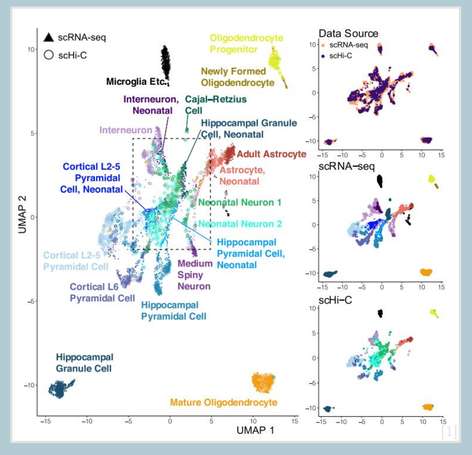
Statistical Machine Learning Methods for Integrating Single-Cell Genomics Data

Jake Blackwell, Paimon Goulart, Claire Lu
Faculty Contact: Dr. Wenxiu Ma, Postdoc: Rui Ma
University of California, Riverside

Introduction

Single-cell high-throughput chromatin conformation (**scHi-C**) data measures contact frequency between two genomic regions to facilitate understanding of its 3D genome organization. Similarly, single-cell RNA sequencing (**scRNA-seq**) examines the gene expression of every gene in an individual cell. Thus, the integration of **scHi-C** data and **scRNA-seq** data allows for the research and study of chromatin organization and gene expression within individual cells. For instance, the integrated data is applicable in disease identification by revealing outlier clusters in the integration graph. This can indicate abnormal cellular behavior and lead to further investigation.

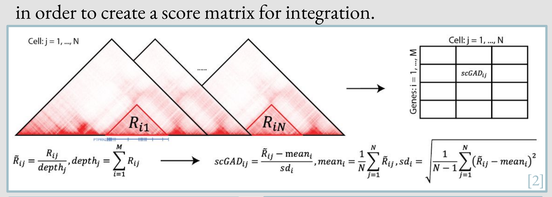


In integration, we project the scHi-C and scRNA-seq data onto the same subspace to better understand the relationship between chromatin organization and gene expression. However, comparing the raw data is difficult because of the three-dimensional nature of scHi-C data and the one-dimensional format of scRNA-seq data. Thus, we transform our scHi-C data for compatible integration with scRNA-seq data. We focus on two algorithms, **scGAD** (single cell Gene Associating Domain) and **MUDI** (Multichannel Data Integration), that are capable of integrating the scHi-C and scRNA-seq data. We replicate the integration processes of both scGAD and MUDI to conduct a comparative analysis, evaluating the advantages and disadvantages of each algorithm.

Methodology

scGAD

We need to process the scHi-C data and transform it to be comparable to the scRNA-seq data for integration. The scGAD algorithm is capable of processing scHi-C data by taking large sets of cell data and grouping its chromosomal interactions by gene names in order to create a score matrix for integration.



	readID	chr1	pos1	chr2	pos2
0	.	chr1	3017457	chr1	47222788
1	.	chr1	3017717	chr1	4179983
2	.	chr1	3018777	chr1	6953743
3	.	chr1	3026467	chr1	31175939
4	.	chr1	3026467	chr1	31176323

"chr"	"s1"	"s2"	"strand"	"gene_name"
"chr1"	3276124	3741721	"-"	"Xkr4"
"chr1"	4069780	4479464	"-"	"Rp1"
"chr1"	4561154	4567577	"-"	"Sox17"
"chr1"	4843429	4855962	"-"	"Mrpl15"

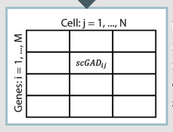
The raw chromosomal data for a single cell (left) and gene annotations (top). The gene annotations are used to assign gene names to each chromosome based on their positional data.

"chrA"	"binA"	"chrB"	"binB"	"count"	"geneA"	"geneB"
chr1	3310000	chr1	3310000	1	Xkr4	Xkr4
chr1	3320000	chr1	3350000	1	Xkr4	Xkr4
chr1	3330000	chr1	3330000	2	Xkr4	Xkr4
chr1	3330000	chr1	3340000	1	Xkr4	Xkr4
chr1	3370000	chr1	3370000	1	Xkr4	Xkr4
chr1	3380000	chr1	3380000	1	Xkr4	Xkr4

Updated dataset with a resolution of 10KB and gene names assigned to chrA and chrB (previously chr1/chr2) based on their binA and binB (previously pos1/pos2)

"gene_name"	"interactions"
Wwp1	1
Wwp2	9
Wwtr1	4
Xiap	1
Xirp2	9
Xkr4	26
Xkr6	12
Xkr9	1

The chromosomal data is grouped by their assigned gene names and the number of interactions represents the sum of the count column for each gene.



Update the score matrix, where the intersection of a gene name and cell file name ($scGAD_{ij}$) is the number of interactions. This will give us our raw score matrix, but in order to prepare this data for integration, we need to apply statistical analysis to the matrix.

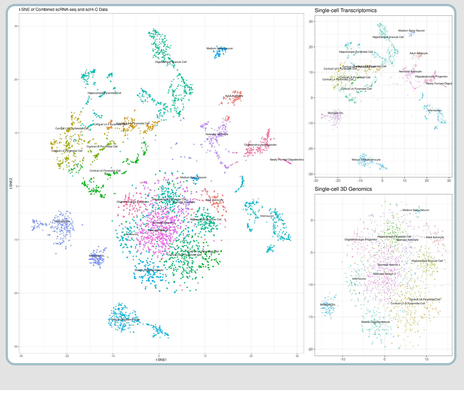
MUDI

In order to integrate scHI-C data and scRNA-seq data, the MUDI algorithm takes a different approach. First, it identifies scHi-C clusters using the schicluster library and scRNA-seq clusters using Seurat. The algorithm then integrates the Chromatin Accessibility Domains (CADs) of each scHi-C cluster with the differentially expressed genes (DEGs) of each scRNA-seq cluster to calculate integration scores. The integration score for a gene, denoted as $\mathbf{I}_g$ is defined by the formula shown below. From this integration score, we can identify different Topologically Integrated SubPopulations (TISPs) of genes.

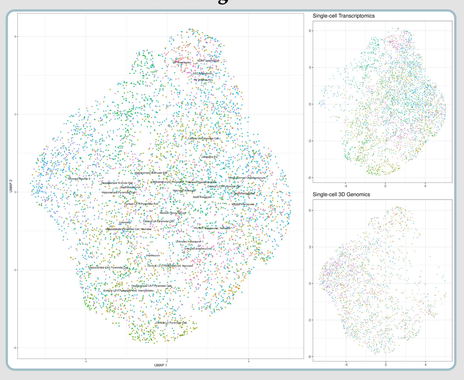
$$I_g = \frac{F_g E_g}{DR}$$ [3]

Results

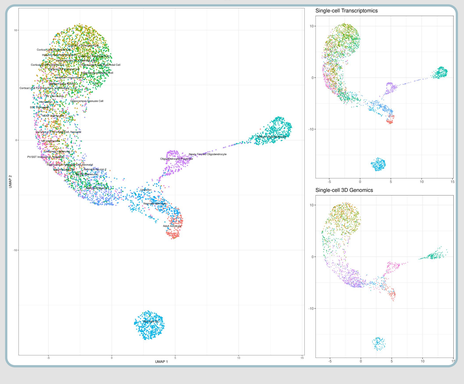
Pre-Integrated Data



Naive Integrated Data



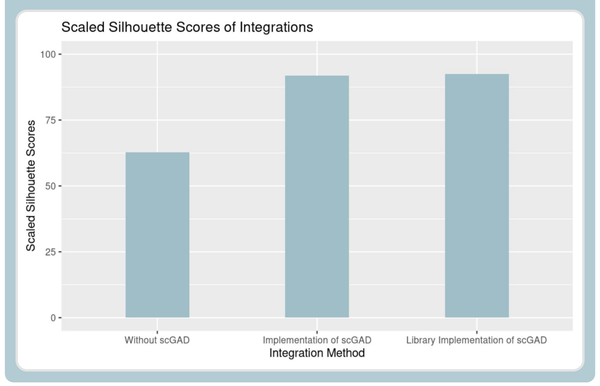
scGAD Integrated Data



Conclusion

Our results demonstrate that our implementation of the scGAD algorithm successfully integrated the scHi-C and scRNA-seq data with 99.4% accuracy. Though we were able to implement the MUDI algorithm, the information produced was not as relevant to this project as the results produced by scGAD.

Since we were able to accurately replicate these algorithms, we can reuse our implementations in future work so that we can analyze new datasets. We can also look into optimizing the runtimes of each algorithm to make them more efficient than their original implementations.



Evaluation between Methodologies

scGAD

- **Advantages**
  - Only takes a few function calls to get results
  - Results are easy to evaluate due to visual outputs
- **Disadvantages**
  - Long processing time
  - Specific format for data input is required

MUDI

- **Advantages**
  - Provides detailed results
  - Faster than our scGAD implementation
- **Disadvantages**
  - Results are difficult to compare to scGAD
  - Poor documentation for algorithm
  - Specific data requirements

Acknowledgements

This work is supported by the DS-PATH Summer Fellowship Program under the National Science Foundation Harnessing Data Revolution Data Science Corps Award #2123444, #2123271, #2123313.

References