Methods for Analyzing Sparse Genetic and Epigenetic Data: Single Cells to Population Levels

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029774">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029774</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Methods for Analyzing Sparse Genetic and Epigenetic Data: Single Cells to Population Levels

A dissertation presented

by

Divy S. Kangeyan

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University
Cambridge, Massachusetts

May 2019
Abstract

This dissertation work is motivated by the large influx of sequencing data: that is, both in terms of the amount and the type of data, where current statistical and computational methods are inadequate in addressing the data manipulation and hence the corresponding scientific questions of interest.

In Chapter 1, we address a current issue regarding a data analysis platform to conduct large amount of Next Generation Sequencing based methylation data. Bisulfite sequencing allows base-pair resolution DNA methylation and has recently been adapted for use in single cells. We present a set of preprocessing pipelines that allow users to ensure 1) reproducibility, 2) scalability, 3) integration with publicly available data, and 4) access to best-practice methods. The workflows produce output for visualization and further downstream analysis. Optional use of cloud computing resources facilitates analysis of large datasets, and integration with existing methylation data.

In Chapter 2, we focus our attention on sparsity in single-cell DNA methylation data. Single-cell DNA methylation analysis has the potential to produce high resolution methylation landscape and elucidate the heterogeneity in methylation. But it suffers from low coverage due to the low quantity of input DNA. We find that on average, only about 5-10% of CpGs are observed in typical single-cell libraries. We show how missingness of methylation status can bias metrics such as mean methylation estimates and clustering analyses. We propose a joint analysis approach that leverages bulk sequencing data, to infer bias-corrected single-cell methylation status.

In Chapter 3, we consider sparsity in the rare variant data and how it can be utilized to infer population structure. Population-substructure in genetic studies is often assessed by principal component analysis of genetic relatedness matrices (GRM). With the general availability of whole-genome sequencing (WGS) platforms, rare variant data
are now widely available. As such data are genetically younger than common variants, they should enable for a fine-scale assessment of the substructure. Here, using the 1,000 genomes project data, we compare the features of Jaccard-based GRMs with standard approaches that utilizes the genetic covariance matrix, with respect to their ability to examine and infer fine-scale population substructure.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title page</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>viii</td>
</tr>
<tr>
<td>1 A (Fire)Cloud-Based DNA Methylation Data Preprocessing and Quality Control Platform</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Methylation assays</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Preprocessing bisulfite sequencing data</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Implementation</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Results &amp; Discussion</td>
<td>4</td>
</tr>
<tr>
<td>1.3.1 Reproducibility</td>
<td>4</td>
</tr>
<tr>
<td>1.3.2 Scalability</td>
<td>5</td>
</tr>
<tr>
<td>1.3.3 Integration with existing datasets</td>
<td>5</td>
</tr>
<tr>
<td>1.3.4 Workflow &amp; output</td>
<td>6</td>
</tr>
<tr>
<td>1.4 TCGA Data Analysis</td>
<td>9</td>
</tr>
<tr>
<td>1.5 Conclusion</td>
<td>11</td>
</tr>
<tr>
<td>2 Modeling missingness in single-cell DNA methylation data</td>
<td>14</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Methods</td>
<td>16</td>
</tr>
</tbody>
</table>
2.2.1 Sparsity of single-cell methylation data ....................................... 17
2.2.2 Ground truth data ......................................................................... 17
2.2.3 Observation rate plots .................................................................... 18
2.2.4 Correction model ........................................................................... 19
2.2.5 Application of the model in real data setting .................................. 20
2.2.6 Distance correction ......................................................................... 21
2.2.7 Extension of single-cell methods for bulk methylation samples ...... 22
2.3 Results ............................................................................................... 23
  2.3.1 Sparsity in single-cell data ................................................................. 23
  2.3.2 Distance correction ........................................................................... 27
  2.3.3 Extension of single-cell methods for bulk methylation samples ...... 28
2.4 Discussion .......................................................................................... 29
2.5 Conclusion .......................................................................................... 31

3 Delineating finer population-substructure with rare variants .......... 32
  3.1 Introduction ........................................................................................ 33
  3.2 Methods .............................................................................................. 35
    3.2.1 Quantitative measures to analyze rare variant data ....................... 35
    3.2.2 1000 Genomes Project (TGP) data ............................................... 39
    3.2.3 Data analysis .................................................................................. 39
  3.3 Results ................................................................................................. 40
  3.4 Discussion ........................................................................................... 43
  3.5 Conclusion .......................................................................................... 46

4 Supplementary materials ...................................................................... 47
  4.1 A (Fire)Cloud-Based DNA Methylation Data Preprocessing and Quality Control Platform .......................................................... 48
    4.1.1 Quality Control metrics ................................................................. 48
  4.2 Modeling missingness in single-cell DNA methylation data .......... 57
    4.2.1 Methylation based on ground truth ............................................... 57
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.2 Observation rate plot within a library</td>
<td>57</td>
</tr>
<tr>
<td>4.2.3 Distance correction results in single-cell DNA methylation data</td>
<td>59</td>
</tr>
<tr>
<td>4.2.4 Correction in the bulk samples</td>
<td>59</td>
</tr>
<tr>
<td>4.3 Delineating finer population-substructure with rare variants</td>
<td>60</td>
</tr>
<tr>
<td>4.3.1 PCA visualization of population structure</td>
<td>60</td>
</tr>
<tr>
<td>4.3.2 Cluster affinity among all sub-populations</td>
<td>60</td>
</tr>
</tbody>
</table>

References 65
Acknowledgments

First, I would like to thank my advisor, Dr. Martin Aryee, for your tireless efforts guiding me through my PhD career. I appreciate your endless patience in discussing obstacles in my projects and striving to deliver good science. I have learned so much from you as a mentor and a scholar.

I would also like to thank Dr. Christoph Lange, for advising me on my third project and providing insightful suggestions as a committee member. In addition, I want to acknowledge Dr. Jeff Miller and Dr. Guo-Cheng Yuan, for your helpful feedback and honest advice. It was a privilege having you all as my PhD dissertation committee members.

I would not have survived the PhD without the help and support of my fellow students at the Department of Biostatistics, especially students in my cohort. Thank you so much for being part of my life as a graduate student and I am proud to have you in my network.

I am forever grateful for my mother (amma), Kirubasakthy Swargaloganathan for providing me and my brothers a better life by sacrificing hers. I also want to thank my brothers and family, it was not easy being an immigrant in this country but I always look up to you to see how much we have achieved and how much more we can achieve.

Finally, I want to thank all my teachers and mentors throughout my life who realized my talents and encouraged me to grow and become a scholar and a scientist.
A (Fire)Cloud-Based DNA Methylation Data Preprocessing and Quality Control Platform

Divy S. Kangeyan
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Martin Aryee
Department of Biostatistics
Harvard T.H. Chan School of Public Health
1.1 Introduction

DNA methylation is an essential component of the epigenetic machinery that regulates gene expression. It involves a chemical modification whereby a methyl group is added to cytosine bases (Suzuki and Bird, 2008). In general, DNA methylation is associated with transcriptional repression. There has been evidence indicating that DNA methylation leads to transcriptional repression by directly interfering with transcription factor binding; a second hypothesis also proposed which says that protein that binds to the modified DNA blocks access to other transcription factors binding and thereby repression transcription. DNA methylation is highly dynamic during development and is also associated with a range of diseases including cancer, autoimmune and neurodegenerative disorders (Okano et al., 1999; Messerschmidt et al., 2014; Baylin and Jones, 2011; Al-Mahdawi et al., 2016).

![Figure 1.1: Methylation usually occurs in CpG sites in the genome and it is facilitated by DNA Methyltransferase enzymes. CpG sites are not uniformly distributed in the genome, there are features that contain large number of CpG sites in smaller genomic regions known as CpG Island. Modified from Jueligeret.al. (2018)](image)

1.1.1 Methylation assays

Several techniques are used to measure the DNA methylation such as methylation specific PCR, methylation arrays, MALDI-TOF and bisulfite sequencing. The gold-standard assays for DNA methylation are based on bisulfite sequencing, where unmethylated cytosines (C) are selectively and efficiently converted to thymines (T) allowing base-pair resolution methylation state to be read out by standard high-throughput sequencing (Frommer et al., 1992). Bisulfite sequencing can be applied to a whole genome library (WGBS).
(Ziller et al., 2014), or in targeted variants that include Reduced Representation Bisulfite Sequencing (RRBS) (Meissner et al., 2005) that enriches for CpG islands, or Hybrid Selection Bisulfite Sequencing (HSBS) (Ziller et al., 2016) that uses capture probes to target a specific set of genomic regions of interest.

Main advantage of WGBS is that it covers the entire genome at a single base resolution. However, WGBS is expensive and if methylation landscape of the entire genome is not of interest, WGBS might not be the optimal approach. In RRBS protocol one or more restriction enzymes will cut the fragment such that both ends of it have a CG in it regardless of their methylation status. Then these fragments are size-selected in order to enrich the number of CpGs observed. Therefore, RRBS protocol reduces the cost of sequencing by targeting CpG rich regions such as CpG islands and promoters. However, as a trade-off little information on CpGs that are in CpG poor regions is obtained. HSBS is mostly used to capture dynamic methylation in differentially methylated regulatory elements.

1.1.2 Preprocessing bisulfite sequencing data

Preprocessing and quality control typically comprise the most computationally intensive portion of bisulfite sequencing data analysis. Due to the large size of raw datasets which may contain >100GB of data for deeply sequenced individual samples, or thousands of cells in single-cell projects (Smallwood et al., 2014). Here we present a set of preprocessing tools for bisulfite sequencing data that facilitate analyses of such datasets, by simplifying the use of large computational compute clusters. We also introduce a new R/Bioconductor package, smeth (Kangeyan, 2018b), that is optimized for QC analysis of large datasets. The pipelines can also be run locally, or on cloud computing infrastructure, providing practically unlimited scalability without requiring local compute resources. The Cloud implementation, in particular, is accessible through a web browser interface and lends itself to both researchers who have technical expertise and to those biologist with limited bioinformatics analysis experience.
1.2 Implementation

The pipelines are designed to go from raw sequencing data to CpG-level methylation estimates. The workflows first perform read alignment and methylation calling in parallel across samples, followed by an aggregation and quality control analysis step. The workflows are implemented in the WDL workflow description language (Voss, 2017; Broad Institute, 2018) and use software packaged into Docker (Docker, Inc., 2018) containers. WDL files are structured such that it contains one or more workflows and within workflows several tasks can exist. Both workflows and tasks will contain input and runtime parameters, and output files. Docker containers are lightweight virtual machines that encapsulate the entire software environment required by the pipeline tools, including their dependencies. In addition to the option of running the WDL workflows locally on a single computer or on an HPC (High-Performance Computing) cluster using job management systems such as LSF (Load Sharing Facility) or SLURM (Simple Linux Utility for Resource Management), we also provide an implementation that is available through the Google Cloud-based FireCloud platform https://software.broadinstitute.org/firecloud/ (Birger et al., 2017). FireCloud is accessible through a web-browser and allows execution of WDL-based workflows on cloud compute resources with scalability that is unlimited for most practical use cases. The scmeth (Kangeyan, 2018b) package used for QC analysis is part of the R/Bioconductor project.

1.3 Results & Discussion

1.3.1 Reproducibility

To guarantee reproducible analyses, we take advantage of two components: First, we use a workflow description language, WDL, that can be executed without modifications on systems ranging from a laptop, to an HPC cluster, to cloud compute resources. This flexibility is provided by the workflow engine, Cromwell (Birger et al., 2017), which has various “back-ends” allowing it to execute workflow tasks on the various platforms. Second, we use Docker containers, lightweight virtual machines, that package the full software
environment required by the pipeline tools. These two components together ensure that identical results are produced across different platforms and across multiple runs of the pipelines.

Table 1.1: Run time and cost estimates for different number of samples based on FireCloud

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Per-sample preprocessing (Hours / $)</th>
<th>Aggregation and QC (Hours / $)</th>
<th>Total (Hours / $)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.98 ($0.93)</td>
<td>0.97 ($0.28)</td>
<td>1.95 ($1.21)</td>
</tr>
<tr>
<td>100</td>
<td>1.47 ($8.99)</td>
<td>6.00 ($0.86)</td>
<td>7.47 ($9.85)</td>
</tr>
<tr>
<td>1000</td>
<td>4.48 ($52.48)</td>
<td>58.01 ($13.74)</td>
<td>62.49 ($66.22)</td>
</tr>
</tbody>
</table>

1.3.2 Scalability

Scalability is achieved through parallelization across samples. For users with an HPC cluster that supports Docker containers, this parallelization is accessible locally. Alternatively, any user can take advantage of the Firecloud platform that uses the Google Compute Cloud as the computing platform. The Google billing model charges per rented minute per machine, which enables all per-sample preprocessing to be performed within a near-fixed total time, regardless of the number of samples, as all samples can be processed in parallel. There are no charges for using the Firecloud platform itself although the user will accrue compute and storage costs billed by Google for resources used in workflow execution (Broad Institute, 2017).

1.3.3 Integration with existing datasets

When analyzing a new dataset, it is often useful to compare the new samples to public data, either from individual published studies or large consortia like TCGA (Weinstein et al., 2013) and TARGET (National Cancer Institute, 2019). These data are often not directly comparable, however, due to differences in preprocessing and other upstream analysis. Applying a uniform processing pipeline is, on the other hand, challenging due to the size of the datasets (e.g. TCGA) making them difficult to download and process. As Firecloud already hosts raw TCGA data, an alternative is to take advantage of our DNA methylation workflow to process both TCGA and the users own data in a uniform manner.
on this platform. The preprocessed data, which is much smaller than the raw sequencing data, can then either be further analyzed using cloud resources, or downloaded for local downstream analysis.

1.3.4 Workflow & output

The methylation workflows follow a two step pattern, with a parallelized per-sample preprocessing step followed by an aggregation and QC step that integrates data across the dataset. Following initial preprocessing with the pipeline default bisulfite-aware aligner Bismark ([Krueger and Andrews, 2011](#)), the following outputs are generated for each input sample:

(i) BAM and BAM index files

(ii) a per-CpG coverage file with unmethylated and methylated read counts

(iii) a bigwig file for visualization, and

(iv) a set of quality assessment metrics such as fraction of aligned reads, bisulfite conversion rate and methylation value distributions.

The aggregation step then prepares the individual sample outputs for downstream analysis by combining them into coverage and methylation matrices, available either as plain text or as a R/Bioconductor bsseq ([Hansen et al., 2012](#)) object that is also annotated with metrics including number of reads, number of covered CpGs and bisulfite conversion rate.

In addition to preprocessed methylation data, comprehensive HTML and plain text quality reports are also generated using tools implemented in the scmeth ([Kangeyan, 2018b](#)) Bioconductor package. The QC report can be used to identify low quality batches or samples, and provides metrics, including number of reads total CpG coverage, bisulfite conversion rate, methylation distribution, genomic feature coverage (e.g. promoters, enhancers), downsampling saturation curve and methylation distributions (Table 1). In order to scale to large sample sizes as is common in single-cell analysis, an on-disk representation of the methylation and coverage matrices as implemented in the bsseq ([Hansen et al., 2012](#)) package is used by default. In order to improve QC analysis run time for large datasets, scmeth ([Kangeyan, 2018b](#)) provides an option to subsample while calculat-
ing metrics. We find that estimates based on using as few as one million of the 28 million CpGs in the human genome are unbiased and stable.
Figure 1.2: End-to-end methylation analysis workflow and some main QC metrics
Table 1.2: Quality control metrics and their interpretation in the context of single-cell methylation

<table>
<thead>
<tr>
<th>QC Metric</th>
<th>Information gained from this metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read metrics</td>
<td>Number of mapped and unmapped reads</td>
</tr>
<tr>
<td>CpG Coverage</td>
<td>Number of CpGs observed with a minimum coverage threshold</td>
</tr>
<tr>
<td>M-bias</td>
<td>Average methylation by position across reads. Deviation from uniformity typically indicates a problem with library construction or data preprocessing.</td>
</tr>
<tr>
<td>Downsampling saturation curve</td>
<td>CpG coverage as a function of number of reads. A rising curve indicates that we would expect to observe additional CpGs from deeper sequencing of a library.</td>
</tr>
<tr>
<td>CpG discretization</td>
<td>Useful in single-cell analysis, this represents the fraction of CpGs with non-binary methylation status</td>
</tr>
<tr>
<td>Feature level coverage</td>
<td>The fraction of key genomic features (e.g. promoters, CpG Islands), covered with at least 1 CpG</td>
</tr>
<tr>
<td>Bisulfite conversion rate</td>
<td>The proportion of non CpG context Cs that were converted to T. This should be close to 100% in most mammalian tissues</td>
</tr>
<tr>
<td>CpG density distribution</td>
<td>The CpG density distribution around observed CpGs is typically similar across samples indicating coverage of similar genomic regions</td>
</tr>
<tr>
<td>Methylation distribution</td>
<td>Sample-to-sample deviations in the distribution of methylation values can indicate technical artifacts</td>
</tr>
</tbody>
</table>

We used 1000 single-cell RRBS samples with a median of 872,223 reads (range of 5,437 to 4,165,149) to calculate the running time and cost estimates for the workflows. For example, processing the full set of 1000 samples using default options took 62 hours and accrued $66 of Google Cloud charges. (Table 1.1)

1.4 TCGA Data Analysis

We have preprocessed and made available the 47 WGBS samples available from TCGA. These samples were sequenced with a median of 361,777,141 reads (range of 289,476,432 to 955,974,014). We confirmed a high concordance in methylation estimates with the avail-
able BEDgraph files from the NCI Genomic Data Commons (GDC), with a correlation of 0.99 when considering CpGs with a minimum read coverage 10. The raw (FASTQ) data, processed data and workflows are made available in a FireCloud workspace. Raw data can be found in the following workspace aryee-merkin/TCGA_WGBS_hg19 (https://portal.firecloud.org/#workspaces/aryee-merkin/TCGA_WGBS_hg19) which requires TCGA access privilege and processes data can be found in aryee-lab/TCGA_WGBS_hg19_OpenAccess (https://portal.firecloud.org/#workspaces/aryee-lab/TCGA_WGBS_hg19_OpenAccess) which is publicly accessible. We have also made the processed data available via tcgaWGBSData.hg19, an experimental data package in bioconductor (Kangeyan, 2018a).

Figure 1.3: Correlation between methylation in publicly available TCGA data and the methylation obtained from FireCloud processing.
The workflows are pre-configured with the quantity of compute resources (e.g. memory and number of CPU cores) to request from either an HPC system (e.g. LSF) or the cloud environment for each analysis step, but these can be altered by the user if a different trade-off between run time and cost is desired (Birger et al., 2017).

![Figure 1.4: dna-methylation workspace in FireCoud](image)

1.5 Conclusion

We have developed a set of preprocessing and quality assessment pipelines for Bisulfite sequencing-based DNA Methylation analysis. By leveraging Docker containers and a workflow language that can be executed both locally and in the cloud, the pipelines produce reproducible output across different platforms and user environments. This also has the benefit of facilitating comparisons across datasets such as between local user data and data from public repositories (e.g. TCGA) as identical preprocessing can be guaranteed. We have also introduced the `scmeth` (Kangeyan, 2018b) R/Bioconductor package that implements QC functions optimized for large methylation datasets, such as those common
in single-cell analyses. We take advantage of the pipelines portability by providing an implementation in the Google Cloud-based FireCloud platform, which enables any user the ability to scale to very large datasets without local compute capacity restraints. We believe that these tools will be useful as the scale of DNA methylation datasets grow, and that they will serve as a template for tools for other types of large genomic data.
Modeling missingness in single-cell DNA methylation data

Divy S. Kangeyan
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Martin Aryee
Department of Biostatistics
Harvard T.H. Chan School of Public Health
2.1 Introduction

Large influx of genomic data in the past decade has led researchers to build efficient tools to process them and obtain manageable size dataset to gain biological insights. In the previous chapter, I introduced a preprocessing and quality control platform to process methylation data. However, in addition to the size of the data, type of genomic data we obtain is also changing. One of the different types of data that has emerged is single-cell sequencing data. Compared to standard bulk genome sequencing, single-cell sequencing is performed from the genomic material obtained from one cell and the genomic signal would be specific to an individual cell instead of aggregate of large number of cells.

Like other sequencing technologies, conventionally bisulfite sequencing has been conducted in bulk samples where DNA is obtained from a large number of cells and then epigenomic and genomic inferences are made in terms of average of all those cells. However, bulk sequencing methods usually do not have the resolution to obtain information at the individual cell level. Therefore, single-cell methylation sequencing is a preferred method to estimate the methylation pattern at individual cell level (Tanay and Regev, 2017). Single-cell analysis has proven to be a powerful method to study rare cell populations, gene expression in various phases of the cell development and many other processes (Guo et al., 2013; Owens, 2012). In a similar manner, single-cell methylation analysis assays have the potential to reveal many of the developmental patterns, disease progression (i.e. lineage trees with higher resolution), and other temporal changes in the methylome (Chen et al., 2018b). Single-cell analysis can play a vital role in understanding genomic and epigenomic information simultaneously and assessing the epigenomic heterogeneity of multiple cell types in a groundbreaking manner (Macaulay and Voet, 2014).

Single-cell methylation analysis would be an appropriate method to observe highly dynamic methylation changes and detect temporal changes of the methylation. In single-cell methylation data methylation pattern of individual cells were measured separately. Therefore it would also be a great tool to quantify heterogeneity in the methylation landscape among the cells. There have already been several single-cell methylation studies
conducted to assess the methylation dynamics in early embryonic stem cells (Guo et al., 2013), understand and classify neuronal subtypes (Luo et al., 2017) and drug-induced remodeling of the methylation landscape (Farlik et al., 2015). As in many other single-cell assays, methylation analysis also has the fundamental problem of dropout or missing signal of interest (Smallwood et al., 2014; Farlik et al., 2015). In the case of methylation data, a vast majority of CpGs are not seen hence their methylation statuses are unknown. Methylation status of each CpG is ascertained from cellular DNA; however, each single cell has about 5 pg of DNA (Schwartzman and Tanay, 2015). Sample preprocessing with this small amount of DNA further exacerbates the problem of missing signals. Therefore methylation status from only a small fraction of the genome is observed.

Although single-cell methylation data provides higher resolution information about individual cells, high level of missingness can lead to biased conclusions where the methylation metrics are measured only based on the observed CpGs. This is exacerbated by the fact that intrinsically methylation levels are influenced by the CpG density (number of CpGs in a specified bases), i.e. in most cell types regions with low CpG density tend to have higher methylation and regions with high CpG density have lower methylation (Deaton and Bird, 2011; Chen et al., 2018a).

In this paper, we analyze the impact of missingness in single-cell methylation data using bulk methylation data as a reference. Then we show how mean methylation, a cell specific metric and cell-to-cell distance, a pairwise metric can be corrected. The issue of missing or unobserved methylation is also seen in different bulk methylation assays. Therefore we extend the same method to compare different methylation assays from bulk samples.

2.2 Methods

Single-cell methylation data is represented as a matrix with each row corresponding to a CpG locus and each column corresponding to an individual cell. Let $C$ represent the coverage matrix with $N$ cells and $K$ CpGs.
$C = \begin{bmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{K1} & \cdots & c_{KN} \end{bmatrix}$

$c_{kn}$ represents number of reads covering CpG $k$ in cell $n$. Let $M$ represent the methylation matrix for the same set of cells and CpGs.

$M = \begin{bmatrix} m_{11} & \cdots & m_{1N} \\ \vdots & \ddots & \vdots \\ m_{K1} & \cdots & m_{KN} \end{bmatrix}$

$m_{kn}$ represents the methylation (proportion of methylated reads compared to total number of reads) at CpG $k$ in cell $n$.

Both matrices are highly sparse due to missingness. If a specific CpG is missing then it will be represented with a 0 in the coverage matrix and with an NA in methylation matrix.

2.2.1 Sparsity of single-cell methylation data

First we assessed the sparsity of each cell using coverage matrix. sparsity for cell $n$ is defined as below:

$$\text{sparsity}_n = \frac{\sum_{i=1}^{K} I(c_{in} = 0)}{K}$$

$1 - \text{sparsity}_n$ would be the proportion of CpGs with at least one read coverage in cell $n$, which we defined as the observation rate.

2.2.2 Ground truth data

In our analysis of single-cell methylation data, ground truth is considered the oracle that has the knowledge of the methylation statuses for all CpGs. However, this oracle does not exist in real data analysis, therefore we used methylation statuses from bulk data as the best approximation to the ground truth. Based on the ground truth, we estimated the mean methylation for observed and unobserved CpGs in single-cell samples. Methylation levels in observed and unobserved CpGs should be similar if there is no systematic
difference between the two sets of CpGs. We used this assumption as a diagnostic tool to assess the difference in methylation between observed and unobserved CpGs.

Overall difference in methylation only conveys how global methylation differs between observed and unobserved CpGs. However, methylation differs by CpG density where CpGs in low density regions in general have higher methylation and CpGs in high density regions have lower methylation. Hence missingness can impact different CpG density regions differently. We confirmed this by first estimating the CpG density in the ground truth data. CpG density was defined as the number of CpGs in a 1 kb window in the genome. Therefore CpG density of a specific CpG is the number of adjacent CpGs in a 1kb window. In order to visualize the different levels of methylation CpG density regions are defined in terms of intervals of length 5. In each region we calculated the methylation in the ground truth data. In addition to that we also calculated the methylation in CpGs that are observed and unobserved in single-cell data by CpG density region.

2.2.3 Observation rate plots

To further analyze the pattern of missingness we constructed an observation rate plot. Observation rate is the proportion of CpGs in each CpG density region. Observation rate in region $r$ for an arbitrary cell is defined as below:

$$\text{observation rate}_r = \frac{\text{Number of CpGs observed in region } r}{\text{Number of CpGs in region } r}$$

Most CpGs are seen in the low density regions and fewer CpGs are seen in the high density regions. This leads to an observation rate curve in the reference genome with a shape that has higher mass in the low density region and a longer tail. We constructed the observation rate plot for multiple single-cell libraries (Table 2.1). Through this plot we assessed the variation in observation rates in cells within a library and across different libraries.
2.2.4 Correction model

Through ground truth based analysis we observed that there was a significant and directional difference in methylation between CpGs that are observed and CpGs that are not observed. Hence we utilized the strength of the ground truth data set to learn how methylation is impacted by missingness and use this knowledge to correct for mean methylation in single-cell data.

We propose a two-step method where in the first step we learn the impact of the missingness from the ground truth data and in the second step we correct the mean methylation estimates in single-cell data. Through this two-step process we preserve the variation in individual cells and at the same time correct the methylation for unobserved loci in the data.

Let $\mu$ be the methylation in ground truth data. $M$ is the methylation matrix where $m_{kn}$ corresponds to methylation in CpG locus $k$ in cell $n$. $z_{kn}$ is defined as the indicator for missingness where it is defined as the following,

$$z_{kn} = \begin{cases} 1, & \text{if } m_{kn} \text{ is NA} \\ 0, & \text{otherwise} \end{cases}$$

CpG density is denoted as $C$ and it will take an integer value between 1 and the maximum CpG density in the data.

In addition, we also added an interaction term $D$ between CpG density and missingness. The interaction will account for the difference in methylation between observed and unobserved loci by CpG density region. Using these covariates in the first step, we fit a model on the ground truth data in a piecewise fashion, where the pieces are defined based on the CpG density region.

$$\mu = \alpha_{\text{bulk}} + \beta_{\text{bulk}} Z_n + \gamma_{\text{bulk}} C + \delta_{\text{bulk}} D_n$$

The model above estimated $\beta_{\text{bulk}}$ and $\delta_{\text{bulk}}$, both of which assess the difference in methylation due to missingness in various contexts.
In the second step, another model is fit on the single-cell samples:

\[ M_n = \alpha^{sc} + \gamma^{sc}C \]

In this model \( \gamma^{sc} \) assesses the region specific methylation in a specific single-cell.

Finally, the methylation for unobserved loci is estimated by calculating \( \tilde{M}_n = \alpha^{sc} + (\gamma^{sc} + \delta^{bulk})C + \beta^{bulk} \). Overall mean methylation is calculated by combining predicted methylation and observed methylation. Figure 2.1 illustrates the workflow of the correction model.

For our model we defined CpG density regions with interval lengths of 10 and any CpG with a density above 120 was binned into one bin as there are fewer CpGs in very high CpG density regions.

**Figure 2.1:** Workflow of the two-step model to correct for mean methylation in single-cell data. In single-cell data 0 and 1 represents unmethylated and methylated CpGs and ? represents methylation in unobserved CpGs.

### 2.2.5 Application of the model in real data setting

To assess the missingness we applied our analysis and the correction model in multiple datasets. Datasets include 95 human embryonic stem-cells from HUES64 cell line, 4 library pools of Chronic Lymphocytic Leukemia (CLL) cells. Out of the four library pools
three were prepared with single-digest library preparation and one of them was prepared with double-digest library preparation, 1 library pool containing 25 CD19 B-cell samples. For the embryonic stem cells we used a HUES64 bulk sample as the ground truth, for CLL and B-cell samples we used similar bulk sample as the ground truth. We also used 9 bulk RRBS samples each from B-cell and CLL samples. Table 2.1 summarizes all the samples used in the analysis.

Single-cell samples have lower coverage of CpGs compared to bulk samples. However, cells with a CpG coverage of less than 10,000 CpGs are excluded from our analysis as they indicate failure in sequencing. For the correction analysis we only included samples that had the minimal coverage of 10,000 CpGs.

Table 2.1: samples used in the analysis.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Number of Cells</th>
<th>Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUES64 ESC</td>
<td>95</td>
<td>Double digest</td>
</tr>
<tr>
<td>CLL</td>
<td>25</td>
<td>Double digest</td>
</tr>
<tr>
<td>CLL</td>
<td>75</td>
<td>Single digest</td>
</tr>
<tr>
<td>CD19 B-cells</td>
<td>25</td>
<td>Double digest</td>
</tr>
<tr>
<td>CLL</td>
<td>9</td>
<td>Bulk RRBS</td>
</tr>
<tr>
<td>CD19 B-cells</td>
<td>9</td>
<td>Bulk RRBS</td>
</tr>
</tbody>
</table>

2.2.6 Distance correction

In single-cell methylation data, cell-to-cell distance would measure the difference in methylation landscape between two cells. Higher distance indicates the cells have divergent methylation landscape and smaller distance indicates cells have similar landscape. Therefore, cell-to-cell distance would be an ideal metric to measure the cellular heterogeneity. However due to missingness distance between two cells could only be calculated based on observed CpGs in both cells. Hence the distance would only reflect the pairwise complete observation. Regions that are highly observed due to protocol would contribute more to overall distance. Therefore we implemented a distance correction model where the distance contributed by each CpG density region would be weighted by the total number of CpGs in each region. We use correlation distance as the measure to calculate distance between two cells. Without any correction cell-to-cell distance between two cells
\( i \) and \( j \) would be:

\[
D_{ij} = \text{Distance}(M_i, M_j)
\]

However, this distance measure would be calculated based on CpGs that are observed in both cell \( i \) and \( j \). Therefore in single-cell methylation data, distance would be weighted more toward regions that have more shared CpGs. Cells that prepared with the same protocol would show lower distance compared to cells prepared with different protocol despite the cell type or other biological differences.

We introduce a correction method that mitigates the influence of regions that has high level of observation. Correction for the distance uses region based distance matrices, say the distance matrix in region \( k \) is \( D_k \) and the number of CpGs in region \( k \) is \( n_k \). Then the weighted distance matrix would be generated as:

\[
D_w = \sum_{k=1}^{K} \frac{n_k}{N} D_k
\]

Weighted distance metric gives higher weight to regions that have more CpGs in the reference genome and lower weight to regions that have fewer CpGs in the reference genome. This weighting scheme reduces the weight on regions that are widely observed due to protocol or other technical reasons and represents the biological differences between two cells.

### 2.2.7 Extension of single-cell methods for bulk methylation samples

As mentioned in Chapter 1, various methylation assays exist to measure methylation in specific regions of interest. For example Reduced Representation Bisulfite Sequencing (RRBS) assay obtains methylation statuses in CpG Islands and higher CpG density regions. Although these type of assays would provide region specific methylation, they would not provide overall sample specific metric such as mean methylation. We extend our single-cell mean methylation correction to bulk RRBS samples to obtain overall sample mean methylation metric.
2.3 Results

2.3.1 Sparsity in single-cell data

Average overall observation rate (1 - sparsity) was 3.3% (range: 10e-5% - 12.4%) for all the single-cell samples. Observation rate varied by different library. Samples prepared with single digest library protocol had lower observation rate than those prepared with double digest library protocol (Figure 2.2). Average overall observation rate in samples prepared with single digest library protocol was 1.44% and for samples prepared with double digest library it was 3.46%.

![Box plot showing observation rate by different library and protocol](image)

Figure 2.2: Observation rate differed by different library and protocol. However overall observation rate was quite low.

In HUES64 bulk sample, which we used as the ground truth data for the human embryonic stem (HES) cells, we observed that the methylation was higher in low CpG density regions, mid CpG density regions had lower methylation and high CpG density regions had the lowest methylation levels (Figure 2.3). Using this ground truth data we observed a significant difference in methylation between observed and unobserved CpGs in HES cells. Observed CpGs had a lower methylation compared to unobserved CpGs. Mean methylation values for the observed CpGs across samples had higher variance and variance was lower for methylation in unobserved CpGs (Figure 2.4). We observed similar trend where mean methylation in CpGs observed in double-digest protocol was higher.
and mean methylation in CpGs observed in single digest protocol was lower. (Supplementary Figure 4.9)

Figure 2.3: Methylation by the CpG density indicates that low density regions have higher methylation and as the CpG density increases methylation decreases.

When this comparison was stratified according to CpG density regions we saw a pattern where medium CpG density regions had the highest difference in methylation between observed and unobserved CpGs. Low and high CpG density regions also showed difference in methylation levels, however the difference is not as high as that in the medium CpG density regions (Figure 2.5). These observations demonstrated that there is a difference in methylation level between observed and unobserved CpGs and the difference varied by CpG density regions.

Based on these observations we evaluated whether CpGs are selectively observed in certain CpG density regions. For this purpose we generated the observation rate plots. Observation rate plot showed that higher proportion of CpGs is observed in low density regions and low proportion of CpGs is observed in high CpG density regions. Also the observation rate varied from cell to cell in the same library (Supplementary Figure 4.9).

Among the libraries the observation rate curves were shifted along the CpG density regions which indicated that some libraries had relatively higher proportion of CpGs from medium and high density regions while some libraries had higher proportion of
Figure 2.4: Using ground truth data as the reference we observed varying levels of methylation in observed vs. missing CpGs.

Figure 2.5: In addition to the overall differences in methylation, observed and missing CpGs also had varying levels of methylation in different CpG density regions. The difference was highest in the medium density regions where the methylation was more dynamic.
CpGs from low CpG density regions (Figure 6).

Figure 2.6: Observation rate varies by cells within a library and also varies by different library pool. Each line represents an individual cell and different libraries are represented with different colors.

Since we saw clear disparity among CpGs that are observed and CpGs that are not observed we applied our correction model where we used 95 HES cells as single-cell samples and HUES64 sample as the bulk cell, for CLL and B-cell single-cell sample we used the bulk sample of the same cell type as the ground truth. When we applied the correction model to HES cells we saw that overall methylation did not change significantly. However when the methylation was stratified by CpG density the change was significant in specific regions, especially in the medium density regions. Low and high density regions did not show significant change due to adjustment. This could explain the reason for non-significant change in the overall methylation because large proportion of CpGs are in the low density region and when overall methylation is calculated the contribution by those regions was higher compared to regions that showed significant differences before and after adjustment. Among different B-cell and CLL libraries we saw that the overall
change was quite significant in the libraries with very low observation rates. However when compared with multiple single digest and a double digest libraries CLL cells had high variation in their methylation but after correction cells in single digest and double digest had similar methylation values regardless of the coverage. Therefore after correction we saw similar methylation level in all the CLL cells.

### 2.3.2 Distance correction

Distance matrix without any correction showed that CLL cells were clustered together based on their protocol and B-cell cluster had high variance. One of the outlier sample from double-digest CLL library clustered together with the B-cell samples. Upon further examination, we identified that outlier cell had the lowest coverage of all the CLL double digest samples. Hence we generated distance matrix by CpG density region as defined in the mean methylation correction model. Clustering based on different region showed clear difference in how the cells were clustered. In low CpG density region there were two distinct clusters within CLL cells, however in the medium CpG density region there
was no distinct clusters within CLL cells and in high density regions distinction between CLL and B-cell clusters reduced drastically.

We combined distance matrix from various CpG density region to generate a weighted distance matrix. This distance matrix produced a clustering that clearly separated B-cell and CLL cell types and CLL cells were separated by mean methylation values instead of the protocol (Supplementary Figure 4.11).

2.3.3 Extension of single-cell methods for bulk methylation samples

In bulk methylation assays, Whole Genome Bisulfite Sequencing assay covers the entire genome however other assays such as Reduced Representation Bisulfite Sequencing covers high density CpG Island regions. Therefore we used our model to obtain global mean methylation from RRBS samples using appropriate WGBS sample as the ground truth.

For the analysis we used 18 RRBS samples nine CLL and nine B-Cell samples and for ground truth data we used corresponding CLL and B-cell WGBS samples. CLL - RRBS samples had an average coverage of 2.817 million CpGs (Range: 2.245 million - 3.302 million) and Bcell - RRBS samples had an average coverage of 2.896 million CpGs (Range: 2.548 million - 3.069 million)

Correlation between observed CpGs in bulk RRBS and bulk WGBS data was relatively high for both B-cells (median correlation: 0.846) and CLL cells (median correlation: 0.742). When a coverage threshold of 10 (i.e. CpGs with at least 10 reads covering them) was imposed the correlation further increased, median correlation for B-cell and CLL were 0.95 and 0.841 respectively.

We used the same 2-step correction model as in single-cell methylation data to obtain the global methylation of RRBS samples. After correction overall methylation in CLL and B-cells showed similar methylation values as in whole genome data (Supplementary Figure 4.12).
2.4 Discussion

Through coverage analysis we showed that single-cell methylation data is highly sparse and methylation statuses of only small fraction of CpGs are observed. This has been observed in several single-cell methylation studies [Farlik et al., 2015; Luo et al., 2017; Mulqueen et al., 2018]. Although bulk sample is not the perfect representation of an oracle for all the cells, it is still a mixture of multiple single-cells. Using bulk sample as the ground truth we showed that there is a clear distinction between the observed and unobserved CpGs. This difference was further observed when the methylation was stratified according to CpG density regions. These observations supported the claim that the observed CpGs are systematically different from the unobserved CpGs and the location of the CpG based on CpG density influences this difference. Also the difference was higher in the medium CpG density regions which have more dynamic methylation status.

Observation rate plots provided useful insight in comparing multiple cells from the same library and multiple libraries together. Observation rate plots showed that cells within a library had small fluctuation in the proportion of the CpGs are seen in each CpG density regions; however, this fluctuation was higher between libraries. This indicates there could be technical differences among libraries which leads to where the CpGs are observed. The libraries we included in our analysis showed high variation in the observation rates, this could not only affect the overall methylation as the methylation differs by CpG density region but also cell-to-cell distance in methylation where cells of same type would show different methylation by library/protocol preparation.

Ground truth based diagnostics explained above showed that single-cell DNA methylation data indeed has certain level of bias and to perform further analysis these biases need to be corrected or substantially reduced.

Our correction model takes all these observations into account. The piecewise regression model is a flexible model that takes varying methylation across CpG density into account. In addition the interaction term accounts for the differences by CpG density region. Overall the correction model encompasses what is observed in terms of single-cell and ground truth data.
Applying the correction model in various datasets provided better insights into how the correction reduces the bias in the data. In the HES cells overall mean methylation did not change significantly; however, region specific methylation showed a significant difference in mean methylation in medium CpG density regions which are usually more dynamic as the methylation has large range in medium density regions. In the CLL datasets correction model was applied to cells prepared with single digest and double digest protocol. Before correction cells prepared with single digest protocol had lower observation rate and lower methylation and cells prepared with double digest protocol had higher observation rate and higher methylation.

This was also supported by the observation rate plots for cells prepared single digest protocol further to the right with higher mass in the tail of the curve and cells prepared with double digest protocol had higher mass in the lower CpG density regions. After correction cells from both protocols had similar methylation values as they are of the same type of cells but at the same time preserving the variation between cells. Hence the correction reduced the technical variation due to protocol while preserving the biological variation.

Mean methylation is a metric that is specific to a single-cells and cell-to-cell distance is a pairwise-metric that measures the distance between two cells. High missingness in single-cell methylation data affects the cell-to-cell distance more as distance can only be calculated based on observed CpGs in a pair of cells. Our proposed solution of weighted distance matrix still uses the pairwise complete distance, however, we only calculate distance within a specific region. These region specific distance matrices provided distinct clustering patterns and when they were combined together the clustering showed clear distinction between cell types and effect of protocol was reduced.

Although, main motivation of our analysis is to understand the bias due to observation in single-cell methylation data and to propose correction method for different metrics we saw that our methods could be extended to bulk data where only fraction of the CpGs are observed. In single-cell methylation data missingness is due to the small amount of starting material in bulk assays such as RRBS and HSBS specific regions in the genome are targeted. However regardless of the mechanism of missingness, we have shown that our
model could be applied to bulk RRBS samples using bulk WGBS sample as the ground truth. We obtained global mean methylation for CLL and B-cell samples based on our analysis. Hence our method can be used to global methylation metrics regardless of the assay type.

2.5 Conclusion

Single-cell methylation provides high resolution information on methylation at individual cellular level. However small amount of starting material and other technical challenges lead to methylation of vast majority of CpGs being unknown. Hence single-cell methylation data would be highly sparse due to missingness in the methylation status. In order to gain biological insights sparseness of the data and challenges that arises with the sparseness should be addressed. In this chapter, we show the challenges associated with single-cell methylation data and propose solution to correct methylation metric such as mean methylation and cell-to-cell distance. After correction we obtain global level mean methylation and cell-to-cell distance that emphasizes the biological distinction of the cell types. Methods presented here would be useful when undertaking single-cell methylation data analysis both in diagnosing biases and also correcting them if needed.
Delineating finer population-substructure with rare variants

Divy S. Kangeyan
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Christoph Lange
Department of Biostatistics
Harvard T.H. Chan School of Public Health
3.1 Introduction

Sparsity in genomic datasets can manifest due to various reasons. In the previous chapter we elaborated on sparsity in single-cell methylation data that is attributed to the scarcity of starting DNA material. However, sparsity could be introduced due to biologically rare observations. In this chapter, we will focus on rare variant data which has high sparsity due to the fact that rare variants are observed at a very low frequency in haplotypes. We will explore methods to generate fine population structure based on rare variant data that takes sparsity into account.

Understanding and quantifying population substructure in genetic datasets is fundamental to answer many research questions, as undetected substructure can lead to biases/confounding in the statistical analysis, e.g., association mapping (Cardon and Palmer, 2003; Ptak and Przeworski, 2002). The knowledge of the substructure can also be important information by itself, as it facilitates a better understanding of the ancestry/relatedness of the study participants (Chakraborty and Jin, 1993; Astle et al., 2009). Consequently, statistical methodology to detect and to adjust for population substructure has been a very active research topic, rapidly changing with the development of new genotyping platforms, e.g., genome-wide SNP chips, whole-genome sequencing, etc (Price et al., 2006, 2010).

When genotyping capabilities in large-scale studies were still very limited, approaches such as genomic control (Devlin and Roeder, 1999) and STRUCTURE (Pritchard et al., 2000) were developed that assessed population substructure based on a very limited set of genetic markers. With the arrival of whole genome-wide association studies (GWAS), the assessment of population substructure at the whole genome-level became feasible, using data on thousands of common variants which were now available (Francioli et al., 2014; Gudbjartsson et al., 2015; Van den Eynden et al., 2018). The principal idea for GWAS data is to compute genetic similarity matrices between study subjects. Subjects are typically compared based on their empirical genetic variance/covariance which is computed based on several thousands of common loci that are not in linkage-disequilibrium. Principal component analysis (Hotelling, 1933) approaches are then ap-
plied to the genetic similarity matrices, in order to obtain a set of eigenvectors that ade-
quately reflect the population-substructure in the sample. The eigenvectors can be in-
cluded in the statistical model as covariates to guard the analysis results against con-
founding. In the mixed model based approach, candidate SNP and other phenotypic
characteristics are added as the fixed effect and kinship matrix based on genetic similar-
ity is incorporated as the covariance structure of the random effects (Kang et al., 2010;
Zhang et al., 2010). Both of the above methods work well in the common variant setting
however, their performance for rare variant data is not well studied and there are theoret-
ical concerns about the application of such approach in this context due to the sparseness
of the data.

As whole-genome sequencing have become now a standard research tool, whole-
genome rare-variant data on very large number of loci are now available in many studies.
It is estimated that vast majority of the variants observed in the genome are rare variants
(Nelson et al., 2012). Excess of rare genetic variations is attributed to an explosive growth
rate in human population over the past 400 generations (Keinan and Clark, 2012). Rare
variants tend to be genetically younger than common variants (Mathieson and McVean,
2014). Therefore they contain higher information regarding the population substructure
and provide better insights into dependence structure among study subjects. Further-
more, rare variants are considered to shed more light into unexplained heritability of
phenotypic traits (Manolio et al., 2009).

Hence, one would like to utilize rare-variant information to adjust/guard association
analysis against confounding due to population substructure. However, empirical vari-
ance/covariance matrices are not necessarily numerical stable for rare variant data (Lin
and Tang, 2011). Most rare-variant association studies therefore, while analyzing associ-
ations with rare variants, rely on adjustments for population-substructure that are based
only on common variant data (Bonnefond et al., 2012; Diogo et al., 2013; Johansen et al.,
2010). To address this issue, alternative approaches based on the Jaccard-Index have been
suggested (Prokopenko et al., 2015). Jaccard-indices are especially suited for sparse-data
formats and their calculations are computationally faster than the ones for empirical vari-
ance/covariance matrices (Prokopenko et al., 2015; Schlauch et al., 2017), enabling their
application to large panels of rare variants.

In this chapter, we compare approaches to infer population substructure from rare and common variant data, using the 1000 Genome Project (1000 Genomes Project Consortium and others, 2015). Using principal component analysis, we examine how well the known population structures of the data are reflected in the PC-plots when the similarity matrices are computed based on the variance/covariance matrices or the Jaccard Index. In particular, we select here a weighted Jaccard approach, as it converges asymptotically towards the kinship coefficients. We define formal measures of clustering and assess the performance of both approaches for different data types.

3.2 Methods

3.2.1 Quantitative measures to analyze rare variant data

Rare variant data can be structured as a matrix with columns being haplotype or genotype and rows being individual variants. Since the variants are rare, overall only 1% of the haplotypes have a specific variant, therefore haplotype matrix would be highly sparse. Hence rare variant data present the same issues associated with sparse matrix analysis. Here we present a few measures to construct similarity matrix from the rare variant data, measure the cluster affinity within a population group and information content of the haplotypes.

Constructing genetic relatedness in rare variant setting

Genetic relatedness is a measure of association between two haplotype or genotypes. When genetic relatedness is calculated among multiple haplotypes it will result in the Genetic Relatedness Matrix (GRM). Jaccard distance based method and variance/covariance based method are two ways to generate GRM.

Jaccard distance \([\text{jaccard}]\) 1908 between two binary vectors \(v\) and \(u\) of length \(K\) can be calculated as in eqn (3.1)

\[
Jaccard(u, v) = \frac{\sum_{k=1}^{K} I_{uk=1}I_{vk=1} - \sum_{k=1}^{K} I_{uk=0}I_{vk=1} + \sum_{k=1}^{K} I_{uk=1}I_{vk=0}}{\sum_{k=1}^{K} I_{uk=0}I_{vk=1} + \sum_{k=1}^{K} I_{uk=1}I_{vk=1} + \sum_{k=1}^{K} I_{uk=1}I_{vk=0}}
\] 

(3.1)
Based on the equation it is clear that Jaccard distance does not account for the scenario where kth element of both vectors are zero. As stated rare variant data is highly sparse and two haplotypes are more likely not have a mutation for certain variant. Therefore Jaccard distance would be an appropriate measure for relatedness in rare variant data as it will not inflate the similarity between two vectors just based on zero observations in both vectors.

We used the weighted Jaccard distance proposed by Schlauch et al. (Schlauch et al., 2017) to construct the Genetic Relatedness Matrix (GRM). Weighted Jaccard distance adds additional weight to the Jaccard distance where the rarer variants are upweighted and common variants are down weighted. Weights are the inverse of the probability of observing two haplotypes/genotypes with a mutation for a specific variant.

GRM could also be constructed from variance and covariance matrix of the Haplotype matrix. First the haplotype matrix is standardized and then singular value decomposition (SVD) is performed on the new matrix. After SVD, columns in U-matrix will be the principal components. Standardization and decomposition steps are further explained below. Say the haplotype matrix is $H$ and value for SNP $i$ and haplotype $j$ is denoted as $h_{ij}$.

row mean for ith row:
$$
\mu_i = \frac{1}{N} \sum_{i=1}^{N} h_{ij}
$$

posterior estimate of the unobserved underlying allele frequency:
$$
p_i = \frac{1 + \sum_{i=1}^{N} h_{ij}}{2 + 2N}
$$

Each row of the haplotype matrix will be standardized as below:
$$
X_i = \frac{h_{ij} - \mu_i}{\sqrt{p_i(1-p_i)}}
$$

Singular Value Decomposition of $X$:
$$
X = U \Sigma V
$$

each column of $U$ corresponds to principal component of the matrix $X$. 36
Quantifying cluster affinity among population groups

GRM is helpful to visualize the population structure and also to construct covariates that could be used to avoid confounding due to ancestry in association studies. However given the population group, GRM could also be used to measure whether similar population groups have high level of relatedness. To assess the cluster affinity we constructed a ratio measure similar to the ratio between sum of squared distance within the cluster and sum of squared distance among different clusters. Following is the definition of the ratio measure.

Given a Genetic Relatedness Matrix S, similarity between haplotype $i$ and $j$ is denoted as $s_{ij}$.

say haplotype $i$ belongs to Group G and there are $n$ total haplotype in group G.

In total there are $N$ haplotypes

$x_i^{WG}$ - mean similarity within the same group

$$x_i^{WG} = \sum_{\{j \in G \text{ and } j \neq i\}} \frac{s_{ij}}{(n - 1)}$$

$x_i^{OG}$ - mean similarity outside the group

$$x_i^{OG} = \sum_{\{j \notin G\}} \frac{s_{ij}}{(N - n)}$$

Overall ratio measure:

$$r_i = \frac{x_i^{WG}}{x_i^{OG}}$$

This ratio measure can be interpreted as the similarity of a specific haplotype with other haplotypes within the group (super- or sub-population) compared to the similarity of the same haplotype with all other haplotypes outside the group. Ratio measure above 1 for a haplotype indicates that that specific haplotype is more similar to other haplotypes within the group than haplotypes from other groups. Ratio measure below 1 indicates that the haplotype of interest is less similar to its assigned group and more similar to haplotypes from other groups.
Information content of rare and common variants

Information contributed by each variant is essential in building the genetic architecture of populations. Variants with higher information content are useful in understanding unrecognized population structure. Information theoretic framework has been used previously to assess the information gain of individual variants in population structure analysis (OConnor et al., 2014). We quantified the information content of variants based on mutual information.

Mutual information is an entropy based measure that quantifies the dependence between two variables. Entropy for a random variable \( X \) is defined as:

\[
H = - \sum_i P(x_i) \log(P(x_i))
\]

Entropy is a measure of information where an event with higher uncertainty contains high level of information. Mutual information is the difference between entropy of an event and conditional entropy of the same event given another event. Conditional entropy \( H(Y|X) \) is a measure of uncertainty in \( Y \) given the realization of \( X \).

Mutual information between two random variables \( X \) and \( Y \) is defined as:

\[
I(X; Y) = H(Y) - H(Y|X)
\]

where \( H \) is the entropy

Through mutual information we measure the information content of haplotype \( X \) given that we have haplotype \( Y \). If the mutual information is higher between two random variables then less information is contributed by one variable given the other variable is observed. In the case of rare events mutual information is generally lower since less information is contributed by one rare event given that you know the occurrence of another rare event. We used a relative measure since absolute measure of information might vary from dataset to dataset and less interpretable.

In our analysis we estimated the mutual information contributed by a haplotype given another haplotype. Then we generated a pairwise mutual information matrix for all the haplotypes in common and rare variant settings.
3.2.2 1000 Genomes Project (TGP) data

To visualize the population substructure and to assess the information content of the rare variants we conducted a series of analyses on the 1000 Genomes Project (TGP) Consortium Phase 3 data (1000 Genomes Project Consortium and others, 2015). The TGP data contains 2504 individuals from five super-populations and 26 sub-populations across various regions of the world which makes it an ideal dataset to conduct population substructure analysis.

3.2.3 Data analysis

We defined rare SNPs as those with MAF (Minor Allele Frequency) of less than or equal to 1% and common SNPs as those with MAF greater than 1%. According to this definition, we separate SNPs as rare and common variant set from chromosome 1 in TGP data. From common and rare SNPs, we selected around 100,000 SNP after linkage disequilibrium pruning to conduct further analysis. In both sets of SNPs we applied the weighted Jaccard distance approach and eigen decomposition based method to generate the GRM that shows similarity between each pair of haplotypes. Principal Component Analysis (PCA) was applied to GRM to visualize grouping of super- and sub-populations. In the case of rare variant data considerable number of variants would only occur in one haplotype, these variants are called singletons.

We applied this analysis to different subset of haplotypes in order to observe certain populations of interest at higher resolution. First the analysis was applied to all haplotypes, then the analysis was applied to different super-populations in order to visualize the distinct sub-population structure within super-populations. In addition we also applied this analysis to all the haplotypes without the outlier populations. For example the African (AFR) super-population is an outlier population compared to other super-populations and within the African super-population the Luhya people in Webuye, Kenya (LWK) are considered an outlier sub-population compared to other African sub-populations (1000 Genomes Project Consortium and others, 2015). When all super-populations or all sub-populations within a super-population is included in the analysis
usually one or two outlier groups capture most of the variation. Therefore we conducted this analysis to observed how rare and common variants depict the substructure in the absence of one or two groups that captures most of the variation within multiple groups.

In addition to visualizing the population substructure, we also quantitatively compared the level of clustering and information content between common and rare variants.

3.3 Results

Based on our analysis we observed that PCA of common variants across all samples yielded the well known visualization (Prokopenko et al., 2015; Glusman et al., 2017) where the first PC separated African super-population from all other super-populations and the second PC separated East Asian super-population from other three (AMR, EUR and SAS) super-populations. This was true for visualization generated by weighted Jaccard analysis and eigen decomposition based visualization. However, rare variant based PCA showed different PCA visualization based on how the GRM was constructed. Weighted Jaccard based PCA showed that most of the variation is captured by the African super-population and the first PC separated the African super-population from every other super-population while the second PC separated the LWK, an East African sub-population from other African groups (Figure 3.1; Supplementary Figure 4.13). Higher PCs showed distinct clusters of other super-populations such as EAS, EUR and SAS (Supplementary Figure 4.14). PCA based on eigenstrat in rare variant setting showed similar picture as in common variant setting except the variation among admixed American haplotypes was low. (Figure 3.2)

When the PCA was conducted without the African super-population there was a clear distinction between common and rare variant settings. In weighted Jaccard based analysis common variants PCA did not show any clear distinction among super-populations while rare variants based PCA showed distinct clusters of super-populations (Supplementary Figure 4.15). In eigenstrat based analysis rare and common variants shows distinct clusters for EUR, SAS and EAS super population; however, common variants based PCA showed high variation in AMR super population and this was not seen in rare vari-
ants based PCA. (Supplementary Figure 4.16)

Figure 3.1: Population structure in common and rare variants in all the haplotypes using weighted Jaccard based methods

Within the African super-population rare variants based PCA showed distinction between LWK, an East African sub-population from all other West African or West African based diaspora sub-populations (Supplementary Figure 4.17). Also, East African sub-populations showed a spectrum where African diaspora populations ACB (African Caribbeans in Barbados) and ASW (Americans of African Ancestry in SW USA) were in the middle and rest of the the west African groups ESN (Esan in Nigeria), MSL (Mende in Sierra Leone), GWD (Gambian in Western Divisions in the Gambia) and YRI (Yoruba in Ibadan, Nigeria) were on the other end of the spectrum. When we conducted the analysis without LWK sub-population further distinction was observed within West African super-populations. For example ESN and YRI sub-populations from Nigeria/gulf of Guinea were clustered together while GWD and MSL sub-populations from western coast of Africa were clustered together. Also the diaspora population of ACB and ASW situated in the middle between the two groups (Figure 4.18). In both analysis common variants based data did not show distinct clusters in weighted Jaccard based analysis however eigenstrat based analysis showed some level of clustering, however, in this PCA visualization the two diaspora population ACB and AWS captured the most variation.

In the East Asian super-population we saw a clear distinction of the sub-populations
Figure 3.2: Population structure in common and rare variants in all the haplotypes using Eigen decomposition based methods

in the rare variant context (Supplementary Figure 4.19). We also observed a spectrum across five sub-populations within East Asian super-population. Japanese in Tokyo, Japan (JPT) were in one end of the spectrum and two groups of Chinese sub-population, Han Chinese in Beijing, China (CHB) and Southern Han Chinese (CHS) were in the middle and the two southeast asian populations Kinh in Ho Chi Minh City, Vietnam (KHV) and Chinese Dai in Xishuangbanna, China (CDX) were in the other end of the spectrum.

These visualization depicted a clear difference between rare variant and common variant data. In order to quantitatively assess this separation we used the ratio measure. Rare variants showed higher ratio measure across all super-population indicating higher level of within group clustering. Ratio measure was higher in the African and East Asian super-population and lowest in the American super-population in the rare variant context (Supplementary Figure 4.20). Within the African super-population, highest ratio measure was observed in LWK sub-population. Interestingly, the ratio measure was relatively higher in the ACB sub-population and absolutely higher in ASW sub-population in the common variant data. It should be noted that ACB and ASW are diasporic population from Western Africa (Figure 3.3). In all other sub-populations ratio measure was higher in rare variants.

We used information theoretic based framework to estimate the information contributed by variants. We constructed pairwise mutual information matrix for all the haplotypes in TGP data. Mutual information matrix based on common variants showed a
Figure 3.3: Ratio measure in common and rare variants context among sub-population across the African super-populations. Ratio measure is a comparison metric for genetic similarity within the group compared to outside the group. Most African sub-populations show higher ratio measure when the similarity is measured with rare variants instead of common variants. Diaspora groups such as ACB and ASW show higher ratio measure in common variant data.

stark difference compared to the matrix generated with rare variants. Mutual information matrix based on common variants depicted that within a super-population mutual information is higher and when haplotype in one super-population to another haplotype in different super-population mutual information was lower. Rare variants still showed similar patterns, however overall mutual information was lower. Notably, compared to all other non-African super-population, haplotypes in American super-population showed lower mutual information indicating the high level of admixture within the American super-population (Figure 3.4).

3.4 Discussion

Rare variants based GWAS data has been increasing rapidly due to advance in sequencing technology. Since rare variants are genetically younger it is hypothesized that they contain higher information than common variants regarding the ancestry and population structure. Our series of analyses visually and quantitatively elaborate the information
content of rare and common variants in the context of population substructure.

When the weighted Jaccard was used in the common variant data we obtained the same 1000 Genomes Project visualization (Figure 3.1) indicating the method preserves the population structure identically to that obtained through eigen decomposition. However in rare variant setting, most variation was captured by the African super-population which agree with the fact that African genome has higher level of haplotype diversity (Gabriel et al., 2002; Ionita-Laza et al., 2009). In addition to that LWK, an East African sub-population was separated from every other African sub-population group which could be attributed to the fact that modern human originated in eastern part of Africa and then migrated from there to western parts of Africa and further into other parts of the world (Supplementary Figure 4.10). This PCA visualization not only agrees with the out of Africa theory (Nei, 1995) but also as humans migrated the haplotype diversity decreased in a process similar to serial dilution (Conrad et al., 2006). Figure 3.1 shows a clear distinction between rare and common variant data, where rare variants cluster the group in distinct fashion however most of the variability is captured by American super-
population with high level of admixture in the common variant setting.

In sub-populations among African and East Asian super-population we saw clear
distinction in the rare variants (Supplementary figure 4.13; 4.15). For example outlier
populations LWK and JPT was clearly separated from other sub-populations and there
was a clear alignment of other groups based on the geography. All of the analysis con-
ducted in super-populations, sub-populations and sub-populations without the outlier
population visually demonstrated a clear difference between haplotypes in the rare and
common variant context such that haplotypes in the rare variant context showed higher
resolution structure of the groups.

Further ratio measure and mutual information are used as quantitative measures to
compare the strength of clustering and information contributed by rare and common vari-
ants. Ratio measure should be interpreted as comparison of within group similarity vs
similarity to other groups. Ratio measure higher than 1 indicates that a specific hap-
lotype is closely affiliated with haplotypes within the group than haplotypes in other
groups and ratio measure below 1 indicates that a specific haplotype has closer affinities
to external groups than its own group. Also beyond 1 and below 1 indicates strength of
affinity within and outside group. At super-population level all groups had higher ratio
measure in the rare variant context than in the common variant context indicating haplo-
types based on rare variants showed closer affinity within a super-population. Most sub-
population showed similar pattern, however in ASW ratio measure in common variant
context is higher and in ACB and PUR (Puerto Ricans from Puerto Rico) sub-populations
ratio measure in common variants was relatively higher compared to ratio measure in
rare variant context. Strikingly, all these population groups are highly admixed and/or
diasporic population. Therefore even simple measure such as this ratio measure was able
to delineate the difference in ancestry based on rare and common variants.

Mutual information was used to quantify the information content of haplotypes. It
is expected that within the same sub- or super-population mutual information will be
higher because haplotype will share more information content within the same group.
However we show that mutual information was quite different between rare and com-
mon variants. In common variants overall mutual information was higher and also mu-
tual information within super-population was higher. However in rare variants mutual information was in general lower indicating that each individual haplotype contributes more information as the dependence is lower between two haplotypes.

### 3.5 Conclusion

Through a series of analyses we have shown how finer population structure can be obtained in the rare variant context. Further, in the rare variant context we observed clear grouping of sub-populations that could not be obtained from the common variant data. Group similarity ratio measure was introduced to quantify the within group similarity for specific haplotype and mutual information was introduced to quantify information content of a haplotype given another haplotype. In both measures we saw that rare variants showed a distinct pattern compared to common variants and specifically rare variants showed higher within group similarity and overall lower mutual information content. These observations reinforced the fact that rare variants contain higher information content. Also observations in our analyses showed that ancestry information obtained from rare variants would be more informative to correct for population substructure.
Supplementary Materials
4.1 A (Fire)Cloud-Based DNA Methylation Data Preprocessing and Quality Control Platform

4.1.1 Quality Control metrics

1. Bisulfite conversion rate

Bisulfite conversion rate quantifies the successful conversion of the Cs to T's in the CpG context in bisulfite sequencing.

\[
\text{Bisulfite Conversion Rate} = \frac{\text{Total number of C to T conversion}}{\text{Total number of C's in non-CpG context}}
\]

In bulk sequencing, the cutoff for this measure is usually above 99%. Small amount of DNA in single-cell sequencing should not affect the conversion rate, hence similar cutoff can be applied in single-cell methylation data analysis. Therefore, removing samples with bisulfite conversion rate less than 99% will avoid any erroneous inference.

2. Methylation Bias (M-Bias) plot

A Methylation bias plot provides measurement of average methylation by position across the read. For CpGs in each position, the average methylation is calculated and ideally this should be a horizontal line since the methylation status should be uniform across the read. Due to low read count, in single-cell sequencing M-bias plot shows jagged pattern instead of a horizontal line. There could be deviation at the beginning and end of the read due to lower quality bases at the end of the read and adapter contamination. This could be overcome by removing bases at the end or beginning of the read. However, low quality single-cell samples could that show erratic pattern with preferential methylation in certain positions of the read should be excluded from further analysis.

However, cells with lower quality exhibit erratic pattern even for single cell samples.

3. Downsampling plot

Downsampling is a technique to assess the efficiency of CpG capture. Downsam-
Figure 4.1: Distribution of bisulfite conversion rate for a set of single-cell samples; in this sample set all samples have high bisulfite conversion rate.
Figure 4.2: Methylation bias across both reads in paired-end sequencing. Methylation bias in the end of the read show high fluctuation, however on average it is a horizontal line across the read.
pling saturation analysis provides information on the number of CpGs that can be captured with a fraction of the reads from a specific sample. If adequate amount of CpGs are observed with a given read count, then the downsampling curve reaches a horizontal asymptote. If the sample does not have sufficient coverage, then downsampling curve will be line close to x axis and if the sample does not achieve saturation it will be a straight line with positive slope.

![Saturation Analysis Plot](image)

Figure 4.3: Downsampling plot for one thousand single cells shows that most of the cells attain saturation beyond 75 % of the read coverage. Flat lines in the bottom of the plot indicate cells that failed and green dots depict the median CpGs covered across different fraction of reads.

4. **CpG Density plot**

   CpG Density plot depicts the fraction of CpGs observed along various CpG den-
CpG density is defined as number of CpGs seen in a certain length of the genome. By default `scmeth` (Kangeyan, 2018b) defines it as 1 kb region. However, the region can be defined according to the user. A CpG density plot is a useful diagnostic to assess what regions in the genome are observed.

Figure 4.4: CpG Density plot shows that the vast majority of the CpGs covered in these samples are in low or medium CpG density regions.

5. **Feature Level Coverage**

Feature level coverage shows the number of genomic features with at least one CpG. Genomic features in the QC report include CpG Island, CpG Shores, CpG Shelves, Intergenic region, exons and introns.
Figure 4.5: CpG coverage in different genomic features from all the samples in the sample set.
6. CpG Discretization Plot

CpG discretization plot is a diagnostic specific for single-cell methylation data. It gives the percentage of CpGs that have non-binary methylation with certain threshold. Ideally all CpGs should have a binary methylation level in single-cell data; however, due to sample quality there might be CpGs with non-binary methylation. CpG discretization is defined as below:

\[ |m| = \begin{cases} 
1, & \text{if } m \geq 0.8 \\
0, & \text{if } m \leq 0.2 \\
\text{undefined, i.e. discarded} & \text{if } 0.2 < m < 0.8
\end{cases} \]

Figure 4.6: CpG discretization plot shows that median percentage of CpGs with non-binary methylation is around 1.2 %.
7. **Read Information Plot**

Read information plot shows the total number of reads, mapped and unmapped reads.

![Read mapping stats](image)

Figure 4.7: Read coverage plot for a samples set containing 1000 cells. Most cells have high mapping rate.

8. **CpG Coverage & Methylation Plot**

CpG Coverage shows the total number of CpGs with read coverage of at least one and methylation plot shows the mean methylation for all the samples.
Figure 4.8: Median CpG coverage for the set of samples is 250,000.
4.2 Modeling missingness in single-cell DNA methylation data

4.2.1 Methylation based on ground truth

![Methylation based on ground truth](image)

Figure 4.9: Methylation based on ground truth

As seen in the figure above observation rate plot has similar shape within the same library pool, however individual cells still show variation in high CpG density regions.

4.2.2 Observation rate plot within a library

As seen in the figure above observation rate plot has similar shape within the same library pool, however individual cells (indicated by lines) still show variation in high CpG density regions.
Figure 4.10: Observation rate plot in HES cells from the same library

Figure 4.11: Clustering based on the distance matrix before (left) and after (right) correction. Before correction cells are mostly clustered according to protocol. After correction the cells are separated by cell type and methylation levels.
4.2.3 Distance correction results in single-cell DNA methylation data

Distance correction led to cells being clustered based on cell type and methylation levels instead of protocol. Also B-cells had higher variance before correction and after correction CLL cells had higher variance.

4.2.4 Correction in the bulk samples

![Box plots showing methylation levels before and after correction](image)

Figure 4.12: After correction B-cell samples had higher methylation than CLL samples as expected based on the bulk Whole Genome data.

Based on RRBS assay b-cells had lower methylation compared to CLL cells; however, after correction b-cells had higher methylation as expected from the whole genome methylation data.
4.3 Delineating finer population-substructure with rare variants

4.3.1 PCA visualization of population structure

PCA plots of population structures in various groups with and without the outlier populations.

Figure 4.13: Population structure in all haplotypes shows that haplotype of African origin show large haplotype diversity. Second principal component separates east African sub-population of LWK from all other west African populations.

4.3.2 Cluster affinity among all sub-populations

Cluster affinity across all sub-population in the TGP data. Panels are for five super-populations.
Figure 4.14: Higher order PCs for the haplotypes shows that EAS, EUR and SAS super-populations are clearly separated along 3rd principal component.

Figure 4.15: Within the African super-population, rare variants based PCA shows higher level of grouping by geography of origin for the sub-populations while common variants based PCA does not show distinct grouping.
Figure 4.16: Within the African super-population, rare and common variants based PCA show similar distinct clusters. However variation within the American population is considerably reduced as shown in previous Eigenstrat based analysis.

Figure 4.17: Within the African super-population, rare variants based PCA shows higher level of grouping by geography of origin for the sub-populations while common variants based PCA does not show distinct grouping
Figure 4.18: PCA visualization of sub-populations within the African super-population in common and rare variant context when LWK (Luhye in Webuye, Kenya), an east African sub-population which is considered an outlier sub-population.

Figure 4.19: PCA visualization of sub-populations within the East Asian super-population shows better separation of sub-population in the rare variants and groups exhibit clear association with geography of origin whereas common variants show little separation among sub-populations.
Figure 4.20: PCA visualization of sub-populations within the East Asian superpopulation shows better separation of sub-population in the rare variants and groups exhibit clear association with geography of origin whereas common variants show little separation among sub-populations.
References


NATIONAL CANCER INSTITUTE (2019). Target: Therapeutically applicable research to generate effective treatments.


population stratification in sequencing data: a simulation study and an application to the 1000 genomes project. *Bioinformatics* **32** 1366–1372.


