plasma membrane	RT		10	2.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	chromatin remodeling	RT		12	2.4	1.7E-1	1.5
☐	GOTERM_BP_DIRECT	cell proliferation	RT		16	3.2	1.3E-1	1.4
☐	GOTERM_BP_DIRECT	cell migration	RT		13	2.6	2.0E-1	1.4
☐	GOTERM_BP_DIRECT	cellular response to DNA damage stimulus	RT		36	7.2	2.9E-2	1.4
☐	GOTERM_BP_DIRECT	ion transport	RT		14	2.8	1.8E-1	1.4
☐	GOTERM_BP_DIRECT	protein ubiquitination	RT		23	4.6	1.7E-1	1.3
☐	GOTERM_BP_DIRECT	DNA repair	RT		25	5.0	1.7E-1	1.3
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		25	5.0	1.8E-1	1.3
☐	GOTERM_BP_DIRECT	ubiquitin-dependent protein catabolic process	RT		17	3.4	2.8E-1	1.2
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		20	4.0	2.5E-1	1.2
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		18	3.6	2.8E-1	1.2

Fig. B.35 **GSEA analysis on** 500 **genomic regions with the strongest positive LFC_DT between embryonic days** $E7.0$ **and** $E7.5$ **for mESC_Pj. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential expression analysis.**

Current Gene List: scMVI_rna_lw_1
Current Background: all
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

22 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	cytoplasmic translation	RT		37	14.7	6.1E-34	13.9
☐	GOTERM_BP_DIRECT	translation	RT		45	17.9	7.9E-22	5.6
☐	GOTERM_BP_DIRECT	rRNA processing	RT		21	8.4	5.7E-8	4.3
☐	GOTERM_BP_DIRECT	ribosome biogenesis	RT		13	5.2	1.1E-4	3.9
☐	GOTERM_BP_DIRECT	cellular response to leukemia inhibitory factor	RT		11	4.4	1.5E-4	4.4
☐	GOTERM_BP_DIRECT	protein folding	RT		11	4.4	6.1E-4	3.7
☐	GOTERM_BP_DIRECT	negative regulation of neuron apoptotic process	RT		10	4.0	1.5E-3	3.6
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		21	8.4	1.9E-3	2.1
☐	GOTERM_BP_DIRECT	protein stabilization	RT		13	5.2	4.2E-3	2.6
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		13	5.2	6.5E-3	2.4
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		16	6.4	1.2E-2	2.0
☐	GOTERM_BP_DIRECT	innate immune response	RT		10	4.0	2.7E-2	2.3
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		14	5.6	2.8E-2	1.9
☐	GOTERM_BP_DIRECT	mRNA splicing, via spliceosome	RT		10	4.0	2.8E-2	2.3
☐	GOTERM_BP_DIRECT	immune system process	RT		10	4.0	3.7E-2	2.2
☐	GOTERM_BP_DIRECT	response to drug	RT		10	4.0	4.7E-2	2.1
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		22	8.8	5.6E-2	1.5
☐	GOTERM_BP_DIRECT	RNA splicing	RT		13	5.2	7.7E-2	1.7
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		11	4.4	1.2E-1	1.7
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		15	6.0	1.2E-1	1.5
☐	GOTERM_BP_DIRECT	spermatogenesis	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	mRNA processing	RT		13	5.2	2.7E-1	1.3

Fig. B.36 **Gene set enrichment analysis for features with the strongest negative gene expression DDLFC between embryonic days $E6.5$ and $E7.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_rna_up_1
Current Background: all
250 DAVID IDs
Options

Rerun Using Options | Create Sublist

16 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	intraciliary transport	RT		12	4.8	2.3E-11	15.8
☐	GOTERM_BP_DIRECT	determination of left/right symmetry	RT		14	5.6	3.0E-12	13.6
☐	GOTERM_BP_DIRECT	non-motile cilium assembly	RT		11	4.4	7.8E-8	9.6
☐	GOTERM_BP_DIRECT	smoothened signaling pathway	RT		11	4.4	7.1E-7	7.8
☐	GOTERM_BP_DIRECT	cilium assembly	RT		30	12.0	1.1E-17	7.3
☐	GOTERM_BP_DIRECT	cell projection organization	RT		20	8.0	3.3E-10	6.0
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		30	12.0	4.9E-7	2.8
☐	GOTERM_BP_DIRECT	heart development	RT		11	4.4	7.5E-3	2.7
☐	GOTERM_BP_DIRECT	brain development	RT		10	4.0	3.6E-2	2.2
☐	GOTERM_BP_DIRECT	spermatogenesis	RT		12	4.8	3.3E-2	2.0
☐	GOTERM_BP_DIRECT	cell proliferation	RT		10	4.0	6.7E-2	2.0
☐	GOTERM_BP_DIRECT	cell differentiation	RT		19	7.6	1.4E-2	1.8
☐	GOTERM_BP_DIRECT	vesicle-mediated transport	RT		10	4.0	1.4E-1	1.7
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		18	7.2	1.0E-1	1.5
☐	GOTERM_BP_DIRECT	signal transduction	RT		12	4.8	2.0E-1	1.5
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		24	9.6	1.3E-1	1.3

Fig. B.37 **Gene set enrichment analysis for features with the strongest positive gene expression DDLFC between embryonic days $E6.5$ and $E7.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_rna_lw_2
Current Background: all
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

16 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	cytoplasmic translation	RT		53	21.1	2.5E-62	19.8
☐	GOTERM_BP_DIRECT	translation	RT		58	23.1	8.4E-35	7.2
☐	GOTERM_BP_DIRECT	negative regulation of neuron apoptotic process	RT		11	4.4	3.5E-4	4.0
☐	GOTERM_BP_DIRECT	rRNA processing	RT		14	5.6	1.2E-3	2.8
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		16	6.4	5.7E-3	2.2
☐	GOTERM_BP_DIRECT	protein stabilization	RT		11	4.4	2.9E-2	2.2
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		12	4.8	3.8E-2	2.0
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		19	7.6	9.8E-3	1.9
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		10	4.0	8.5E-2	1.9
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		14	5.6	5.4E-2	1.8
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		11	4.4	1.2E-1	1.7
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		19	7.6	1.1E-1	1.4
☐	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		10	4.0	2.6E-1	1.4
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		26	10.4	6.0E-2	1.4
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		14	5.6	2.0E-1	1.4
☐	GOTERM_BP_DIRECT	apoptotic process	RT		16	6.4	2.5E-1	1.3

Fig. B.38 **Gene set enrichment analysis for features with the strongest negative gene expression DDLFC between embryonic days $E5.5$ and $E6.5$ for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_rna_up_2
Current Background: all
250 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

5 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	cilium assembly	RT		21	8.4	3.6E-9	5.0
☐	GOTERM_BP_DIRECT	cell projection organization	RT		10	4.0	6.2E-3	3.0
☐	GOTERM_BP_DIRECT	intracellular protein transport	RT		12	4.8	9.5E-2	1.7
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		31	12.4	4.9E-3	1.7
☐	GOTERM_BP_DIRECT	vesicle-mediated transport	RT		10	4.0	1.6E-1	1.6

Fig. B.39 **Gene set enrichment analysis for features with the strongest positive gene expression DDLFC between embryonic days $E5.5$ and $E6.5$ formESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_rna_up_3
Current Background: all
250 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

8 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	cilium assembly	RT		21	8.4	2.3E-9	5.1
☐	GOTERM_BP_DIRECT	cell projection organization	RT		14	5.6	1.9E-5	4.2
☐	GOTERM_BP_DIRECT	spermatogenesis	RT		13	5.2	1.4E-2	2.2
☐	GOTERM_BP_DIRECT	cell division	RT		16	6.4	8.5E-2	1.6
☐	GOTERM_BP_DIRECT	cell differentiation	RT		15	6.0	1.4E-1	1.5
☐	GOTERM_BP_DIRECT	cell cycle	RT		21	8.4	1.3E-1	1.4
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		14	5.6	2.5E-1	1.3
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		17	6.8	2.3E-1	1.3

Fig. B.40 **Gene set enrichment analysis for features with the strongest positive gene expression DDLFC between embryonic days** $E4.5$ **and** $E5.5$ **for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Help and Manual

Current Gene List: scMVI_rna_lw_3
Current Background: all
250 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

22 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	cell adhesion	RT		21	8.4	1.3E-8	4.7
☐	GOTERM_BP_DIRECT	transmembrane transport	RT		11	4.4	2.6E-3	3.1
☐	GOTERM_BP_DIRECT	positive regulation of cell migration	RT		11	4.4	3.0E-3	3.0
☐	GOTERM_BP_DIRECT	angiogenesis	RT		10	4.0	5.7E-3	3.0
☐	GOTERM_BP_DIRECT	Wnt signaling pathway	RT		11	4.4	4.9E-3	2.8
☐	GOTERM_BP_DIRECT	cell migration	RT		13	5.2	1.9E-3	2.8
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		19	7.6	2.3E-4	2.7
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		14	5.6	5.9E-3	2.4
☐	GOTERM_BP_DIRECT	response to drug	RT		11	4.4	1.8E-2	2.4
☐	GOTERM_BP_DIRECT	ion transport	RT		11	4.4	2.4E-2	2.2
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		16	6.4	1.1E-2	2.0
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		13	5.2	2.5E-2	2.0
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		22	8.8	3.6E-3	2.0
☐	GOTERM_BP_DIRECT	cell differentiation	RT		21	8.4	5.9E-3	1.9
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		31	12.4	2.8E-3	1.7
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		22	8.8	1.8E-2	1.7
☐	GOTERM_BP_DIRECT	ubiquitin-dependent protein catabolic process	RT		11	4.4	1.3E-1	1.6
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		16	6.4	6.6E-2	1.6
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		22	8.8	3.6E-2	1.6
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		15	6.0	1.0E-1	1.5
☐	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		10	4.0	2.4E-1	1.5
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		24	9.6	2.0E-1	1.3

Fig. B.41 **Gene set enrichment analysis for features with the strongest negative gene expression DDLFC between embryonic days** $E4.5$ **and** $E5.5$ **for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_met_up_1
Current Background: all
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

27 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		47	18.7	8.0E-9	2.5
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		33	13.1	1.1E-6	2.6
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		40	15.9	1.5E-6	2.3
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		32	12.7	1.4E-5	2.3
☐	GOTERM_BP_DIRECT	brain development	RT		16	6.4	6.8E-5	3.4
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		26	10.4	1.0E-4	2.3
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		24	9.6	1.0E-4	2.4
☐	GOTERM_BP_DIRECT	gene expression	RT		10	4.0	1.6E-4	4.9
☐	GOTERM_BP_DIRECT	axon guidance	RT		10	4.0	2.1E-4	4.7
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		29	11.6	4.2E-4	2.0
☐	GOTERM_BP_DIRECT	cell differentiation	RT		23	9.2	1.1E-3	2.1
☐	GOTERM_BP_DIRECT	neural tube closure	RT		10	4.0	1.1E-3	3.8
☐	GOTERM_BP_DIRECT	nervous system development	RT		15	6.0	1.3E-3	2.7
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		15	6.0	2.1E-3	2.5
☐	GOTERM_BP_DIRECT	chromatin organization	RT		16	6.4	7.9E-3	2.1
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		16	6.4	1.0E-2	2.1
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		15	6.0	1.1E-2	2.1
☐	GOTERM_BP_DIRECT	signal transduction	RT		16	6.4	2.6E-2	1.8
☐	GOTERM_BP_DIRECT	cell adhesion	RT		10	4.0	3.4E-2	2.2
☐	GOTERM_BP_DIRECT	cell proliferation	RT		11	4.4	4.1E-2	2.0
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		12	4.8	5.5E-2	1.9
☐	GOTERM_BP_DIRECT	apoptotic process	RT		19	7.6	5.7E-2	1.6
☐	GOTERM_BP_DIRECT	protein phosphorylation	RT		17	6.8	7.4E-2	1.6
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		10	4.0	7.5E-2	1.9
☐	GOTERM_BP_DIRECT	phosphorylation	RT		16	6.4	1.4E-1	1.5
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		14	5.6	1.7E-1	1.4
☐	GOTERM_BP_DIRECT	cell division	RT		15	6.0	1.9E-1	1.4

Fig. B.42 **Gene set enrichment analysis for features with the strongest positive DNA methylation DDLFC between embryonic days** $E6.5$ **and** $E7.5$ **for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_acc_up_1
Current Background: all
250 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

8 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		13	5.2	8.1E-2	1.7
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	ubiquitin-dependent protein catabolic process	RT		10	4.0	2.1E-1	1.5
☐	GOTERM_BP_DIRECT	protein ubiquitination	RT		13	5.2	1.5E-1	1.5
☐	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		10	4.0	2.2E-1	1.5
☐	GOTERM_BP_DIRECT	protein transport	RT		23	9.2	6.2E-2	1.5
☐	GOTERM_BP_DIRECT	cell division	RT		15	6.0	1.8E-1	1.4
☐	GOTERM_BP_DIRECT	cell cycle	RT		20	8.0	2.5E-1	1.2

Fig. B.43 **Gene set enrichment analysis for features with the strongest positive chromatin accessibility DDLFC between embryonic days** $E6.5$ **and** $E7.5$ **for mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by DDLFC.**

Current Gene List: scMVI_rna_DT_lw_0
Current Background: all_scMVI_rna_DT
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

14 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	transmembrane transport	RT		10	4.0	7.7E-3	2.9
☐	GOTERM_BP_DIRECT	response to drug	RT		13	5.2	2.0E-3	2.8
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		17	6.8	1.6E-3	2.4
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		26	10.4	6.3E-5	2.4
☐	GOTERM_BP_DIRECT	cell proliferation	RT		11	4.4	3.9E-2	2.1
☐	GOTERM_BP_DIRECT	ion transport	RT		10	4.0	5.2E-2	2.1
☐	GOTERM_BP_DIRECT	spermatogenesis	RT		12	4.8	4.3E-2	1.9
☐	GOTERM_BP_DIRECT	proteolysis	RT		12	4.8	5.8E-2	1.9
☐	GOTERM_BP_DIRECT	intracellular signal transduction	RT		11	4.4	7.3E-2	1.8
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		11	4.4	9.9E-2	1.7
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		10	4.0	1.3E-1	1.7
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		11	4.4	2.4E-1	1.4
☐	GOTERM_BP_DIRECT	signal transduction	RT		12	4.8	2.4E-1	1.4

Fig. B.44 **Gene set enrichment analysis for features with the strongest negative gene expression LFC_DT between embryonic days** $E6.5$ **and** $E7.5$ **in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.**

Current Gene List: scMVI_rna_DT_up_0
Current Background: all_scMVI_rna_DT
251 DAVID IDs

Options

Thresholds: Count 10 EASE 0.3

Display: ☑ Fold Enrichment ☐ Bonferroni ☐ Benjamini ☐ FDR ☐ Fisher Exact ☐ LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

38 chart records Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	heart looping	RT		12	4.8	4.5E-10	12.6
☐	GOTERM_BP_DIRECT	determination of left/right symmetry	RT		13	5.2	1.1E-10	12.2
☐	GOTERM_BP_DIRECT	positive regulation of epithelial cell proliferation	RT		10	4.0	3.4E-7	9.7
☐	GOTERM_BP_DIRECT	BMP signaling pathway	RT		12	4.8	3.6E-8	8.9
☐	GOTERM_BP_DIRECT	axon guidance	RT		10	4.0	1.9E-4	4.8
☐	GOTERM_BP_DIRECT	negative regulation of canonical Wnt signaling pathway	RT		11	4.4	8.6E-5	4.7
☐	GOTERM_BP_DIRECT	positive regulation of MAPK cascade	RT		10	4.0	5.0E-4	4.2
☐	GOTERM_BP_DIRECT	kidney development	RT		10	4.0	6.5E-4	4.1
☐	GOTERM_BP_DIRECT	cell adhesion	RT		18	7.2	1.4E-6	4.1
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		42	16.7	6.3E-14	3.8
☐	GOTERM_BP_DIRECT	cilium assembly	RT		16	6.4	1.8E-5	3.7
☐	GOTERM_BP_DIRECT	positive regulation of cell migration	RT		13	5.2	1.9E-4	3.6
☐	GOTERM_BP_DIRECT	heart development	RT		15	6.0	7.6E-5	3.5
☐	GOTERM_BP_DIRECT	angiogenesis	RT		11	4.4	1.5E-3	3.3
☐	GOTERM_BP_DIRECT	cell projection organization	RT		11	4.4	2.1E-3	3.2
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		22	8.8	5.1E-6	3.1
☐	GOTERM_BP_DIRECT	brain development	RT		13	5.2	2.4E-3	2.8
☐	GOTERM_BP_DIRECT	nervous system development	RT		15	6.0	1.2E-3	2.7
☐	GOTERM_BP_DIRECT	Wnt signaling pathway	RT		10	4.0	1.3E-2	2.6
☐	GOTERM_BP_DIRECT	cell differentiation	RT		28	11.2	6.5E-6	2.6
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		19	7.6	5.9E-4	2.5
☐	GOTERM_BP_DIRECT	cell proliferation	RT		12	4.8	1.6E-2	2.3
☐	GOTERM_BP_DIRECT	signal transduction	RT		17	6.8	1.1E-2	2.0
☐	GOTERM_BP_DIRECT	protein phosphorylation	RT		21	8.4	4.8E-3	2.0
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		10	4.0	7.0E-2	1.9
☐	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		13	5.2	3.4E-2	1.9
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		12	4.8	4.9E-2	1.9
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		11	4.4	6.4E-2	1.9
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		17	6.8	2.8E-2	1.8
☐	GOTERM_BP_DIRECT	intracellular signal transduction	RT		10	4.0	1.4E-1	1.7
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		21	8.4	2.8E-2	1.7
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		16	6.4	5.9E-2	1.6
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		22	8.8	4.2E-2	1.6
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		27	10.8	2.5E-2	1.5
☐	GOTERM_BP_DIRECT	phosphorylation	RT		16	6.4	1.3E-1	1.5
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		27	10.8	5.3E-2	1.4
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		17	6.8	2.9E-1	1.2

Fig. B.45 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days** $E6.5$ **and** $E7.5$ **in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.**

Current Gene List: scMVI_rna_DT_lw_1
Current Background: all_scMVI_rna_DT
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

14 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		26	10.4	6.3E-5	2.4
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		17	6.8	1.6E-3	2.4
☐	GOTERM_BP_DIRECT	response to drug	RT		13	5.2	2.0E-3	2.8
☐	GOTERM_BP_DIRECT	transmembrane transport	RT		10	4.0	7.7E-3	2.9
☐	GOTERM_BP_DIRECT	cell proliferation	RT		11	4.4	3.9E-2	2.1
☐	GOTERM_BP_DIRECT	spermatogenesis	RT		12	4.8	4.3E-2	1.9
☐	GOTERM_BP_DIRECT	ion transport	RT		10	4.0	5.2E-2	2.1
☐	GOTERM_BP_DIRECT	proteolysis	RT		12	4.8	5.8E-2	1.9
☐	GOTERM_BP_DIRECT	intracellular signal transduction	RT		11	4.4	7.3E-2	1.8
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		11	4.4	9.9E-2	1.7
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		10	4.0	1.3E-1	1.7
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	signal transduction	RT		12	4.8	2.4E-1	1.4
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		11	4.4	2.4E-1	1.4

Fig. B.46 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days $E5.5$ and $E6.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.**

Current Gene List: scMVI_rna_DT_up_1
Current Background: all_scMVI_rna_DT
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

38 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	heart looping	RT		12	4.8	4.5E-10	12.6
☐	GOTERM_BP_DIRECT	determination of left/right symmetry	RT		13	5.2	1.1E-10	12.2
☐	GOTERM_BP_DIRECT	positive regulation of epithelial cell proliferation	RT		10	4.0	3.4E-7	9.7
☐	GOTERM_BP_DIRECT	BMP signaling pathway	RT		12	4.8	3.6E-8	8.9
☐	GOTERM_BP_DIRECT	axon guidance	RT		10	4.0	1.9E-4	4.8
☐	GOTERM_BP_DIRECT	negative regulation of canonical Wnt signaling pathway	RT		11	4.4	8.6E-5	4.7
☐	GOTERM_BP_DIRECT	positive regulation of MAPK cascade	RT		10	4.0	5.0E-4	4.2
☐	GOTERM_BP_DIRECT	kidney development	RT		10	4.0	6.5E-4	4.1
☐	GOTERM_BP_DIRECT	cell adhesion	RT		18	7.2	1.4E-6	4.1
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		42	16.7	6.3E-14	3.8
☐	GOTERM_BP_DIRECT	cilium assembly	RT		16	6.4	1.8E-5	3.7
☐	GOTERM_BP_DIRECT	positive regulation of cell migration	RT		13	5.2	1.9E-4	3.6
☐	GOTERM_BP_DIRECT	heart development	RT		15	6.0	7.6E-5	3.5
☐	GOTERM_BP_DIRECT	angiogenesis	RT		11	4.4	1.5E-3	3.3
☐	GOTERM_BP_DIRECT	cell projection organization	RT		11	4.4	2.1E-3	3.2
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		22	8.8	5.1E-6	3.1
☐	GOTERM_BP_DIRECT	brain development	RT		13	5.2	2.4E-3	2.8
☐	GOTERM_BP_DIRECT	nervous system development	RT		15	6.0	1.2E-3	2.7
☐	GOTERM_BP_DIRECT	Wnt signaling pathway	RT		10	4.0	1.3E-2	2.6
☐	GOTERM_BP_DIRECT	cell differentiation	RT		28	11.2	6.5E-6	2.6
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		19	7.6	5.9E-4	2.5
☐	GOTERM_BP_DIRECT	cell proliferation	RT		12	4.8	1.6E-2	2.3
☐	GOTERM_BP_DIRECT	signal transduction	RT		17	6.8	1.1E-2	2.0
☐	GOTERM_BP_DIRECT	protein phosphorylation	RT		21	8.4	4.8E-3	2.0
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		10	4.0	7.0E-2	1.9
☐	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		13	5.2	3.4E-2	1.9
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		12	4.8	4.9E-2	1.9
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		11	4.4	6.4E-2	1.9
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		17	6.8	2.8E-2	1.8
☐	GOTERM_BP_DIRECT	intracellular signal transduction	RT		10	4.0	1.4E-1	1.7
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		21	8.4	2.8E-2	1.7
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		16	6.4	5.9E-2	1.6
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		22	8.8	4.2E-2	1.6
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		27	10.8	2.5E-2	1.5
☐	GOTERM_BP_DIRECT	phosphorylation	RT		16	6.4	1.3E-1	1.5
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		27	10.8	5.3E-2	1.4
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		17	6.8	2.9E-1	1.2

Fig. B.47 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days $E5.5$ and $E6.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.**

Current Gene List: scMVI_rna_DT_up_2
Current Background: all_scMVI_rna_DT
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

38 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	heart looping	RT		12	4.8	4.5E-10	12.6
☐	GOTERM_BP_DIRECT	determination of left/right symmetry	RT		13	5.2	1.1E-10	12.2
☐	GOTERM_BP_DIRECT	positive regulation of epithelial cell proliferation	RT		10	4.0	3.4E-7	9.7
☐	GOTERM_BP_DIRECT	BMP signaling pathway	RT		12	4.8	3.6E-8	8.9
☐	GOTERM_BP_DIRECT	axon guidance	RT		10	4.0	1.9E-4	4.8
☐	GOTERM_BP_DIRECT	negative regulation of canonical Wnt signaling pathway	RT		11	4.4	8.6E-5	4.7
☐	GOTERM_BP_DIRECT	positive regulation of MAPK cascade	RT		10	4.0	5.0E-4	4.2
☐	GOTERM_BP_DIRECT	kidney development	RT		10	4.0	6.5E-4	4.1
☐	GOTERM_BP_DIRECT	cell adhesion	RT		18	7.2	1.4E-6	4.1
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		42	16.7	6.3E-14	3.8
☐	GOTERM_BP_DIRECT	cilium assembly	RT		16	6.4	1.8E-5	3.7
☐	GOTERM_BP_DIRECT	positive regulation of cell migration	RT		13	5.2	1.9E-4	3.6
☐	GOTERM_BP_DIRECT	heart development	RT		15	6.0	7.6E-5	3.5
☐	GOTERM_BP_DIRECT	angiogenesis	RT		11	4.4	1.5E-3	3.3
☐	GOTERM_BP_DIRECT	cell projection organization	RT		11	4.4	2.1E-3	3.2
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		22	8.8	5.1E-6	3.1
☐	GOTERM_BP_DIRECT	brain development	RT		13	5.2	2.4E-3	2.8
☐	GOTERM_BP_DIRECT	nervous system development	RT		15	6.0	1.2E-3	2.7
☐	GOTERM_BP_DIRECT	Wnt signaling pathway	RT		10	4.0	1.3E-2	2.6
☐	GOTERM_BP_DIRECT	cell differentiation	RT		28	11.2	6.5E-6	2.6
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		19	7.6	5.9E-4	2.5
☐	GOTERM_BP_DIRECT	cell proliferation	RT		12	4.8	1.6E-2	2.3
☐	GOTERM_BP_DIRECT	signal transduction	RT		17	6.8	1.1E-2	2.0
☐	GOTERM_BP_DIRECT	protein phosphorylation	RT		21	8.4	4.8E-3	2.0
☐	GOTERM_BP_DIRECT	negative regulation of gene expression	RT		10	4.0	7.0E-2	1.9
☐	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		13	5.2	3.4E-2	1.9
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		12	4.8	4.9E-2	1.9
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		11	4.4	6.4E-2	1.9
☐	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		17	6.8	2.8E-2	1.8
☐	GOTERM_BP_DIRECT	intracellular signal transduction	RT		10	4.0	1.4E-1	1.7
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		21	8.4	2.8E-2	1.7
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		16	6.4	5.9E-2	1.6
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		22	8.8	4.2E-2	1.6
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		27	10.8	2.5E-2	1.5
☐	GOTERM_BP_DIRECT	phosphorylation	RT		16	6.4	1.3E-1	1.5
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		27	10.8	5.3E-2	1.4
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		17	6.8	2.9E-1	1.2

Fig. B.48 **Gene set enrichment analysis for features with the strongest positive gene expression LFC_DT between embryonic days $E4.5$ and $E5.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.**

Current Gene List: scMVI_rna_DT_lw_2
Current Background: all_scMVI_rna_DT
251 DAVID IDs
Options
Thresholds: Count 10 EASE 0.3
Display: ☑Fold Enrichment ☐Bonferroni ☐Benjamini ☐FDR ☐Fisher Exact ☐LT,PH,PT # of Records 1000

Rerun Using Options | Create Sublist

14 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment
☐	GOTERM_BP_DIRECT	lipid metabolic process	RT		26	10.4	6.3E-5	2.4
☐	GOTERM_BP_DIRECT	positive regulation of gene expression	RT		17	6.8	1.6E-3	2.4
☐	GOTERM_BP_DIRECT	response to drug	RT		13	5.2	2.0E-3	2.8
☐	GOTERM_BP_DIRECT	transmembrane transport	RT		10	4.0	7.7E-3	2.9
☐	GOTERM_BP_DIRECT	cell proliferation	RT		11	4.4	3.9E-2	2.1
☐	GOTERM_BP_DIRECT	spermatogenesis	RT		12	4.8	4.3E-2	1.9
☐	GOTERM_BP_DIRECT	ion transport	RT		10	4.0	5.2E-2	2.1
☐	GOTERM_BP_DIRECT	proteolysis	RT		12	4.8	5.8E-2	1.9
☐	GOTERM_BP_DIRECT	intracellular signal transduction	RT		11	4.4	7.3E-2	1.8
☐	GOTERM_BP_DIRECT	in utero embryonic development	RT		11	4.4	9.9E-2	1.7
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		10	4.0	1.3E-1	1.7
☐	GOTERM_BP_DIRECT	regulation of cell cycle	RT		10	4.0	1.8E-1	1.6
☐	GOTERM_BP_DIRECT	signal transduction	RT		12	4.8	2.4E-1	1.4
☐	GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT		11	4.4	2.4E-1	1.4

Fig. B.49 **Gene set enrichment analysis for features with the strongest negative gene expression LFC_DT between embryonic days $E4.5$ and $E5.5$ in mESC_Ag. Figures used to demonstrate potential biological processes initiated from one developmental stage to the next indicated by differential gene expression analysis.**

Bibliography

10X (2020). Pbmc from a healthy donor - granulocytes removed through cell sorting (10k).

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines*. *Cognitive Science*, 9.

Adossa, N., Khan, S., Rytkönen, K. T., and Elo, L. L. (2021). Computational strategies for single-cell multi-omics integration. *Computational and Structural Biotechnology Journal*, 19.

Ahn, J., Heo, S., Lee, J., and Bang, D. (2021). Introduction to single-cell dna methylation profiling methods. *Biomolecules*, 11.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. proceedings of the 2nd international symposium on information theory. *Second International Symposium on Information Theory*.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). Molecular biology of the cell. *Molecular Biology of the Cell*.

Alekseyev, Y. O., Fazeli, R., Yang, S., Basran, R., Maher, T., Miller, N. S., and Remick, D. (2018). A next-generation sequencing primer—how does it work and what can it do? *Academic Pathology*, 5.

Allis, C. D. and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 17.

Amari, S. I. (2016). Information geometry and its applications. *Applied Mathematical Sciences (Switzerland)*, 194.

Amodio, M. and Krishnaswamy, S. (2018). Magan: Aligning biological manifolds. *35th International Conference on Machine Learning, ICML 2018*, 1.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., and Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13.

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). Deepcpg: Accurate prediction of single-cell dna methylation states using deep learning. *Genome Biology*, 18.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. (2020). Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21.

Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C. A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P. J., Kelsey, G., Dean, W., Nichols, J., Stegle, O., Marioni, J. C., and Reik, W. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576.

Argelaguet, R., Cuomo, A. S., Stegle, O., and Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14.

Ashuach, T., Gabitto, M. I., Jordan, M. I., and Yosef, N. (2021). Multivi: deep generative model for the integration of multi-modal data. *bioRxiv*.

Ashuach, T., Reidenbach, D. A., Gayoso, A., and Yosef, N. (2022). Peakvi: A deep generative model for single-cell chromatin accessibility analysis. *Cell Reports Methods*, 2.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H. F., John, R. M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., and Fisher, A. G. (2006). Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, 8.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). (ln) layer normalization. *arXiv:1607.06450*.

Ballantyne, K. N., van Oorschot, R. A., Mitchell, R. J., and Koukoulas, I. (2006). Molecular crowding increases the amplification success of multiple displacement amplification and short tandem repeat genotyping. *Analytical Biochemistry*, 355.

Barnard, J., McCulloch, R., and Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10.

Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein ctcf is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98.

Bengio, Y., Courville, A., Erhan, D., Bengio, Y., and Vincent, P. (2009). Visualizing higher-layer features of a deep network visualizing higher-layer features of a deep network département d'informatique et recherche opérationnelle. *researchgate.net*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J. M., and Stuart, A. (2013). Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Betancourt, M. and Girolami, M. (2015). Hamiltonian monte carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications*.

Bishop, C. M. (2006). Pattern recognition and machine learning solutions to exercis. *Pattern Recognition*, 4.

Biswas, M., Voltz, K., Smith, J. C., and Langowski, J. (2011). Role of histone tails in structural stability of the nucleosome. *PLoS Computational Biology*, 7.

Blackwood, E. M. and Kadonaga, J. T. (1998). Going the distance: A current view of enhancer action. *Science*, 281.

Blair, R. H., Trichler, D. L., and Gaille, D. P. (2012). Mathematical and statistical modeling in cancer systems biology. *Frontiers in Physiology*, 3 JUN.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

Bochkina, N. and Richardson, S. (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics*, 63.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71.

Braak, C. J. T. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67.

Branco, M. R., Oda, M., and Reik, W. (2008). Safeguarding parental identity: Dnmt1 maintains imprints during epigenetic reprogramming in early embryogenesis. *Genes and Development*, 22.

Brockmann, R., Beyer, A., Heinisch, J. J., and Wilhelm, T. (2007). Posttranscriptional expression regulation: What determines translation rates? *PLoS Computational Biology*, 3.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., and Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361.

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357.

Cao, K., Bai, X., Hong, Y., and Wan, L. (2020). Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36.

Cedar, H. and Bergman, Y. (2009). Linking dna methylation and histone modification: Patterns and paradigms. *Nature Reviews Genetics*, 10.

Chan, T. E., Stumpf, M. P., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems*, 5.

Chappell, L., Russell, A. J., and Voet, T. (2018). Single-cell (multi)omics technologies. *Annual Review of Genomics and Human Genetics*, 19.

Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6.

Chen, S., Lake, B. B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37.

Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., and Chen, X. (2018). Umi-count modeling and differential expression analysis for single-cell rna sequencing. *Genome Biology*, 19.

Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., and Reik, W. (2018). Scnmt-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nature Communications*, 9.

Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: Powerful new methods for understanding gene regulation and cell identity. *Genome Biology*, 17.

Clark, S. J., Smallwood, S. A., Lee, H. J., Krueger, F., Reik, W., and Kelsey, G. (2017). Genome-wide base-resolution mapping of dna methylation in single cells using single-cell bisulfite sequencing (scbs-seq). *Nature Protocols*, 12.

Clarke, B., Fokoue, E., and Zhang, H. H. (2009). Principles and theory for data mining and machine learning. *Springer*.

Cohen, I., Poreba, E., Kamieniarz, K., and Schneider, R. (2011). Histone modifiers in cancer: Friends or foes? *Genes and Cancer*, 2.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452.

Colgan, D. F. and Manley, J. L. (1997). Mechanism and regulation of mrna polyadenylation. *Genes and Development*, 11.

Colomé-Tatché, M. and Theis, F. J. (2018). Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7.

Cooper, J., Ding, Y., Song, J., and Zhao, K. (2017). Genome-wide mapping of dnase i hypersensitive sites in rare cell populations using single-cell dnase sequencing. *Nature Protocols*, 12.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91.

Creighton, H. and Waddington, C. H. (1958). The strategy of the genes. *AIBS Bulletin*, 8.

Crick, F. H. C. (1958). On protein syntesis. *Symp. Soc. Exp. Biol. XII*.

Croce, L. D., Raker, V. A., Corsaro, M., Fazi, F., Fanelli, M., Faretta, M., Fuks, F., Coco, F. L., Kouzarides, T., Nervi, C., Minucci, S., and Pelicci, P. G. (2002). Methyltransferase recruitment and dna hypermethylation of target promoters by an oncogenic transcription factor. *Science*, 295.

Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M., and Drummond, D. A. (2015). Accounting for experimental noise reveals that mrna levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genetics*, 11.

Cui, Z., Chang, H., Shan, S., and Chen, X. (2014). Generalized unsupervised manifold alignment. *Advances in Neural Information Processing Systems*, 3.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., and Shendure, J. (2018). A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174.

Dahm, R. (2008). Discovering dna: Friedrich miescher and the early years of nucleic acid research. *Human Genetics*, 122.

Davie, K., Janssens, J., Koldere, D., Waegeneer, M. D., Pech, U., Łukasz Kreft, Aibar, S., Makhzami, S., Christiaens, V., González-Blas, C. B., Poovathingal, S., Hulselmans, G., Spanier, K. I., Moerman, T., Vanspauwen, B., Geurs, S., Voet, T., Lammertyn, J., Thienpont, B., Liu, S., Konstantinides, N., Fiers, M., Verstreken, P., and Aerts, S. (2018). A single-cell transcriptome atlas of the aging drosophila brain. *Cell*, 174.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, 7.

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., and Oudenaarden, A. V. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33.

Dimitriu, M. A., Lazar-Contes, I., Roszkowski, M., and Mansuy, I. M. (2022). Single-cell multiomics techniques: From conception to applications. *Frontiers in Cell and Developmental Biology*, 10.

Dolinoy, D. C., Huang, D., and Jirtle, R. L. (2007). Maternal nutrient supplementation counteracts bisphenol a-induced dna hypomethylation in early development. *Proceedings of the National Academy of Sciences of the United States of America*, 104.

Dong, Y., Su, H., Zhu, J., and Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. *arviX*.

Donnelly, D. and Rogers, E. (2005). Symplectic integrators: An introduction. *American Journal of Physics*, 73.

Doolittle, W. F. (2013). Is junk dna bunk? a critique of encode. *Proceedings of the National Academy of Sciences of the United States of America*, 110.

Dou, J., Liang, S., Mohanty, V., Cheng, X., Kim, S., Choi, J., Li, Y., Rezvani, K., Chen, R., Chen, K., and Org, K. (2020). Unbiased integration of single cell multi-omics data. *bioRxiv*.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, 195.

Duhl, D. M., Vrieling, H., Miller, K. A., Wolff, G. L., and Barsh, G. S. (1994). Neomorphic agouti mutations in obese yellow mice. *Nature Genetics*, 8.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T.,

Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E. C., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., Baren, M. J. V., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K. K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor,

J., and Lochovsky, L. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489.

Duren, Z., Chang, F., Naging, F., Xin, J., Liu, Q., and Wong, W. H. (2022). Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with screg. *Genome Biology*.

Eckersley-Maslin, M., Alda-Catalinas, C., Blotenburg, M., Kreibich, E., Krueger, C., and Reik, W. (2019). Dppa2 and dppa4 directly regulate the dux-driven zygotic transcriptional program. *Genes and Development*, 33.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30.

Eltager, M., Abdelaal, T., Mahfouz, A., and Reinders, M. J. T. (2022). scmoc: single-cell multi-omics clustering. *Bioinformatics Advances*, 2.

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell dna methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Reports*, 10.

Felsenfeld, G. and Groudine, M. (2003). Controlling the double helix. *Nature*, 421.

Forcato, M., Romano, O., and Bicciato, S. (2021). Computational methods for the integrative analysis of single-cell data. *Briefings in Bioinformatics*, 22.

Franklin, R. E. and Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual dna strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89.

Gayoso, A., Lopez, R., Steier, Z. R., Regier, J., Streets, A. M., and Yosef, N. (2019). A joint model of rna expression and surface protein abundance in single cells. *bioRxiv*.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., and Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian data analysis, third edition. *Bayesian Data Analysis, Third Edition*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian data analysis second edition. *Wiley Interdisciplinary Reviews Cognitive Science*, 1.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7.

Gershman, S. J. and Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*, 1.

Ghahramani, I. M. Z. (2005). A note on the evidence and bayesian occam's razor. *Gatsby Unit Technical Report*, 2.

Gibney, E. R. and Nolan, C. M. (2010). Epigenetics and gene expression. *Heredity*, 105.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73.

Goel, G., Chou, I. C., and Voit, E. O. (2006). Biological systems modeling and analysis: A biomolecular technique of the twenty-first century. *Journal of Biomolecular Techniques*, 17.

Goetz, J. J. and Trimarchi, J. M. (2012). Transcriptome sequencing of single cells with smart-seq. *Nature Biotechnology*, 30.

Golbabapour, S., Abdulla, M. A., and Hajrezaei, M. (2011). A concise review on epigenetic regulation: Insight into molecular mechanisms. *International Journal of Molecular Sciences*, 12.

Gong, B., Zhou, Y., and Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biology*, 22.

Granja, J. M., Klemm, S., McGinnis, L. M., Kathiria, A. S., Mezger, A., Corces, M. R., Parks, B., Gars, E., Liedtke, M., Zheng, G. X., Chang, H. Y., Majeti, R., and Greenleaf, W. J. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology 2019 37:12*, 37:1458–1465.

Greer, E. L. and Shi, Y. (2012). Histone methylation: A dynamic mark in health, disease and inheritance. *Nature Reviews Genetics*, 13.

Gu, J., Wang, X., Halakivi-Clarke, L., Clarke, R., and Xuan, J. (2014). Badge: A novel bayesian model for accurate abundance quantification and differential analysis of rna-seq data. *BMC Bioinformatics*, 15.

Guo, H., Zhu, P., Guo, F., Li, X., Wu, X., Fan, X., Wen, L., and Tang, F. (2015). Profiling dna methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nature Protocols*, 10.

Han, L., Zi, X., Garmire, L. X., Wu, Y., Weissman, S. M., Pan, X., and Fan, R. (2014). Co-detection and sequencing of genes and transcripts from the same single cells facilitated by a microfluidics platform. *Scientific Reports*, 4.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G. C., Chen, M., and Guo, G. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell*, 172.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184.

Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., and Tilghman, S. M. (2000). Ctcf mediates methylation-sensitive enhancer-blocking activity at the h19/igf2 locus. *Nature*, 405.

Hayatsu, H. (2008). The bisulfite genomic sequencing used in the analysis of epigenetic states, a technique in the emerging environmental genotoxicology research. *Mutation Research - Reviews in Mutation Research*, 659.

He, L., Davila-Velderrain, J., Sumida, T. S., Hafler, D. A., Kellis, M., and Kulminski, A. M. (2021). Nebula is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Communications Biology*, 4.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, 107.

Hendrich, B. and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-cpg binding proteins. *Molecular and Cellular Biology*, 18.

Hernando-Herraez, I., Evano, B., Stubbs, T., Commere, P. H., Bonder, M. J., Clark, S., Andrews, S., Tajbakhsh, S., and Reik, W. (2019). Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nature Communications*, 10.

Hervouet, E., Vallette, F. M., and Cartron, P. F. (2009). Dnmt3/transcription factor interactions as crucial players in targeted dna methylation. *Epigenetics*, 4.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14.

Hitchcock, D. B. (2003). A history of the metropolis–hastings algorithm. *American Statistician*, 57.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15.

Hoffman, S. L. and Collaborators (2021). Integrating scrna-seq and scatac-seq data.

Holliday, R. and Pugh, J. E. (1975). Dna modification mechanisms and gene activity during development. *Science*, 187.

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of biological chemistry*, 175.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15.

Hu, D., Garruss, A. S., Gao, X., Morgan, M. A., Cook, M., Smith, E. R., and Shilatifard, A. (2013). The mll2 branch of the compass family regulates bivalent promoters in mouse embryonic stem cells. *Nature Structural and Molecular Biology*, 20.

Hu, Q. and Greene, C. S. (2019). Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell rna transcriptomics. *Pacific Symposium on Biocomputing*, 24.

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C. Y., Xue, Z., and Fan, G. (2016). Simultaneous profiling of transcriptome and dna methylome from a single cell. *Genome Biology*, 17.

Hu, Y. G., Hirasawa, R., Hu, J. L., Hata, K., Li, C. L., Jin, Y., Chen, T., Li, E., Rigolet, M., Viegas-Péquignot, E., Sasaki, H., and Xu, G. L. (2008). Regulation of DNA methylation activity through dnmt3l promoter methylation by dnmt3 enzymes in embryonic development. *Human Molecular Genetics*, 17.

Huang, S. (2009). Reprogramming cell fates: Reconciling rarity with robustness. *BioEssays*, 31.

Ikegami, K., Ohgane, J., Tanaka, S., Yagi, S., and Shiota, K. (2009). Interplay between dna methylation, histone modification and chromatin remodeling in stem cells and during development. *International Journal of Developmental Biology*, 53.

Ivanovitch, K., Temiño, S., and Torres, M. (2017). Live imaging of heart tube development in mouse reveals alternating phases of cardiac differentiation and morphogenesis. *Elife*, 6:e30668.

Jackson, C. A., Castro, D. M., Saldi, G. A., Bonneau, R., and Gresham, D. (2020). Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments. *eLife*, 9.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3.

Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33.

Jang, S. J., Ham, M. S., Lee, J. M., Chung, S. K., Lee, H. J., Kim, J. H., Chang, H. C., Lee, J. H., and Chung, D. K. (2003). New integration vector using a cellulase gene as a screening marker for lactobacillus. *FEMS Microbiology Letters*, 224.

Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G. A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., Kim, S. B., Yang, L., Ko, M., Chen, R., Göttgens, B., Lee, J. S., Gunaratne, P., Godley, L. A., Darlington, G. J., Rao, A., Li, W., and Goodell, M. A. (2014). Large conserved domains of low dna methylation maintained by dnmt3a. *Nature Genetics*, 46.

Ji, H. and Liu, X. S. (2010). Analyzing 'omics data using hierarchical models. *Nature Biotechnology*, 28.

Jin, B., Li, Y., and Robertson, K. D. (2011). Dna methylation: Superior or subordinate in the epigenetic hierarchy? *Genes and Cancer*, 2.

Jin, T., Rehani, P., Ying, M., Huang, J., Liu, S., Roussos, P., and Wang, D. (2021). scgrnom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Medicine*, 13.

Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T. M., Childs, R., Levens, D., and Zhao, K. (2015). Genome-wide detection of dnase i hypersensitive sites in single cells and ffpe tissue samples. *Nature*, 528.

John, C. R., Watson, D., Barnes, M. R., Pitzalis, C., and Lewis, M. J. (2020). Spectrum: Fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, 36.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37.

Josling, G. A., Selvarajah, S. A., Petter, M., and Duffy, M. F. (2012). The role of bromodomain proteins in regulating gene expression. *Genes*, 3.

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., and Luo, Y. (2022). Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12.

Kapourani, C. A., Argelaguet, R., Sanguinetti, G., and Vallejos, C. A. (2021). scmet: Bayesian modeling of dna methylation heterogeneity at single-cell resolution. *Genome Biology*, 22.

Kapourani, C. A. and Sanguinetti, G. (2016). Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32.

Kapourani, C. A. and Sanguinetti, G. (2018). Bprmeth: A flexible bioconductor package for modelling methylation profiles. *Bioinformatics*, 34.

Kapourani, C. A. and Sanguinetti, G. (2019). Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biology*, 20.

Karlić, R., Chung, H. R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90.

Kass, S. U., Pruss, D., and Wolffe, A. P. (1997). How does dna methylation repress transcription? *Trends in Genetics*, 13.

Kaut, O., Schmitt, I., Tost, J., Busato, F., Liu, Y., Hofmann, P., Witt, S. H., Rietschel, M., Fröhlich, H., and Wüllner, U. (2017). Epigenome-wide dna methylation analysis in siblings and monozygotic twins discordant for sporadic parkinson's disease revealed different epigenetic patterns in peripheral blood mononuclear cells. *Neurogenetics*, 18.

Kazwini, N. E. and Sanguinetti, G. (2023). Share-topic: Bayesian inerpretable modelling of single-cell multi-omic data. *bioRxiv*, pages 2023–02.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., Guigo, R., Hubbard, T., Kent, J., Lieb, J. D., Myers, R. M., Pazin, M. J., Ren, B., Stamatoyannopoulos, J. A., Weng, Z., White, K. P., and Hardison, R. C. (2014). Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 111.

Kellum, R. and Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, 64.

Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P., and Jones, P. A. (2012). Genome-wide mapping of nucleosome positioning and dna methylation within individual dna molecules. *Genome Research*, 22.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11.

Kilgore, J. A., Hoose, S. A., Gustafson, T. L., Porter, W., and Kladde, M. P. (2007). Single-molecule and population probing of chromatin structure using dna methyltransferases. *Methods*, 41.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12.

Kobayashi, H., Koike, T., Sakashita, A., Tanaka, K., Kumamoto, S., and Kono, T. (2016). Repetitive dna methylome analysis by small-scale and single-cell shotgun bisulfite sequencing. *Genes to Cells*, 21.

Koller, D. and Friedman, N. (2009). Introduction - probabilistic graphical models: principles and techniques. *Probabilistic graphical models: principles and techniques*.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular Cell*, 58.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20.

Krueger, F. and Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27.

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). Dna methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9.

Kumar, A., Rai, P., and Daumé, H. (2011). Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y. C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K. H., Feizi, S., Karlic, R., Kim, A. R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., Jager, P. L. D., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L. H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518.

Lambert, B. (2017). A student's guide to bayesian statistics. *Journal of Chemical Information and Modeling*, 53.

Lander, E. S. (1996). The new genomics: Global views of biology. *Science*, 274.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm,

L., Rubenfield, M., Weinstock, K., Hong, M. L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., Bastide, M. D. L., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409.

Latchman, D. S. (1997). Transcription factors: An overview. *International Journal of Biochemistry and Cell Biology*, 29.

Lee, J. T. (2011). Gracefully ageing at 50, x-chromosome inactivation becomes a paradigm for rna and chromatin control. *Nature Reviews Molecular Cell Biology*, 12.

Leontiou, C. A., Hadjidaniel, M. D., Mina, P., Antoniou, P., Ioannides, M., and Patsalis, P. C. (2015). Bisulfite conversion of dna: Performance comparison of different kits and methylation quantitation of epigenetic biomarkers that have the potential to be used in non-invasive prenatal testing. *PLoS ONE*, 10.

Levene, P. (1919). The structure of yeast nucleic acid. *Journal of Biological Chemistry*, 40.

Lindsey, J. K. (1997). Applying generalized linear models. *Springer*, 1.

Linnarsson, S. and Teichmann, S. A. (2016). Single-cell genomics: Coming of age. *Genome Biology*, 17.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, 462.

Liu, H., Zhang, Z., and Grimm, K. J. (2016). Comparison of inverse wishart and separation-strategy priors for bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling*, 23.

Liu, J., Huang, Y., Singh, R., Vert, J. P., and Noble, W. S. (2019). Jointly embedding multiple single-cell omics measurements. *Leibniz International Proceedings in Informatics, LIPIcs*, 143.

Liu, J. S. (2004). Monte carlo strategies in scientific computing. *Springer New York*, 10.

Loan, C. F. V. (1976). Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13.

Lopez, R., Nazaret, A., Langevin, M., Samaran, J., Regier, J., Jordan, M. I., and Yosef, N. (2019). A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269*.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15.

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., Žiga Avsec, Gayoso, A., Yosef, N., Interlandi, M., Rybakov, S., Misharin, A. V., and Theis, F. J. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15.

Lu, Y., Chan, Y. T., Tan, H. Y., Li, S., Wang, N., and Feng, Y. (2020). Epigenetic regulation in human cancer: The potential role of epi-drug in cancer therapy. *Molecular Cancer*, 19.

Lun, A. T., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-Seq data with bioconductor. *F1000Research*, 5.

Luo, C., Liu, H., Xie, F., Armand, E., Siletti, K., Bakken, T., Fang, R., Doyle, W., Hodge, R., Hu, L., Wang, B.-A., Zhang, Z., Preissl, S., Lee, D.-S., Zhou, J., Niu, S.-Y., Castanon, R., Bartlett, A., Rivkin, A., Wang, X., Lucero, J., Nery, J., Davis, D., Mash, D., Dixon, J., Linnarsson, S., Lein, E., Behrens, M. M., Ren, B., Mukamel, E., and Ecker, J. (2019). Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv*.

Luo, C., Rivkin, A., Zhou, J., Sandoval, J. P., Kurihara, L., Lucero, J., Castanon, R., Nery, J. R., Pinto-Duarte, A., Bui, B., Fitzpatrick, C., O'Connor, C., Ruga, S., Eden, M. E. V., Davis, D. A., Mash, D. C., Behrens, M. M., and Ecker, J. R. (2018). Robust single-cell dna methylome profiling with snmc-seq2. *Nature Communications*, 9.

Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020a). Integrative methods and practical challenges for single-cell multi-omics. *Trends in Biotechnology*, 38.

Ma, Q. and Xu, D. (2022). Deep learning shapes single-cell data analysis. *Nature Reviews Molecular Cell Biology*, 23.

Ma, S., Revenga, M. D. L. F., Sun, Z., Sun, C., Murphy, T. W., Xie, H., González-Maeso, J., and Lu, C. (2018). Cell-type-specific brain methylomes profiled via ultralow-input microfluidics. *Nature Biomedical Engineering*, 2.

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y. C., Regev, A., and Buenrostro, J. D. (2020b). Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Aa, N. V. D., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&t-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12.

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-cell multiomics: Multiple measurements from single cells. *Trends in Genetics*, 33.

Macaulay, I. C., Teng, M. J., Haerty, W., Kumar, P., Ponting, C. P., and Voet, T. (2016). Separation and parallel sequencing of the genomes and transcriptomes of single cells using g&t-seq. *Nature Protocols*, 11.

MacKay, D. J. C. (2004). Information theory, inference, and learning algorithms. *Robotica*, 22.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mrna and protein in complex biological samples. *FEBS Letters*, 583.

Maloy, S. and Hughes, K. (2013). Brenner's encyclopedia of genetics: Second edition. *Brenner's Encyclopedia of Genetics: Second Edition*.

Maniatis, C., Vallejos, C., and Sanguinetti, G. (2022). Scrapl: A bayesian hierarchical framework for detecting technical associates in single cell multiomics data. *PLoS Computational Biology*.

Mann, M. and Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24.

Martin, C. and Zhang, Y. (2005). The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Biology*, 6.

Maxam, A. M. and Gilbert, W. (1980). [57] sequencing end-labeled dna with base-specific chemical cleavages. *Methods in Enzymology*, 65.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic Acids Research*, 33.

Meyer, A. and de Peer, Y. V. (2003). 'natural selection merely modified while redundancy created' - susumu ohno's idea of the evolutionary importance of gene and genome duplications. *Journal of Structural and Functional Genomics*, 3.

Michael, E. and Timothy, S. (2000). *Approximating integrals via Monte Carlo and deterministic methods*, volume 20. OUP Oxford.

Minoura, K., Abe, K., Nam, H., Nishikawa, H., and Shimamura, T. (2021). scmm: Mixture-of-experts multimodal deep generative model for single-cell multiomics data analysis. *SSRN Electronic Journal*.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, 40.

Miura, G. I., Froelick, G. J., Marsh, D. J., Stark, K. L., and Palmiter, R. D. (2003). The d subunit of the vacuolar atpase (atp6d) is essential for embryonic development. *Transgenic Research*, 12.

Mobley, A. S. (2019). Neural stem cells and adult neurogenesis. *Neural Stem Cells and Adult Neurogenesis*.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 30.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K. R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., and Cocho, J. A. (2013). A glimpse into past, present, and future dna sequencing. *Molecular Genetics and Metabolism*, 110.

Moudgil, A., Wilkinson, M. N., Chen, X., He, J., Cammack, A. J., Vasek, M. J., Lagunas, T., Qi, Z., Lalli, M. A., Guo, C., Morris, S. A., Dougherty, J. D., and Mitra, R. D. (2020). Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells. *Cell*, 182.

Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., Sinnamon, J. R., Shendure, J., Trapnell, C., O'Roak, B. J., Xia, Z., Steemers, F. J., and Adey, A. C. (2018). Highly scalable generation of dna methylation profiles in single cells. *Nature Biotechnology*, 36.

Murphy, K. P. (2012). Machine learning - a probabilistic perspective - table-of-contents. *The MIT Press*.

Muskovic, W. and Powell, J. E. (2021). Dropletqc: improved identification of empty droplets and damaged cells in single-cell rna-seq data. *Genome Biology*, 22.

Muto, Y., Wilson, P. C., Ledru, N., Wu, H., Dimke, H., Waikar, S. S., and Humphreys, B. D. (2021). Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature Communications*, 12.

Navis, A. R. (2012). Conrad hal waddington (1905-1975). *Embryo Project Encyclopedia*.

Neri, F., Krepelova, A., Incarnato, D., Maldotti, M., Parlato, C., Galvagni, F., Matarese, F., Stunnenberg, H. G., and Oliviero, S. (2013). Xdnmt3l antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in escs. *Cell*, 155.

Newell-Price, J., Clark, A. J., and King, P. (2000). Dna methylation and silencing of gene expression. *Trends in Endocrinology and Metabolism*, 11.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5.

Nikolov, D. B. and Burley, S. K. (1997). Rna polymerase ii transcription initiation: A structural view. *Proceedings of the National Academy of Sciences of the United States of America*, 94.

Ohlsson, R., Renkawitz, R., and Lobanenkov, V. (2001). Ctcf is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics*, 17.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: Five essential questions. *Nature Reviews Genetics*, 14.

Philpott, M., Cribbs, A. P., Brown, T., and Oppermann, U. (2020). Advances and challenges in epigenomic single-cell sequencing applications. *Current Opinion in Chemical Biology*, 57.

Picelli, S., Åsa K. Björklund, Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10.

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C., Ho, D. L. L., Reik, W., Srinivas, S., Simons, B. D., Nichols, J., Marioni, J. C., and Göttgens, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566.

Ponting, C. P. and Hardison, R. C. (2011). What fraction of the human genome is functional? *Genome Research*, 21.

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *eLife*, 6.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155.

Prokhortchouk, E. and Defossez, P. A. (2008). The cell biology of dna methylation in mammals. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1783.

Qiu, C. and Shendure, J. (2021). A continuous model of early mammalian development. *Nature*, 593.

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., Noble, W. S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell hi-c. *Nature Methods*, 14.

Rautenstrauch, P., Vlot, A. H. C., Saran, S., and Ohler, U. (2022). Intricacies of single-cell multi-omics data integration. *Trends in Genetics*, 38.

Razin, A. (1998). Cpg methylation, chromatin structure and gene silencing - a three-way connection. *EMBO Journal*, 17.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J., Theis, F. J., Uhlen, M., Oudenaarden, A. V., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., and Yosef, N. (2017). The human cell atlas. *eLife*, 6.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, 293.

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, 58.

Reyes, M., Billman, K., Hacohen, N., and Blainey, P. C. (2019). Simultaneous profiling of gene expression and chromatin accessibility in single cells. *Advanced Biosystems*, 3.

Richards, J. E. and Hawley, R. S. (2011). The central dogma of molecular biology. *The Human Genome*.

Riggs, A. D. (1975). X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43.

Robert, C. P. (2007). The bayesian choice: From decision-theoretic foundations to computational implementation. *Elements*, 91.

Robert, C. P. and Casella, G. (2004). Monte carlo statistical methods. *Springer*.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26.

Robinson, S. R. and Dreher, B. (1990). The visual pathways of eutherian mammals and marsupials develop according to a common timetable. *Brain, Behavior and Evolution*, 36.

Rooijers, K., Markodimitraki, C. M., Rang, F. J., de Vries, S. S., Chialastri, A., de Luca, K. L., Mooijman, D., Dey, S. S., and Kind, J. (2019). Simultaneous quantification of protein–dna contacts and transcriptomes in single cells. *Nature Biotechnology*, 37.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33.

Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., and Teichmann, S. A. (2017). The human cell atlas: From vision to reality. *Nature*, 550.

Sadakierska-Chudy, A. and Filip, M. (2015). A comprehensive view of the epigenetic landscape. part ii: Histone post-translational modification, nucleosome level, and chromatin regulation by ncrnas. *Neurotoxicity Research*, 27.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74.

Sass, S., Buettner, F., Mueller, N. S., and Theis, F. J. (2013). A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Research*, 41.

Sawan, C. and Herceg, Z. (2010). Histone modifications and cancer. *Advances in Genetics*, 70.

Schmidt, M. W., Houseman, A., Ivanov, A. R., and Wolf, D. A. (2007). Comparative proteomic and transcriptomic profiling of the fission yeast schizosaccharomyces pombe. *Molecular Systems Biology*, 3.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473.

Schwartzman, O. and Tanay, A. (2015). Single-cell epigenomics: Techniques and emerging applications. *Nature Reviews Genetics*, 16.

Schwarz, G. (1978). "estimating the dimension of a model.". *The Annals of Statistics*, 6.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14.

Shi, Y., Siddharth, N., Paige, B., and Torr, P. H. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32.

Shim, S., Stemke-Hale, K., Tsimberidou, A. M., Noshari, J., Anderson, T. E., and Gascoyne, P. R. (2013). Antibody-independent isolation of circulating tumor cells by continuous-flow dielectrophoresis. *Biomicrofluidics*, 7.

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, 122.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18.

Soneson, C. and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and der Linde, A. V. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64.

Stanojevic, S., Li, Y., Ristivojevic, A., and Garmire, L. X. (2022). Computational methods for single-cell multi-omics integration and alignment. *Genomics, Proteomics & Bioinformatics*, 20:836–849.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177.

Svensson, V. (2020). Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38.

Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. (2017). Power analysis of single-cell rna-sequencing experiments. *Nature Methods*, 14.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell rna-seq in the past decade. *Nature Protocols*, 13.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016). How to train deep variational autoencoders and probabilistic ladder networks. *Advances in Neural Information Processing Systems*, 0.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6.

Tarabay, Y., Kieffer, E., Teletin, M., Celebi, C., Montfoort, A. V., Zamudio, N., Achour, M., Ramy, R. E., Gazdag, E., Tropel, P., Mark, M., Bourc'His, D., and Viville, S. (2013). The mammalian-specific tex19.1 gene plays an essential role in spermatogenesis and placenta-supported development. *Human Reproduction*, 28.

Tilghman, S. M. (1999). The sins of the fathers and mothers: Genomic imprinting in mammalian development. *Cell*, 96.

Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20.

Tyser, R. C., Miranda, A. M., Chen, C.-m., Davidson, S. M., Srinivas, S., and Riley, P. R. (2016). Calcium handling precedes cardiac differentiation to initiate the first heartbeat. *Elife*, 5:e17113.

V, D. J., Ian, L., Dustin, T., Eugene, B., Srinivas, V., Dave, M., Brian, P., Alex, A., Matt, H., and A, S. R. (2017). Tensorflow distributions. *arXiv preprint arXiv:1711.10604*.

Vaissière, T., Sawan, C., and Herceg, Z. (2008). Epigenetic interplay between histone modifications and dna methylation in gene silencing. *Mutation Research - Reviews in Mutation Research*, 659.

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). Basics: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology*, 11.

Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016). Beyond comparisons of means: Understanding changes in gene expression at the single-cell level. *Genome Biology*, 17.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1.

Vandereyken, K., Sifrim, A., Thienpont, B., and Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M. H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q.,

Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291.

Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13.

Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., and Pawitan, Y. (2016). Beta-poisson model for single-cell rna-seq data analyses. *Bioinformatics*, 32.

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, 150.

Waddington, C. H. (2012). The epigenotype. 1942. *International journal of epidemiology*, 41.

Wade, P. A. (2001). Methyl cpg binding proteins: Coupling chromatin architecture to gene regulation. *Oncogene*, 20.

Wang, J., Xu, L., Wang, E., and Huang, S. (2010). The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophysical Journal*, 99.

Wang, K., Li, X., Dong, S., Liang, J., Mao, F., Zeng, C., Wu, H., Wu, J., Cai, W., and Sun, Z. S. (2015). Q-rrbs: A quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses. *Epigenetics*, 10.

Wang, X., Sun, Z., Zhang, Y., Xu, Z., Xin, H., Huang, H., Duerr, R. H., Chen, K., Ding, Y., and Chen, W. (2020). Brem-sc: A bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic Acids Research*, 48.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11.

Watson, J. D., Baker, T. A., Bell, S. P., Grann, A., Levine, M., and Losick, R. (2015). Molecular biology of the gene (7th edition). *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York*.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171.

Watt, F. and Molloy, P. L. (1988). Cytosine methylation prevents binding to dna of a hela cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & development*, 2.

Wei, Z., Yang, A., Rocha, L., Miranda, M. F., and Nathoo, F. S. (2022). A review of bayesian hypothesis testing and its practical implementations. *Entropy*, 24.

Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., Guo, Y. E., Hnisz, D., Jaenisch, R., Bradner, J. E., Gray, N. S., and Young, R. A. (2017). Yy1 is a structural regulator of enhancer-promoter loops. *Cell*, 171.

Welch, J., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. (2018). Integrative inference of brain cell similarities and differences from single-cell genomics. *bioRxiv*.

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2017). Matcher: Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology*, 18.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8.

Wu, M. and Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 2018-December.

Yang, L., Liu, J., Lu, Q., Riggs, A. D., and Wu, X. (2017). Saic: An iterative clustering approach for analysis of single cell rna-seq data. *BMC Genomics*, 18.

Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Yes, but did it work?: Evaluating variational inference. *35th International Conference on Machine Learning, ICML 2018*, 12.

Yuan, M., Chen, L., and Deng, M. (2022). Clustering single-cell multi-omics data with MoClust. *Bioinformatics*, 39(1). btac736.

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: Simulation of single-cell rna sequencing data. *Genome Biology*, 18.

Zerbino, D. R. and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18.

Zhang, Z., Cui, F., Lin, C., Zhao, L., Wang, C., and Zou, Q. (2021). Critical downstream analysis steps for single-cell rna sequencing data. *Briefings in Bioinformatics*, 22.

Zhao, Z. and Shilatifard, A. (2019). Epigenetic modifications of histones in cancer. *Genome Biology*, 20.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8.

Zhou, J. X. and Huang, S. (2011). Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends in Genetics*, 27.

Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M. M., Hu, M., and Ren, B. (2019). An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural and Molecular Biology*, 26.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of single-cell rna sequencing methods. *Molecular Cell*, 65.

Zuo, C. and Chen, L. (2021). Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings in Bioinformatics*, 22.