*Leiden University*

MSc Research Project report

# Exploring Single Nucleotide Variants as shared features for unpaired Genome-Transcriptome integration

MASTER OF SCIENCE

In Biology,

Molecular Genetics and Biotechnology

BY

*Jorrit van Uhm*

*S3378101*

*Supervision by:*

*Associate Prof. M.P. Chien**

*MT. Lopez-Cascales*

* contact for lab journal

***The Chien lab, Erasmus MC***

*Dr. Molewaterplein 40*

From 25-09-2023
To 25-03-2024
30 EC
Following format: BMC Bioinformatics

# ABSTRACT

**Background**:
Single-cell sequencing techniques have revolutionized biological research, enabling the detailed study of cellular heterogeneity within complex systems. The integration of genomic and transcriptomic data, obtained through methods like single-cell whole-genome sequencing and single-cell RNA sequencing, is essential for a comprehensive understanding of biological processes. This project aims to analyze the impact of genomic instability on tumor heterogeneity by developing a method to integrate single-cell whole-genome sequencing and single-cell RNA sequencing data and subsequent visualize the results of copy number profile and DEG clusters respectively. This integration will utilize the SIMBA framework, a graph neural network, focusing on shared features such as single nucleotide variants (SNV) , identified by Monopogen, to address the challenges of copy number variation based approaches.

**Results:**

- SIMBA successfully maintained core cluster structures when applied to an unpaired multi-omics dataset, although cluster density decreased with fewer cells.
- Initial profiling across experimental batches revealed a total of 5,933 shared SNV with subsequent selection yielding 2,636 SNV for integration.
- Omic-specific clustering demonstrated distinct cellular phenotypes based on differential gene expression and copy number patterns.
- Multi-omics integration highlighted limited alignment between transcriptomic and genomic clusters, indicating potential experiment biases and the need for further refinement.
- Analysis of cluster distances revealed statistically significant differences, particularly between clusters driven by distinct copy number profiles, albeit derived from not well integrated data.

**Conclusion:**
This thesis highlights the challenges in integrating genomic and transcriptomic data, particularly when using unpaired datasets. While single nucleotide variants provided a quantifiable basis for integration, its utility in capturing the complex interplay between genomic alterations and transient transcriptional states was limited. Our findings underscore the need for innovative approaches that account for experiment biases and explore alternative strategies for multi-omics integration beyond traditional, quantifiable genomic features.

**Keywords:** multi-omics, single-cell sequencing, data integration, cancer research, genomic instability, tumor heterogeneity, scWGS, scRNAseq, single nucleotide variants

# CONTENT

# 1 BACKGROUND

## 1.1 SEQUENCING AND INTEGRATION

The evolution of sequencing technologies has revolutionized our understanding of biological systems. While early methods provided foundational insights, they were limited in throughput and resolution. The subsequent development of next-generation sequencing (NGS) offered a significant increase in throughput, enabling the bulk analysis of cell populations. With the emergence of single-cell sequencing technologies, such as scRNAseq and single-cell whole genome sequencing (scWGS), has enabled researchers to better dissect cellular heterogeneity (1). The Chien lab (2) developed functional single cell selection and isolation (fSCS) named FUNseq, a technique extending SORT-seq (3). By integrating a phototagging microscope, cellular phenotypes can be directly linked with single-cell resolution and downstream sequencing methods. This provides more detail for mapping tumor heterogeneity, enabling a clearer distinction between cell identity and the underlying transcriptomic/genomic factors driving those identities.

While the analysis of a single-cell omics dataset can be highly informative, it inherently provides only a partial view of complex biological systems (4–6). The integration of multiple omics datasets, including genomics, transcriptomics, proteomics, among others, offers the possibility of finding deeper biological insights inaccessible when examining these datasets in isolation (7). A wide array of computational integration methods have been developed in recent years, demonstrating notable progress (8). Traditionally, many integration approaches relied on statistical modelling techniques. However, with the increasing capabilities and accessibility of machine learning, neural networks (NNs), particularly graph neural networks (GNNs), have gained significant popularity for multi-omics integration (9–12). These methods have demonstrated success in cancer research, where studies integrating epigenomic (methylation and Assay for Transposase-Accessible Chromatin; ATAC) and transcriptomic profiles have helped identify new diagnostic biomarkers and potential drug targets (13). However, there is a need for a methodological approach that effectively integrates genomic and transcriptomic data, allowing a more complete understanding of the complex biological mechanisms that drive cancer, particularly in the relationship between Copy Number Variations (CNVs) and gene expression levels (14,15). This integration poses unique challenge due to the inherent differences in the types of input data: genomic data reflects the underlying DNA sequence, while transcriptomic data captures the dynamic expression of genes. Additionally, when these data come from unpaired cells, the gene expression might not yet reflect the genomic variations of the daughter cell (16). Effectively addressing these challenges enables a broader spectrum of data to be used for integration, especially from unpaired cells.

## 1.2 CANCER AS A RESEARCH MODEL

Chromosomal instability (CIN), the tendency for increased rates of structural and numerical chromosomal changes, is a hallmark of many cancer types (17). This instability, which encompasses changes in chromosome number and structure, can drive the accumulation of genetic mutations, influencing critical processes in cancer development and progression like metastasis, immune evasion, and therapeutic resistance (14,15,18). CIN arises from a wide range of defects, including errors in DNA replication, compromised cell cycle checkpoints, and failures in DNA repair pathways (19).

Substantial statistical evidence underscores the link between aneuploidy (abnormal chromosome number) and malignancies, such as lung tumors (where approximately 65% show aneuploidy) and glioblastomas (91% aneuploidy) (20,21).

Modeling the complex and dynamic heterogeneity observed in tumors presents a fundamental challenge for cancer research. Multi-omics approaches offer powerful analytical tools for addressing this problem. The integration of scWGS and scRNAseq data,

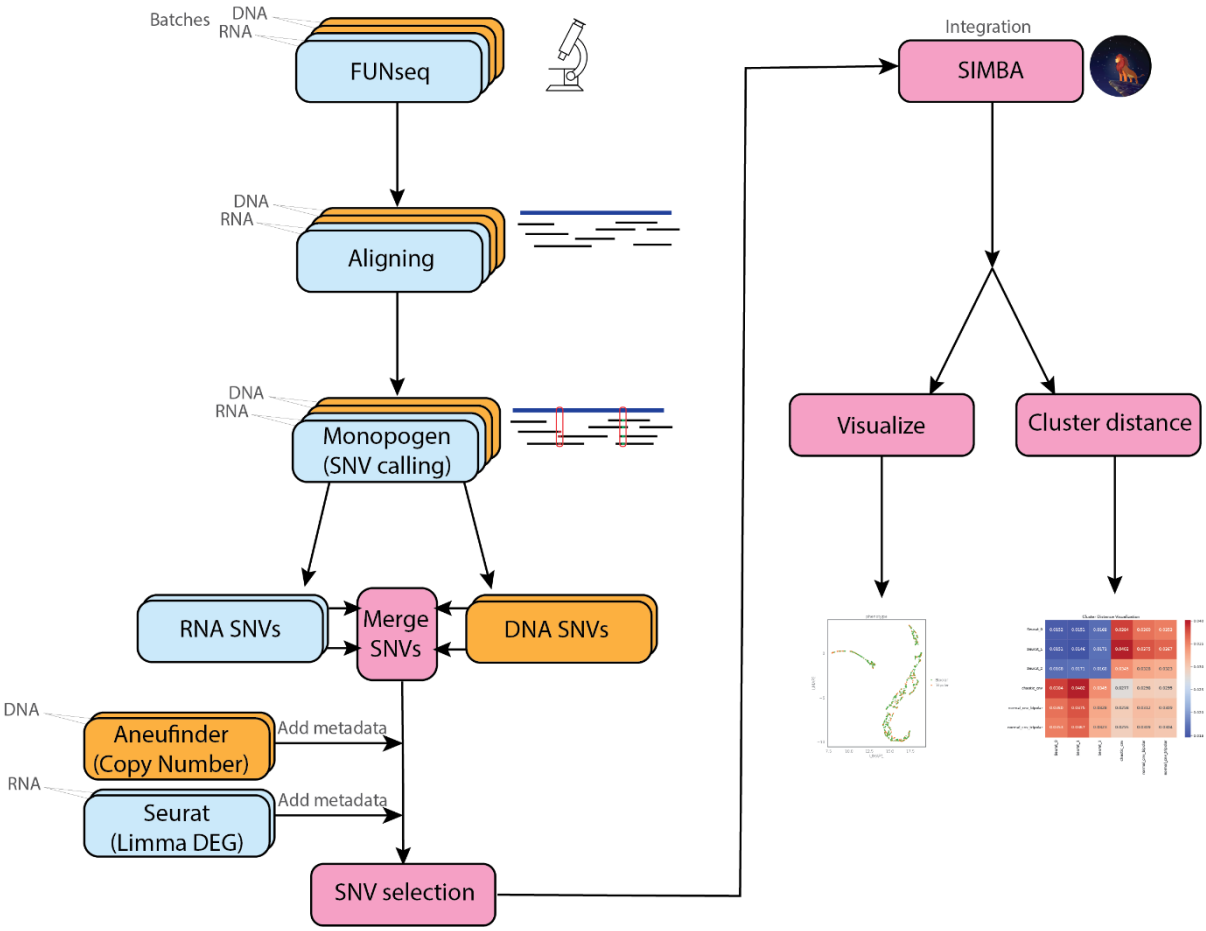



***Figure 1. Overview of the integration workflow**: Our study used MCF10A cell line to explore chromosomal instability (CIN) through FUNseq, sorting cells with tripolar and bipolar division patterns, and controls into 384-well plates. We profiled mRNA via SORT-seq and performed WGS using NlaIII-seq. RNA and DNA samples were processed using STARsolo and BWA, respectively, with quality checks and alignment against GRCh38.p14. Single nucleotide variants were identified with Monopogen and cross-analyzed in h5ad format, enabling the integration of genomic and transcriptomic data. SNV selection was performed to establish a basis for downstream integration. Seurat and Limma were used for DEG clustering and AneuFinder for CNV calling. Integration was achieved with SIMBA, creating embeddings for complex data assessment, and cluster distances were visualized to evaluate the relationships between different cell clusters.*

using CIN-prone cell lines such as MCF10A (22,23), might provide unique insights into how CIN-related genomic alterations manifest within the diverse transcriptomic profiles found within a tumor. Importantly, micronuclei and multipolarity influence gene expression levels within cells, adding another layer of complexity to the tumor microenvironment (TME) and impacting aspects such as oncogenesis, tumor progression, and therapeutic responses (24). This complexity underlines the need for advanced data analysis techniques, like multi-omics integration.

## 1.3 Research Objectives and Hypotheses

The objective of this research project is to develop an approach for the integration of scWGS and scRNAseq data (from different experiments, following FUNseq). This approach is made to potentially capture how genomic instability shapes tumor heterogeneity, ultimately leading to more effective cancer diagnostics and therapeutics. Traditional omics integration approaches often rely on CNV analysis, which can be challenging to accurately infer in single-cell sequencing data, and might be unreliable if sequencing occurs right after cell-division (16,25). Previous integration attempts using Aneufinder (26), to call genomic CNVs; NumBat (27), to infer genomic CNVs from transcriptomic reads; and CloneAlign (28), for integration, have proven to be unsuccessful in this setting. Therefore, the focus will be on exploring alternative integration strategies that leverage shared features like single

nucleotide variants (SNVs) as the basis for integration.

**Objective 1:** Data Preparation will entail the identification and annotation of SNVs.
Additionally, essential preprocessing steps will be performed to ensure the quality and compatibility of the multi-omics datasets.

**Objective 2:** Data Integration will prioritize the development or careful adaptation of methodologies designed for the integration of scWGS and scRNAseq data, together with a comprehensive analysis of the expected challenges and potential solutions.

# 2 METHODS

In this section we discuss the methods used for integration. Starting with data collection, followed by pre-processing and feature selection. Finally integration and visualization. A schematic overview is given in Figure 1.

## 2.1 DATASET

Our study utilized cells from the MCF10A cell line, recognized for its utility in investigating non-tumorigenic human mammary epithelial cell behavior (22,23). To isolate cells demonstrating chromosomal instability (CIN), the FUNseq pipeline was employed (29). This technique facilitated the identification of cells undergoing unusual division patterns, specifically those forming tripolarity, as well as a subset of bipolar cells and randomly selected controls. Subsequent to their identification, these cells were sorted into 384-well plates using Fluorescence-Activated Cell Sorting (FACS). After sorting, the library was prepared.

The initial phase of our experiment involved mRNA profiling through SORT-seq (30), executed by Single Cell Discoveries (31), followed by WGS employing NlaIII-seq (32), conducted by the Single Cell Core facility (33).

In total, the dataset encompasses four 384-well plates, evenly divided between RNA and DNA samples. A detailed description and allocation of the samples are presented in Table 1.
*All these steps were performed by other members from the Chien lab, prior to the start of this project.*

## 2.2 PRE-PROCESSING OF RNA SAMPLES

For the preprocessing of RNA samples, we used STARsolo, an extension of STAR designed for single-cell data (version 2.7.10a) (34). Reads were aligned against the GRCh38.p14 human reference genome (35), using a comprehensive annotation file to facilitate splice junction discovery and gene quantification. The full command prompt with parameters is in 'star_align.sh'. Below are the most crucial parameters explained.

The handling of cell barcodes and unique molecular identifiers (UMIs) is critical for correct interpretation of the data. The reads from R2 files contained the cell barcodes and UMIs, indicating the origin of each transcript, while R1 files were used for sequence information. We specified the start position and length of cell barcodes (CB) and UMIs within the reads to accurately demultiplex the data and correct for PCR duplicates, ensuring the reliability of transcript quantification.

The '--soloFeatures GeneFull' parameter was employed to quantify gene expression levels across the entire gene body, rather than just at the 3' end, providing a more comprehensive view of gene activity. To manage multimappers, which are reads that map to multiple locations in the genome, we used the '--soloMultiMappers Unique EM' setting, allowing only uniquely mapped reads to contribute to gene expression counts directly, while employing an Expectation-Maximization (EM) algorithm to distribute multimappers in proportion to the unique mapping evidence.

The output from STARsolo is the gene counts per cell (raw and filtered), and a binary alignment map (BAM) file for all the reads. The BAM file was split into individual files per cell, based on the barcode (script: 'split_bamfiles.py'). Finally the read groups were adjusted so that each BAM file contained the cell ID, with phenotype.

## 2.3 PRE-PROCESSING OF WGS SAMPLES

For the pre-processing of our scWGS samples, we adopted the protocol established by the van Oudenaarden group (36). The process began with a quality check of the raw

**Table 1:** Overview of cells used in the project. The data is selected based on live cell imaging and observing the desired phenotype. These cells are subsequently phototagged and sorted using FACS sort. Finally they are sequenced.

| Run/plate | Omic | Tripolar | Bipolar | Control |
|---|---|---|---|---|
| **s143** | RNA | 54 | 160 | 144 |
| **s145** | RNA | 92 | 148 | 0 |
| **CHI-006** | DNA | 81 | 144 | 144 |
| **CHI-007** | DNA | 50 | 144 | 168 |

sequencing data, using FastQC (37) to identify any issues. Special attention was given to NlaIII restriction enzyme (RE) sites, CATG (32).

Following the quality assessment, we proceeded with demultiplexing and trimming the reads, utilizing the BuysDB scripts for both tasks. This step was crucial for removing adapter sequences and ensuring that only high-quality sequences were retained for alignment.

The alignment of these processed reads to the same human reference genome as the scRNAseq reads, GRCh38.p14, was performed using the Burrows-Wheeler Aligner (38) (BWA, version 0.7.17-r1188).

Next, duplicate reads were identified and removed using Picard tools (39) (version 3.1.0), a necessary step for minimizing biases in variant calling. Additionally, read groups were assigned to organize the data more efficiently and facilitate downstream processing.

Completing the pre-processing, we split the aligned BAM files into individual cell-specific files (script: 'split_bamfiles.py'). This was done to isolate the data for each cell, mirroring the approach taken with our RNA sequencing samples, and allowing for cell-specific genomic analyses. This step ensures that each cell's genomic information is distinctly categorized, laying the groundwork for in-depth single-cell genomic exploration.

## 2.4 Single Nucleotide Variant calling and processing

We used SNVs as features for integration. We employed Monopogen (40), a tool specifically designed for sparse single-cell sequencing data. Monopogen's architecture seems particularly well-suited for our analysis, allowing for the efficient detection of SNVs across both scWGS and scRNA-seq datasets. This dual applicability facilitates the integration of genomic and transcriptomic data within a singular analytical framework, setting the stage for the downstream integration using SNVs as a shared feature (script: 'Monopogen.sh').

Monopogen begins by preprocessing BAM files for quality control, followed by SNV discovery using pooled read alignments and genotype likelihood calculations with Samtools mpileup (41). The tool refines germline variant calls using linkage disequilibrium data from the 1KG3 database (42) and employs sequencing error modeling to differentiate between germline and somatic SNVs.

Subsequent to SNV calling, the variant call format (VCF) files generated were converted into the h5ad format (43), leveraging a custom script (script: convertVCF_h5ad.py). The h5ad format, rooted in the hierarchical h5 standard, allows the management of extensive datasets composed of multidimensional arrays. This

**Table 2:** Overview of SNV identification across four samples (RNA: s143 and s145; DNA: CHI-006, CHI-007). The table presents the total number of cells analyzed, the count of SNVs detected, and the proportion of SNVs with allele frequencies (AF) greater than 0.1 and 0.3. High allele frequency thresholds provide insights into prevalent genetic variations within each sample.

| | s143 | s145 | CHI-006 | CHI-007 |
|---|---|---|---|---|
| ***No.* Cells** | 214 | 240 | 225 | 194 |
| ***No.* SNVs** | 33753 | 64331 | 2366432 | 1530854 |
| ***No.* SNVs (AF > 0.1)** | 25103 (74%) | 48989 (76%) | 2194868 (93%) | 1416918 (93%) |
| ***No.* SNVs (AF > 0.3)** | 19792 (59%) | 38681 (60%) | 1796796 (76%) | 1170323 (76%) |

format is especially advantageous for organizing scientific data efficiently, facilitating both the inclusion of enriched metadata and faster data retrieval. In our converted dataset, the matrix encodes which genotype a particular cell (columns) has for what SNV (row). This format follows a straightforward numerical scheme, where '1' signifies a homozygous reference (0/0), '2' a heterozygous variant (0/1), and '3' a homozygous variant (1/1). Cell metadata incorporated includes extra information, labelled as phenotype (bipolar, tripolar, or control), batch information, and cell barcodes, while SNV metadata encompasses chromosomal location, reference alleles, alternate alleles, and allele frequency (AF). Utilizing the AnnData format, our methodology ensures seamless integration and manipulation of comprehensive genomic data alongside phenotype metadata.

## 2.5 Selecting SNVs

We applied a selection process to identify which SNVs are suitable to use as foundation for integration. The following process is repeated for each individual SNV, where the genotypes across datasets are compared.

Given a set of genotypes $G = \{1, 2, 3\}$, where the numbers refer to the genotype matrix from the SNV matrix, for each $gt \in G$, we calculated the proportion ($P$) per dataset $D \in \{dna, rna\}$ in:

$$P_{gt}^{D} = \frac{S_{gt}^{D}}{\sum_{i \in G} S_{i}^{D}}$$

where $S_{gt}^{D}$ are the counts of a particular genotype within a dataset and $\sum_{i \in G} S_{i}^{D}$ represents the total occurrences of that SNV within a dataset.

Next, these proportions were used to calculate a ratio as follows:

$$R_{gt} = \frac{P_{gt}^{dna}}{P_{gt}^{rna}}$$

We then applied the following constraints:

$R_{gt} \in [1 - a, 1 + a]$, which ensures that the ratio is within a-interval from 1.

$P_{gt}^{D} > b$ , which ensures that small proportions of any genotype within any dataset are disregarded.

$P_{gt}^{D} \leq 1 - c$ , which ensures that there is variation within the genotype of that SNV.

Within the constraints, $a$ regulates the allowed variation of proportions between datasets. We set $b$ as a small threshold to filter out noise attributed to very low proportions of cells, ensuring that only biologically relevant variants are considered. The parameter $c$, a very small number, is determined empirically to allow for sufficient intra-genotype variation while excluding outliers that may represent sequencing errors or other artifacts. To choose

the right values, multiple heatmaps of the SNV profile were created to show the selection of accepted and rejected SNVs.

The final selection of SNVs, as determined by these criteria, is then subjected to downstream integration.

### 2.6 SEURAT DEG CLUSTERING

To get the differentially expressed genes (DEGs), we had to use the gene counts instead of the BAM file from the STARsolo output. This was done following a script used for previous analyses in the lab. Quality control (QC) is essential to filter out low-quality cells with erroneous transcript counts, which if unaddressed, can adversely affect downstream analyses. We utilized the Seurat library (44), which calculates QC metrics like nFeature_RNA and ncount_RNA to identify potential empty wells or doublets, and additional metrics like percent_MT and percent_RIBO to assess mitochondrial and ribosomal RNA content, as outlined in Ilicic et al. (2016) (45). Following QC, we performed normalization using Seurat's NormalizeData and SCTransform to correct for sequencing depth variance and prevent overfitting, respectively (46). Highly variable features were identified using the FindVariableFeatures method, and scaling was applied to facilitate PCA, despite the ongoing debate over its necessity in RNAseq data analysis. We did not perform batch correction as per recommendations specific to our sequencing setup. Differential gene expression analysis was conducted using Limma lmfit (47), and clustering of the final selected samples was achieved using Seurat clustering.
The seurat clusters (*Seurat_0*, *Seurat_1*, *Seurat_2*) were then assigned to the respective cell in the AnnData metadata dataframe.

### 2.7 ANEUFINDER CNV CALLING

For the analysis of CNVs in scWGS data, individual BAM files of scWGS cells were analyzed using AneuFinder (26). This tool leverages either a Hidden Markov Model (HMM) or a binary bisection method for the precise estimation of copy numbers, breakpoints, and hotspots. Our approach, as derived from previous experiments in our lab, included the utilization of a specific script for both the identification of CNVs and the implementation of comprehensive QC.

This QC protocol was structured in two distinct stages to ensure data quality. Type I QC focused on removing low-quality cell clusters to refine the dataset, while Type II QC imposed strict thresholds based on spikiness (< 0.4) and Bhattacharyya distance (> 0.8),which measures the degree of overlap between two statistical samples of populations, to filter out cells not meeting quality standards. The CNV profile of the remaining cells were visualized using genome-wide heatmaps. Cells exhibiting significant fluctuations in their copy number were labeled as '*chaotic_cnv*', whereas cells with stable profiles were categorized as '*normal_cnv*' along with a phenotype suffix. This classification was finally assigned to their respective cells in the AnnData metadata dataframe.

### 2.8 SIMBA

To further investigate the complex interplay between genotype, gene expression, and copy number variations, we employed SIMBA (Single cell embeddings along with features) (48). SIMBA is a method to embed cells along with their defining features such as gene expression, transcription factor binding sequences and chromatin accessibility peaks into the same latent space. The joint embedding of cells and features allows SIMBA to perform various types of single cell tasks, including but not limited to single-modal analysis (e.g. scRNA-seq and scATAC-seq analysis), multimodal analysis, batch correction, and multi-omic integration.

We used SIMBA's generalized graph generation for automatic matching of entities and relations between datasets beyond the pre-specified features. Using this, we could infer the similarities between cells from our omics modalities based on the SNV profile. The result was a matrix where columns represented RNA cells, rows represented DNA cells, and values indicated similarity scores. SIMBA then modeled the single-cell data as a heterogeneous graph. Cells (from both

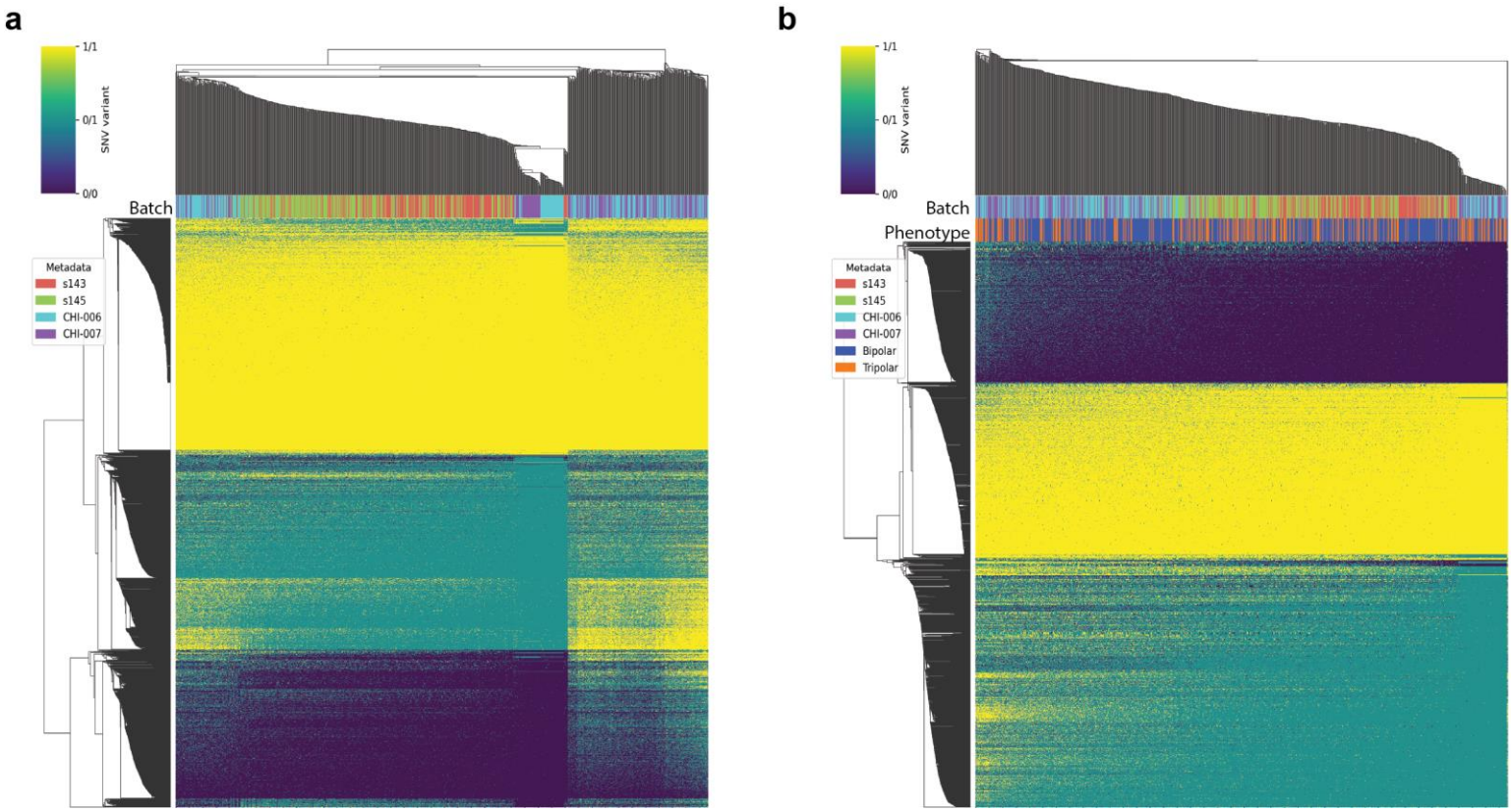



***Figure 2: Heatmap of SNV profile. (a)*** *5933 SNVs, where each color in the heatmap represents a genotype. On top the batch is shown.* ***(b)*** *The filtered heatmap consists of 2636 SNVs, besides the batches, it also contains the mitosis type (Tripolar/Bipolar). Rows are SNVs and columns are cells.*

modalities) became individual nodes, while relevant features (SNVs) formed additional nodes. Edges are established between cell nodes and feature nodes based on their observed relationships. SIMBA employed a graph neural network (GNN) to learn embeddings for each node in the graph. Through iterative updates influenced by the graph structure, the GNN generated representations for DNA cells, RNA cells, and SNV features in a shared latent space.

Multiple UMAPs were created, with the SNVs being hidden as they are shared features. This allowed us to explore different parameter combination and determine the optimal number of nearest neighbors for our analysis. This choice was important because the number of nearest neighbors significantly impacts UMAP visualizations. With too few neighbors, local data structure can be overemphasized. On the other hand, using too many neighbors can result in over-generalization. By setting the neighbors is to 15, we aimed to strike a balance between preserving local structure and revealing global trends within our dataset. Different variables were visualized as colors: 1) Seurat clusters and CNV clusters, 2) batches, and 3) phenotypes.

## 2.9 CLUSTER DISTANCES

To assess the relationship between Seurat and CNV clusters, the analysis involves calculating Euclidean distances using 50-dimensional embeddings from SIMBA. This approach provides a quantitative measure of similarity or dissimilarity between clusters. The average distances are visualized in a heatmap, where the diagonal shows the mean intra-cluster distance, indicating the homogeneity within clusters. Furthermore, a violin plot is used to display the full range of distance distributions, offering a nuanced view of cluster separations. Given the non-normal distribution of the data, as determined by the Shapiro-Wilk test, the Mann-Whitney U test, coupled with the Benjamini-Hochberg procedure for multiple test correction, is employed to assess statistical significance.

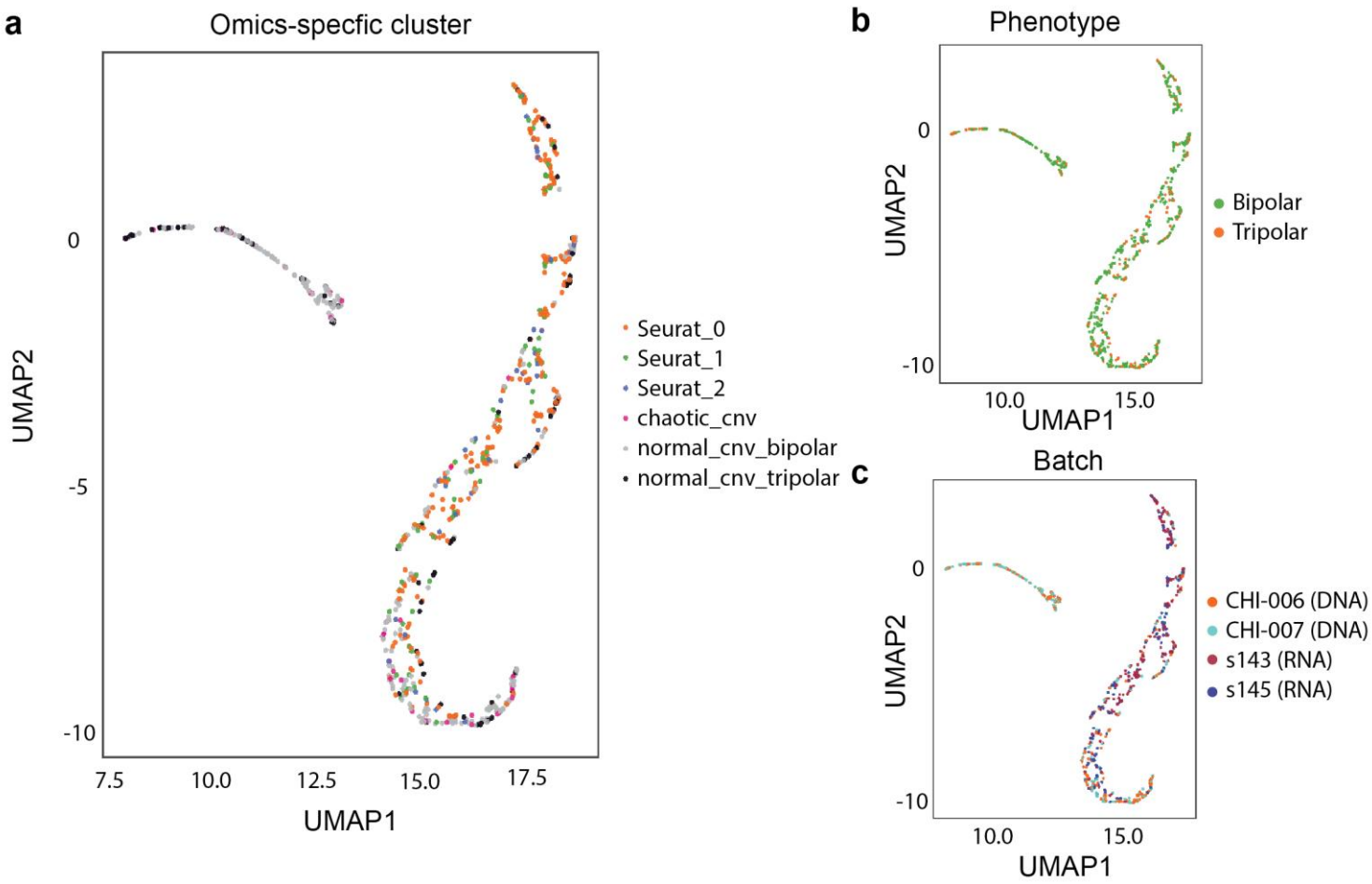



***Figure 3.*** *Representation of cell embeddings using UMAP.* ***(a)*** *Illustrates the omics-specific clustering of cells, with colors representing different clusters identified using the Seurat algorithm and two clusters characterized by copy number variations (CNVs), one labeled as 'chaotic_cnv' and the others as 'normal_cnv_bipolar' and 'normal_cnv_tripolar'.* ***(b)*** *depicts the phenotype classification of cells, with colors distinguishing between 'Bipolar' and 'Tripolar' states.* ***(c)*** *Shows batch effects across different cell samples, labeled as 'CH-006', 'CH-007', 's143', and 's145'. All UMAP visualizations were generated using SIMBA with a setting of 15 nearest neighbors for dimensionality reduction.*

### 2.10 Code Availability

The scripts used for the data analysis in this study are available in a public Github repository: https://github.com/Uhm-J/MSc_multiomics-integration.

# 3 Results

This thesis explores the integration of transcriptomic and genomic data to investigate cellular phenotypes within our dataset. We begin by assessing the performance of SIMBA with unpaired multi-omics data. Subsequently, we profile SNVs to establish a basis for cross-modality analysis. Omic-specific clustering reveals distinct phenotypes, and integrated analysis highlights the relationships and differences between these clusters.

### 3.1 SIMBA subsample results

Upon assessment of SIMBA with an adapted dataset from a 10x genomics example—where the dataset was specifically modified to simulate unpaired multi-omics by including only one type of omic data per cell—the UMAP visualizations maintained the core cluster structures, despite the dataset's simplification. This observation is visually represented in Supplementary Figures 1a (11909 paired cells) and 1b (11909 unpaired cells), where despite the reduction to unpaired status, the essential cluster structures are preserved. The identification of 2713 shared features between scATAC peaks (in genes) and gene counts in scRNA provides a promising starting point for exploring data integration. Further subsampling to 1,600 cells resulted in sparser

cluster distributions and a more linear arrangement, as depicted in Supplementary Figure 1c.

### 3.2 SNV PROFILING AND FILTERING

To establish a foundation for robust multi-omics analysis, our initial focus centered on the quantification and characterization of SNVs within our datasets. Table 2 provides a detailed overview of the SNVs identified across individual samples, delineating the total counts and the proportions with allele frequencies (AF) above 0.1 and 0.3. Supplementary Table 1 further explores the intersections of these SNVs between datasets.

To ensure data quality, we merged DNA datasets and intersected them with RNA datasets using an AF threshold >0.1, yielding 5,933 shared SNVs (Figure 2a). Subsequently, we applied a selection process to assess consistency across batches, focusing on the proportion of genotypes for each SNV. This additional layer of filtering refined it to 2,636 high-confidence SNVs (Figure 2b), ensuring that the variations selected for subsequent analysis were not only prevalent but also stable across experimental batches. The rejected SNVs are presented to illustrate the selectivity of the filtering process (Supplementary Figure 2).

### 3.3 OMIC-SPECIFIC CLUSTERS

Our results delineate distinct cellular phenotypes, leveraging insights from both transcriptomic and genomic datasets. In Supplementary Figure 3, UMAP plots illustrate clusters defined by differential gene expression, showcasing the transcriptional diversity within our cell population.

Complementing the clusters acquired by transcriptomics, Supplementary Figure 4 introduces a genomic perspective with a heatmap displaying aneuploidy across chromosome bins. Through DNA CNVs analysis performed by Aneufinder, we categorized cells into three distinct genomic groups: 'normal_cnv_bipolar', 'normal_cnv_tripolar', and 'chaotic_cnv'. This genomic classification, based on CNV patterns, serves as a pivotal indicator of underlying genomic instability and variation, although it remains an approximation of the complex genomic landscape.

Integrating cell metadata with these omic-specific clusters [1] enables a comprehensive visualization of our findings. Supplementary Table 2 enumerates the cell count within each cluster category.

### 3.4 MULTI-OMICS INTEGRATION

The integrated UMAP visualization in Figure 3 showed important features of the integration multi-omics dataset. Notably, there was a distinct separation between a subset of DNA samples and the RNA cells (a). While the phenotypic distinction between bipolar and tripolar cells was not clear-cut on the UMAP (b), the omic-specific clustering showed a clear delineation for one group of DNA cells (c). Another subset of DNA samples showed better integration with the RNA cells. Cluster distance quantification

The quantification of distances between omic-specific clusters helps to decipher the structure and consistency within our integrated multi-omics dataset. Figure 4a illustrates the mean distances between each pair of clusters, shedding light on the varying degrees of proximity and, thus, the similarity between clusters. This analysis is underpinned by individual distance measurements, as shown in Supplementary Figures 5.

Figure 4b visualizes the distribution of distances for pairwise cluster comparisons, revealing a generally significant differentiation between clusters, as confirmed by the Mann-Whitney U-test with Benjamini-Hochberg

[1] The term 'omic-specific clusters' refers to the CNV labels and 'Seurat' clusters for genomic and transcriptomic data, respectively

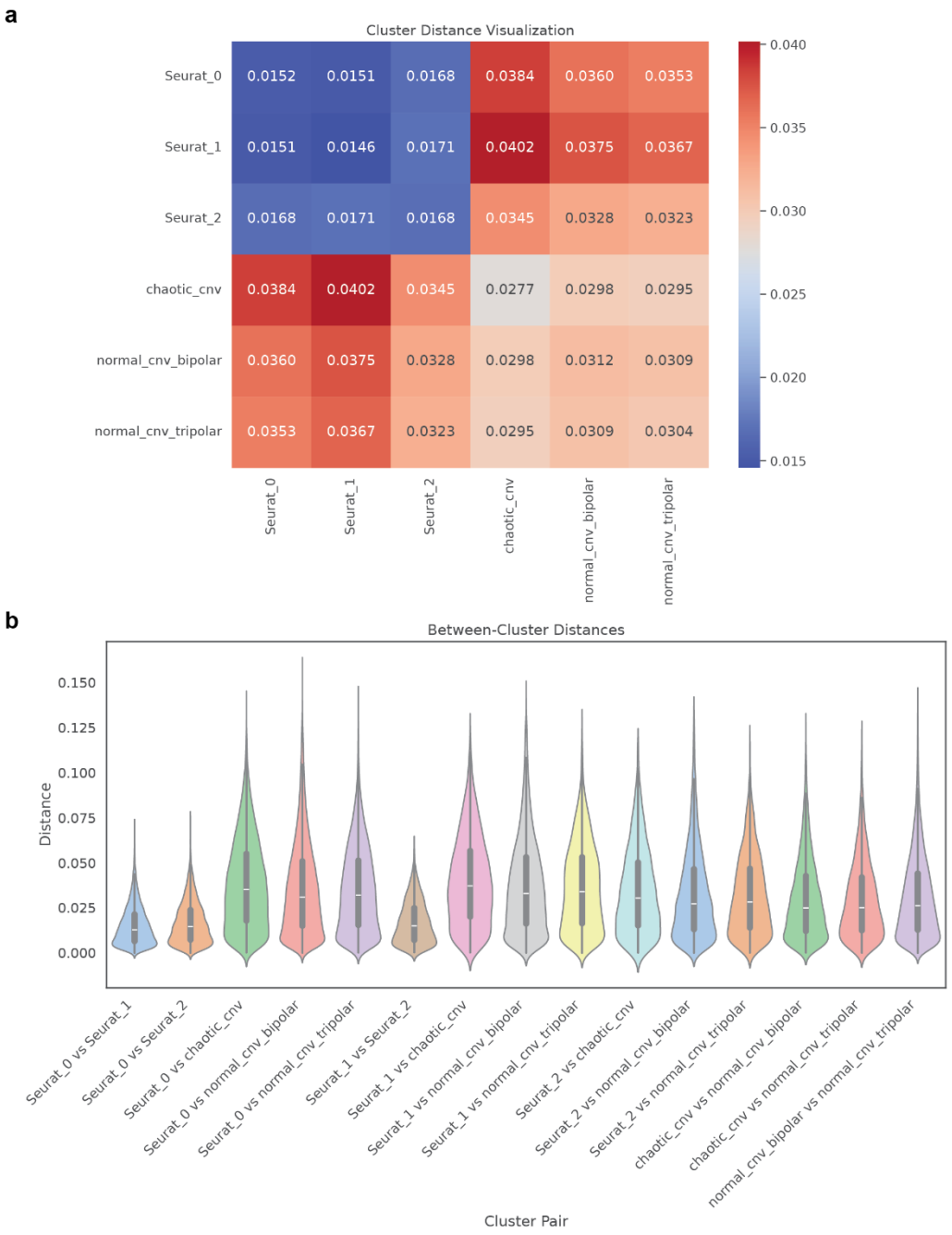



***Figure 3.*** *Euclidean distance between omic-specific clusters.* ***(a)*** *Heatmap visualizing the Euclidean distances between clusters. The diagonal depicts the intra-cluster distances.* ***(b)*** *The violin plot of the distances between clusters.*

correction (details in Supplementary Table 3). Notably, cells grouped by transcriptomic data within Seurat clusters exhibit significantly lower distances compared to those grouped within CNV clusters, highlighting a closer similarity among transcriptomically defined groups. The largest mean distances were observed between specific Seurat and *chaotic_cnv* clusters [mean difference: 0.286, p: 1.12e-303], with identical distances noted between certain Seurat clusters and both *normal_cnv_bipolar* and *normal_cnv_tripolar* groups. This pattern persists with *Seurat_2*, where the average distance to *chaotic_cnv* exceeds that to *normal_cnvs*, though these distances are less pronounced than those observed with *Seurat_0* and *Seurat_1* clusters.

# 4 DISCUSSION

## 4.1 UNPAIRED SIMBA PERFORMANCE

In important aspect is the assessment whether SIMBA is fit for unpaired integration with lower sample sizes. After running the integration, the core cluster structures remain identical when being unpaired (Supplementary Figure 1). The relationship between cell types as well, even in the subsampled dataset. This finding confirms SIMBA doesn't rely on the data from different modalities to be paired. However, it should be noted that the clusters themselves are becoming less 'circular' and more 'linear'. A possible explanation for this, is that the clusters become less dense, and thus decreasing intra-cluster variation. Another possibility is that the number of features is getting too large compared to the number of cells. In our paired integration, the cell-to-feature ratio stands at 8.7:1, contrasting with a halved ratio in the unpaired data and further diminishing to 0.6:1 in the subsampled set. This gradient of ratios underscores a potential optimal threshold, suggesting greater ratios facilitate more defined and biologically informative cluster configurations. Overall, the performance of SIMBA on unpaired cells is successful, with a note that the number of features should be balanced in relation to the number of cells, as well a more 'linear' cluster might appear.

## 4.2 ROLE OF SNVS

Initial attempts to integrate our datasets resulted in highly linear UMAP visualizations, reminiscent of those seen in Supplementary Figure 1c, albeit more pronounced. This observation led to the hypothesis that the cell-to-feature ratio was insufficient. In response, we employed a SNV selection technique aiming not only to curtail the feature count but also to mitigate variation between modalities, thereby fostering increased compatibility. This approach culminated in the identification of 2,636 SNVs, a refined set for subsequent analyses.

Interestingly, heatmap clustering still delineated clear modality-based separations, despite the reduction in feature count (Figure 2b). A peculiar observation was the positioning of a subset of cells from scWGS closer to those from scRNAseq. This was contrary to expectations of a more evenly distributed separation and may be attributed to the pronounced similarity within this subset of cells. Notably, our analysis did not reveal a direct correlation between these SNVs and mitosis type or chromosomal instability (CIN). Specifically, the SNVs assessed were those present within coding regions due to their detectability through RNA overlap. Given that structural variants (SVs) and their associated SNVs, which are typically enriched around 1 Mb proximal to SV sites and are indicative of DNA repair deficiencies, are predominantly located outside of coding regions, we did not foresee a link between coding region SNVs and mitotic behaviors (49,50).

One anticipated advantage of employing SIMBA's graph neural network was its potential to model the nuanced relationships between differing modalities (12). Indeed, Figure 3c hints at SIMBA's capability to capture more complex inter-modal interactions, though a significant divide between the bulk of scWGS and scRNAseq samples persists. This divide underscores the inherent challenges in bridging the gap between genomic and transcriptomic landscapes, even when advanced analytical tools like SIMBA are utilized.

## 4.3 OMIC-SPECIFIC CLUSTERING

The exploration of omic-specific clustering revealed unexpected patterns that diverged from our initial hypotheses. Contrary to the expectation that Seurat clusters would align closely with specific CNV profile clusters, our analysis revealed a lack of direct correlation between these omic-specific clusters. This discrepancy prompts a critical examination of underlying factors influencing clustering outcomes and the integration of multi-omics data.

A significant observation is the potential impact of experiment effects on the clustering patterns. Given that batches across modalities are processed at different times, it is plausible that these procedural variations introduce

unique lineage signals into the data. Such effects were initially expected to be minimal or negligible; however, the distinct clustering outcomes suggest a more pronounced influence than anticipated. This divergence underscores the complexity of integrating multi-omics data and highlights the necessity of accounting for batch effects and temporal variations in experimental design and analysis.

The strategy employed to refine the integration of SNVs, through ratio filtering, aimed at reducing feature complexity and enhancing modality compatibility, presents its own set of challenges. While this approach successfully narrowed down the SNV count to 2,636 variants, the resulting heatmap clustering still exhibited modality-specific separations. Notably, a minor subset of cells from scWGS demonstrated closer affinity to the scRNAseq cells (Figure 3).

This observation may hint at underlying similarities among a specific group of cells, yet the overarching separation indicates a broader disconnect between genomic and transcriptomic profiles. Furthermore, the attempt to mitigate modality distinctions by filtering out SNVs that accentuate these differences inadvertently introduces a new bias, skewing the dataset towards an overcorrection. This adjustment, while well-intentioned, may obscure genuine biological distinctions between modalities, complicating the interpretation of integration outcomes.

### 4.4 CLUSTER DISTANCES

Our examination of cluster distances served as a validation step, reinforcing our preceding observations. The analysis focused on the mean intra-cluster distances—represented on the diagonal in Figure 4a—and revealed that these distances are consistent with the distances observed between different clusters within the same modality, both for RNA-RNA and DNA-DNA cluster comparisons. The uniformity in cluster distances in both RNA and DNA clusters underscores a modality-specific clustering effect.

Further validation came from the statistical analysis, where the Mann-Whitney U test results presented in Table 5 confirmed the significance of the distribution differences between clusters. This significance is attributed primarily to the existence of two distinct DNA cell clusters, each characterized by unique Copy Number Variation (CNV) profiles. The presence of these diverse CNV profiles within the DNA clusters introduces a pronounced variance, underscoring the heterogeneity of genomic alterations and their impact on cluster formation.

### 4.5 SNVS AS FOUNDATION FOR INTEGRATION

Despite the methodological rationale behind employing SNVs for integration, from a biological perspective, the decision warrants re-evaluation. Integrating multi-omics data using SNVs, while technically feasible, might not always effectively capture the complex interplay between genomic and transcriptomic variations in a manner that accurately reflects true biological processes. This concern arises partly from the inherent limitations of SNVs in representing the broader genomic context, especially considering the nuanced regulatory and functional implications of genomic alterations.

An essential consideration in this discourse is the distinct nature of genomic and transcriptomic information—the former being largely static, reflecting the genetic blueprint of a cell without direct indication of its current state, except for mutations, and the latter being highly dynamic, varying in response to environmental cues and cellular conditions (51). This dynamic nature of the transcriptome serves as a real-time reflection of the cell's state, offering insights into the cell's functional activity at any given moment. The static nature of the genome, while fundamentally informative, does not capture these transient states, providing a more unchanging backdrop against which cellular dynamics unfold.

Moreover, it is important to note that, in our study, all cells were sequenced post-cell division, implying a somewhat homogenized state regarding cell cycle phase. However, this does not fully encapsulate the entirety of a cell's condition or its potential trajectory. Previously, Copy Number Variations (CNVs)

were explored as a basis for integration, given their significant impact on gene expression and cellular phenotype. The exploration of SNVs emerged as an alternative avenue, driven by a desire to exhaust all potential methods for multi-omics integration. This approach, however, was undertaken with reservations regarding its biological significance, further complicated by the contrasting characteristics of the genomic and transcriptomic data.

### 4.6 FUTURE DIRECTIONS

In light of the challenges and insights garnered from our current study, several promising avenues for future research emerge. Firstly, an exploration utilizing single-cell Genomics and Transcriptomics sequencing (scG&T-seq), where genome and transcriptome data are paired from the outset, presents a compelling opportunity (52). This approach would not only circumvent the limitations associated with unpaired data but also allow for a direct assessment of potential biases in the integration process, leveraging the 'ground truth' inherent in paired data to validate the effectiveness of integration methodologies. Furthermore, the significant impact of experimental variables—such as batch effects and the variability introduced by temporal variance—on clustering outcomes has been underscored. This observation points to the necessity for rigorous investigations into the optimization of experimental designs and analytical strategies to mitigate these effects, ensuring that the data accurately reflects biological realities rather than artifacts. Additionally, there is a pressing need to explore alternative methods for genome-transcriptome integration. Moving beyond conventional strategies to incorporate innovative computational techniques or novel biological insights could unveil more nuanced and biologically relevant patterns within multi-omics datasets. By embracing these directions, future research can advance our understanding of cellular mechanisms at a molecular level, enhancing the precision and applicability of multi-omics analyses in uncovering complex biological phenomena.

# 5 CONCLUSION

In conclusion, this thesis explores multi-omics integration, particularly through the lens of SNVs, and it has unveiled both the potential and the limitations of current methodologies in bridging the complex interplay between genomic and transcriptomic data. Despite the methodological rationale and the initial promise of these genomic markers, our efforts to achieve a meaningful integration have confronted significant challenges. The static nature of genomic data, set against the dynamic transcriptome, presents an inherent obstacle to capture the full spectrum of cellular states and processes. Furthermore, our attempts to utilize conventional quantifiable elements of genomic data, previously CNVs and in this project SNVs, have not yielded the depth of insight anticipated, suggesting that these markers alone may not suffice for effective multi-omics integration of these modalities.

This realization calls for a shift in our approach to multi-omics integration. It underscores the necessity of going beyond traditional quantifiable genomic features and embracing innovative computational techniques and novel biological insights.

Our study, while highlighting the challenges and limitations of current integration strategies, also illuminates the path toward future exploration. It advocates for a broadened perspective and a relentless pursuit of alternative approaches that promise to enhance our understanding of complex biological systems.

# 6 ACKNOWLEDGEMENTS

I would like to express my gratitude to Pin-Rui (Ray) Su from the MP Chien lab for his invaluable contribution to cell selection, sequencing, and Seurat Clustering. My thanks also go to my supervisor Mayte Lopez-Cascales for her assistance and support throughout this project. I am especially thankful to Li You for providing access to high-end computing resources and for the guidance in formalizing the SNV selection process.

# 7 REFERENCES


1. Wen L, Tang F. Recent advances in single-cell sequencing technologies. Precis Clin Med. 2022 Feb 24;5(1).
2. http://www.mpchienlab.org/ [Internet]. 2024. Chien Lab.
3. You L, Su PR, Betjes M, Rad RG, Chou TC, Beerens C, et al. Linking the genotypes and phenotypes of cancer cells in heterogenous populations via real-time optical tagging and image analysis. Nat Biomed Eng. 2022 Mar 17;6(5):667–75.
4. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform. 2018 Nov 27;19(6):1370–81.
5. Bravo González-Blas C, Quan X, Duran-Romaña R, Taskiran II, Koldere D, Davie K, et al. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Mol Syst Biol. 2020 May 19;16(5).
6. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018 Jan 11;36(1):70–80.
7. Wekesa JS, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Front Genet. 2023 Jul 20;14.
8. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016 Dec 20;17(S2):S15.
9. Chen H, Ryu J, Vinyard ME, Lerer A, Pinello L. SIMBA: single-cell embedding along with features. Nat Methods. 2023 May 29;
10. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022 Oct 2;40(10):1458–66.
11. Wen H, Ding J, Jin W, Wang Y, Xie Y, Tang J. Graph Neural Networks for Multimodal Single-Cell Data Integration. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2022. p. 4153–63.
12. Yi HC, You ZH, Huang DS, Kwoh CK. Graph representation learning in bioinformatics: trends, methods and applications. Brief Bioinform. 2022 Jan 17;23(1).
13. Ortega MA, Poirion O, Zhu X, Huang S, Wolfgruber TK, Sebra R, et al. Using single-cell multiple omics approaches to resolve tumor heterogeneity. Clin Transl Med. 2017 Dec 28;6(1).
14. Sansregret L, Vanhaesebroeck B, Swanton C. Determinants and clinical implications of chromosomal instability in cancer. Vol. 15, Nature Reviews Clinical Oncology. Nature Publishing Group; 2018. p. 139–50.
15. van Jaarsveld RH, Kops GJPL. Difference Makers: Chromosomal Instability versus Aneuploidy in Cancer. Vol. 2, Trends in Cancer. Cell Press; 2016. p. 561–71.

16. Funnell T, O'Flanagan CH, Williams MJ, McPherson A, McKinney S, Kabeer F, et al. Single-cell genomic variation induced by mutational processes in cancer. Nature. 2022 Dec 1;612(7938):106–15.

17. Bakhoum SF, Cantley LC. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. Cell. 2018 Sep;174(6):1347–60.

18. Wilhelm T, Said M, Naim V. DNA Replication Stress and Chromosomal Instability: Dangerous Liaisons. Genes (Basel). 2020 Jun 10;11(6):642.

19. Heng HH, Bremer SW, Stevens JB, Horne SD, Liu G, Abdallah BY, et al. Chromosomal instability (CIN): what it is and why it is crucial to cancer evolution. Cancer and Metastasis Reviews. 2013 Dec 19;32(3–4):325–40.

20. Choma D, Daurès JP, Quantin X, Pujol JL. Aneuploidy and prognosis of non-small-cell lung cancer: a meta-analysis of published data. Br J Cancer. 2001 Jul 3;85(1):14–22.

21. Duijf PHG, Schultz N, Benezra R. Cancer cells preferentially lose small chromosomes. Int J Cancer. 2013 May 15;132(10):2316–26.

22. Soule HD, Maloney TM, Wolman SR, Peterson WD, Brenz R, McGrath CM, et al. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. Cancer Res. 1990 Sep 15;50(18):6075–86.

23. Witt AE, Hines LM, Collins NL, Hu Y, Gunawardane RN, Moreira D, et al. Functional Proteomics Approach to Investigate the Biological Activities of cDNAs Implicated in Breast Cancer. J Proteome Res. 2006 Mar 1;5(3):599–610.

24. Gao C, Su Y, Koeman J, Haak E, Dykema K, Essenberg C, et al. Chromosome instability drives phenotypic switching to metastasis. Proc Natl Acad Sci U S A. 2016 Dec 20;113(51):14793–8.

25. Zheng M, Hu Y, Gou R, Wang J, Nie X, Li X, et al. Integrated multi-omics analysis of genomics, epigenomics, and transcriptomics in ovarian carcinoma. Aging. 2019 Jun 29;11(12):4198–215.

26. Bakker B, Taudt A, Belderbos ME, Porubsky D, Spierings DCJ, de Jong T V., et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biol. 2016 Dec 31;17(1):115.

27. Gao T, Soldatov R, Sarkar H, Kurkiewicz A, Biederstedt E, Loh PR, et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. Nat Biotechnol. 2023 Mar 26;41(3):417–26.

28. Campbell KR, Steif A, Laks E, Zahn H, Lai D, McPherson A, et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. 2019 Dec 12;20(1):54.

29. You L, Su PR, Betjes M, Rad RG, Chou TC, Beerens C, et al. Linking the genotypes and phenotypes of cancer cells in heterogenous populations via real-time optical tagging and image analysis. Nat Biomed Eng. 2022 Mar 17;6(5):667–75.

30. Peterman N, Levine E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. BMC Genomics. 2016 Dec 9;17(1):206.

31. https://www.scdiscoveries.com/ [Internet]. 2024. Single Cell Discovery.

32. Morgan RD, Camp RR, Wilson GG, Xu S yong. Molecular cloning and expression of NlaIII restriction-modification system in E. coli. Gene. 1996 Jan;183(1–2):215–8.

33. https://www.singlecellcore.eu/ [Internet]. 2024. Single-Cell Core.

34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

35. Genome Reference Consortium. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/. 2022. Human Genome Assembly GRCh38. Patch 14. .

36. Buys de Barbanson, Marloes, Vivek Bhardwaj, Jake Yeung, J-PTRson, Anna Alemany. BuysDB/SingleCellMultiOmics. Zenodo; 2023.

37. Andrews S. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/;

38. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

39. BroadInstititute. Picard tools. https://broadinstitute.github.io/picard/;

40. Dou J, Tan Y, Kock KH, Wang J, Cheng X, Tan LM, et al. Single-nucleotide variant calling in single-cell sequencing data with Monopogen. Nat Biotechnol. 2023 Aug 17;

41. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987–93.

42. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

43. Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat Biotechnol. 2023 May 10;41(5):604–6.

44. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun;184(13):3573-3587.e29.

45. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016 Dec 17;17(1):29.

46. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019 Dec 23;20(1):296.

47. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47–e47.

48. Chen H, Ryu J, Vinyard ME, Lerer A, Pinello L. SIMBA: single-cell embedding along with features. Nat Methods. 2023 May 29;

49. Matsuno Y, Kusumoto-Matsuo R, Manaka Y, Asai H, Yoshioka K ichi. Echoed induction of nucleotide variants and chromosomal structural

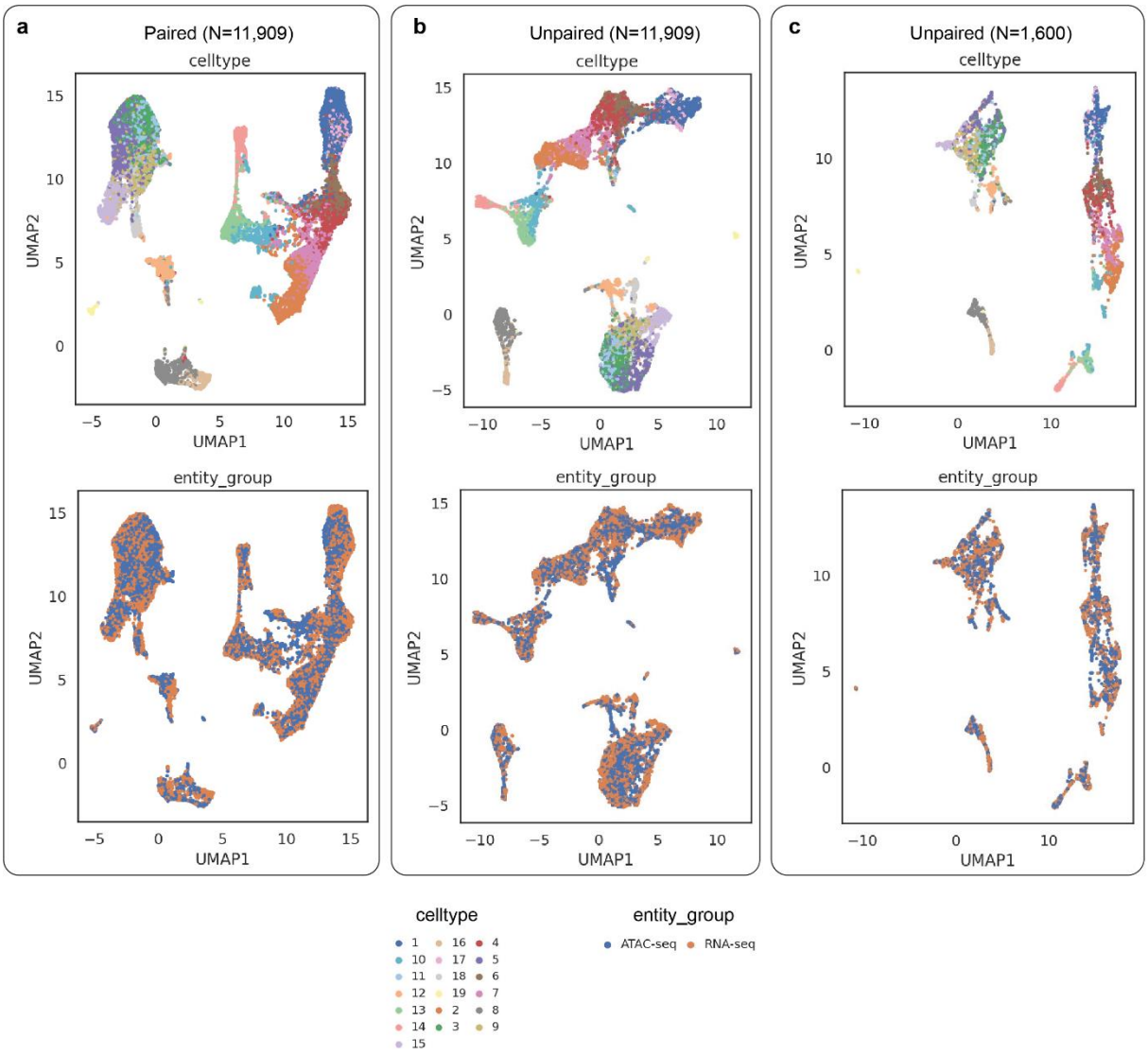
variants in cancer cells. Sci Rep. 2022 Dec 5;12(1):20964.

50. Thompson SL, Bakhoum SF, Compton DA. Mechanisms of Chromosomal Instability. Current Biology. 2010 Mar;20(6):R285–95.

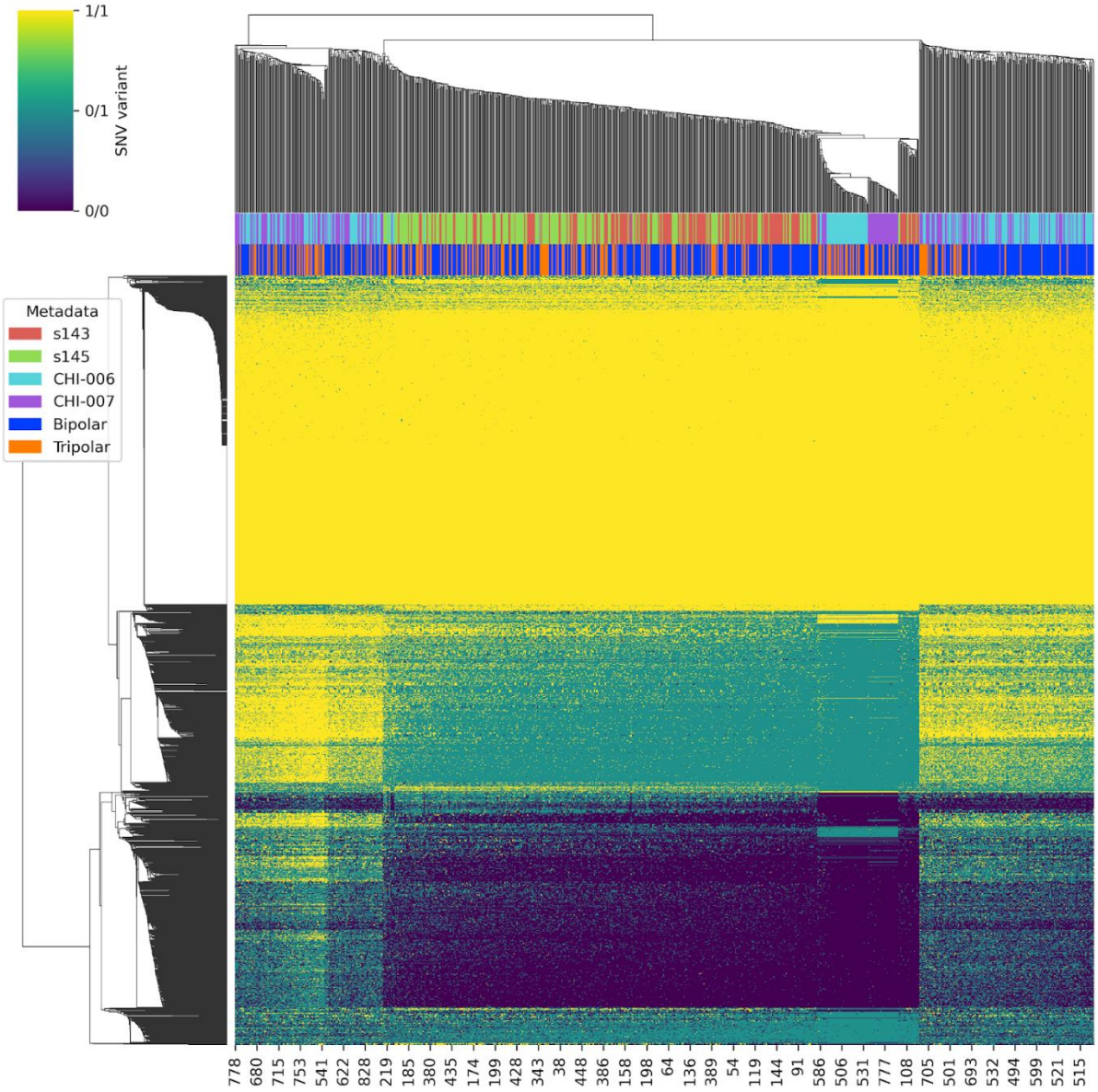
51. Khodadadian A, Darzi S, Haghi-Daredeh S, sadat Eshaghi F, Babakhanzadeh E, Mirabutalebi SH, et al. Genomics and Transcriptomics: The Powerful Technologies in Precision Medicine . Int J Gen Med. 2020 Sep;Volume 13:627–40.

52. Macaulay IC, Teng MJ, Haerty W, Kumar P, Ponting CP, Voet T. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nat Protoc. 2016 Nov 29;11(11):2081–103.

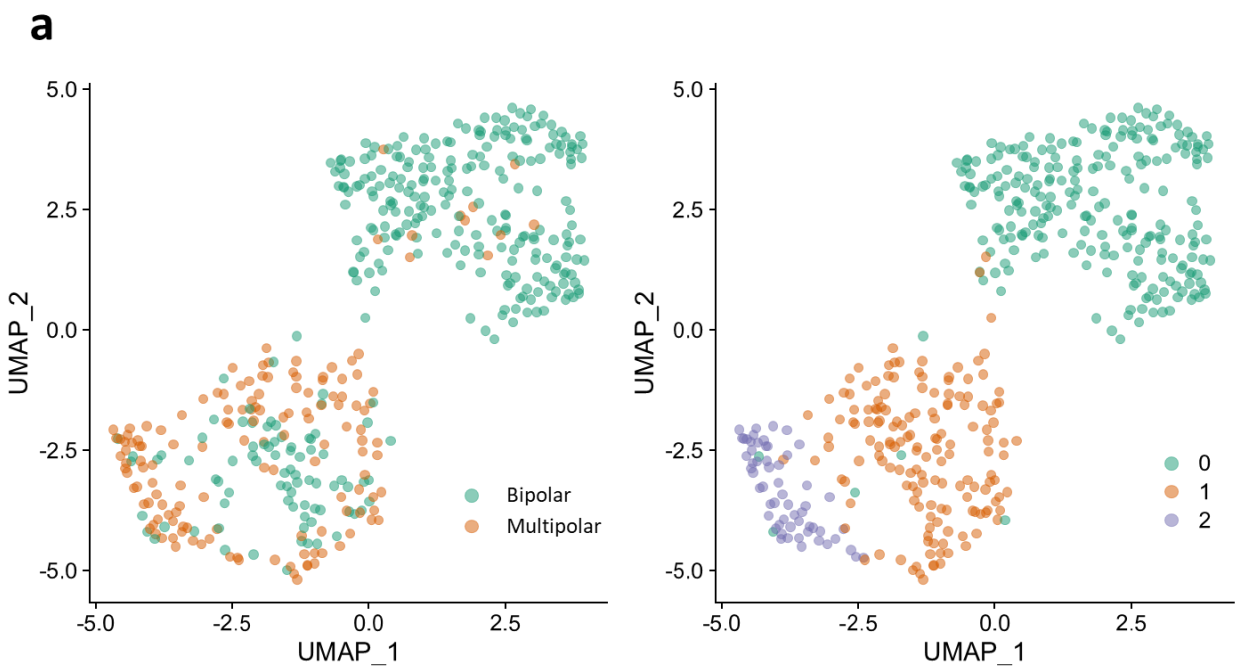
# SUPPLEMENTARY DATA



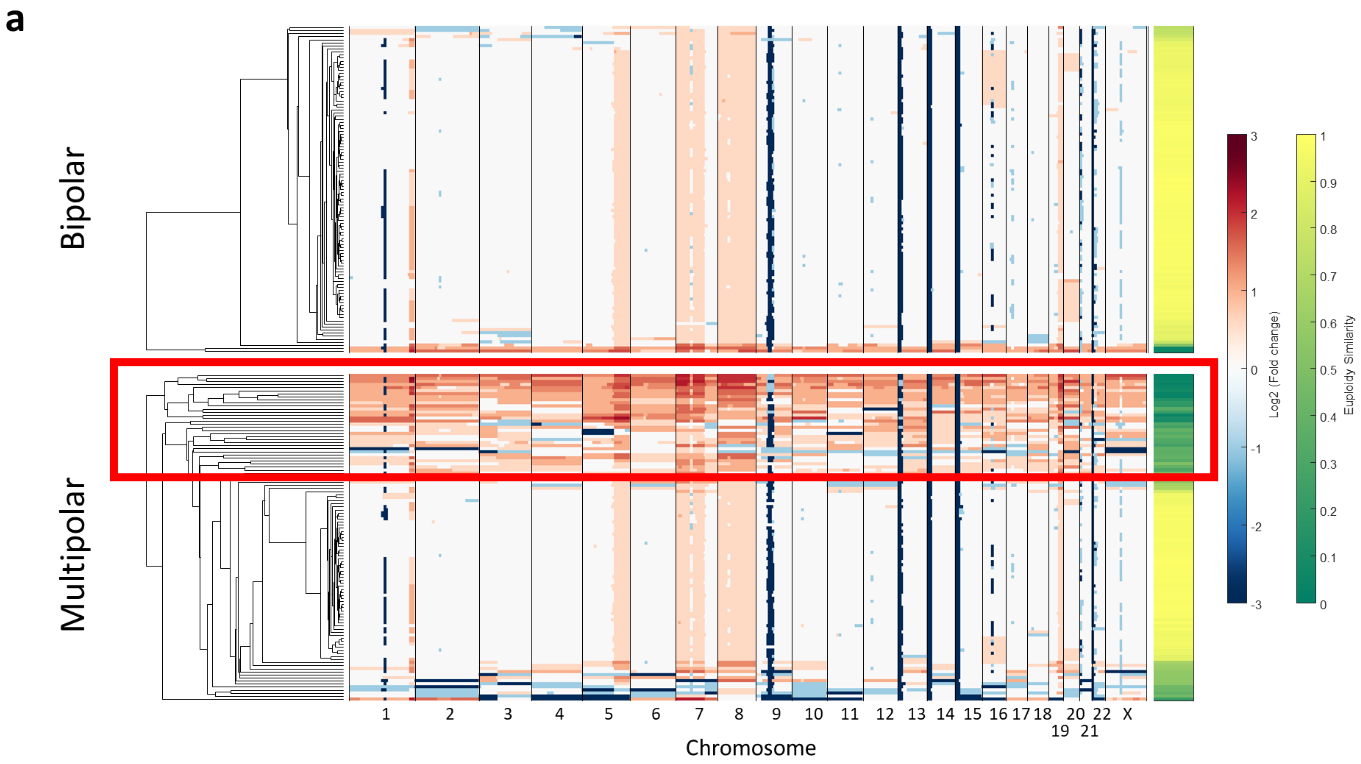

**Supplementary Figure 1. UMAP visualizations representing the integration of single-cell data by SIMBA. (a)** Shows the complete paired dataset with ATAC-seq (blue) and RNA-seq (orange) data from the same cells, demonstrating the potential for precise integration. **(b)** Displays the unpaired full dataset of 11,909 cells, indicating the algorithm's ability to handle extensive data. **(c)** Illustrates the unpaired and subsampled set of 1,600 cells, showcasing SIMBA's flexibility in working with smaller, more manageable datasets.



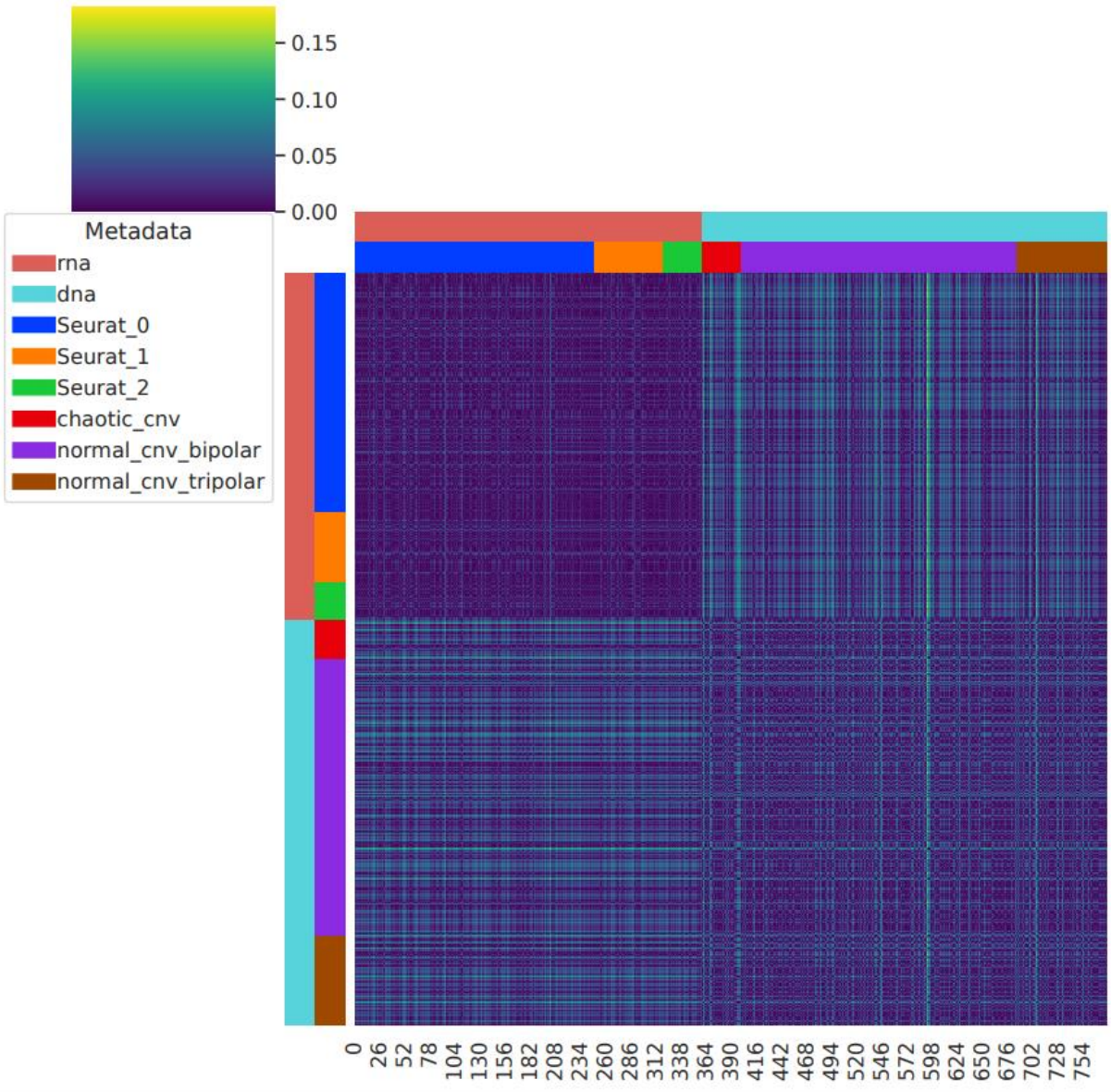

**Supplementary Figure 2. Heatmap of rejected SNVs,** where each color in the heatmap represents a genotype. The SNVs are rejected based on an inconsistent ratio between datasets, and enough total variation. On top the additional metadata of the batch and phenotype are shown.




**Supplementary Figure 3: Comparative UMAP visualizations based on differentially expressed genes (DEGs)**. On the left, the plot illustrates the distribution of DEGs in the multidimensional space, indicating potential relationships and subgroupings. The right plot shows clusters as identified by the Seurat algorithm, highlighting the distinct groupings within the data based on gene expression profiles. Each color represents a unique cluster, revealing the inherent data structure and potential biological significance behind the gene expression variations.




**Supplementary Figure 4: CNV profile Heatmap** This heatmap represents the predicted relative copy number variations (CNVs) identified by Aneufinder across different cell populations. The section highlighted in red delineates the subset of CNVs categorized as 'chaotic_cnv,' showcasing regions of the genome with notably erratic copy number changes that contrast with the more stable regions observable throughout the rest of the heatmap.




**Supplementary Figure 5: Euclidean Distance Heatmap between Omic-specific clusters.** Each row and column is a cell. Rows and columns are mirrored. The lower the distance, the closer the two cells are.

**Supplementary Table 1:** Intersection of SNVs with allele frequency (AF) greater than 0.1 between datasets. This matrix highlights the number of shared SNVs between each pair of samples, indicating the extent of common genetic variations and providing a basis for comparative analyses across different omics modalities.

| | s143 | s145 | CHI-006 | CHI-007 |
|---|---|---|---|---|
| **s143** | X | 18021 | 18160 | 11731 |
| **s145** | 18021 | X | 35608 | 23070 |
| **CHI-006** | 18160 | 35608 | X | 1211366 |
| **CHI-007** | 11731 | 23070 | 1211366 | X |

**Supplementary Table 2:** Cell distribution across omic-specific clusters. The table enumerates the final cell counts within each RNA and DNA cluster identified. RNA clusters are derived from Seurat clustering post-transcriptomic integration, whereas DNA clusters are categorized by copy number variation (CNV) patterns, using Aneufinder, into bipolar, tripolar and chaotic CNV patterns.

| Omic | Cluster Label | Cell Count |
|---|---|---|
| RNA | Seurat_0 | 247 |
| RNA | Seurat_1 | 72 |
| RNA | Seurat_2 | 40 |
| DNA | normal_cnv_bipolar | 286 |
| DNA | normal_cnv_tripolar | 93 |
| DNA | chaotic_cnv | 40 |

**Supplementary Table 3: Pairwise comparisons of distances.** Calculated using Mann-Whitney U with multiple test correction.

| *Comparison* | *p-value* | *Significant* | *Mean Distance 1* | *Mean Distance 2* | *Mean difference* |
|---|---|---|---|---|---|
| *('Seurat_1', 'Seurat_0') vs ('Seurat_1', 'Seurat_2')* | 0.814271 | FALSE | 0.014846 | 0.01527 | 0.000424 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_1', 'normal_cnv_bipolar')* | 0 | TRUE | 0.014846 | 0.032338 | 0.017492 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_1', 'normal_cnv_tripolar')* | 0 | TRUE | 0.014846 | 0.03421 | 0.019364 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_1', 'chaotic_cnv')* | 0 | TRUE | 0.014846 | 0.041218 | 0.026372 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_0', 'Seurat_2')* | 0.326955 | FALSE | 0.014846 | 0.015289 | 0.000443 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_0', 'normal_cnv_bipolar')* | 0 | TRUE | 0.014846 | 0.034638 | 0.019792 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_0', 'normal_cnv_tripolar')* | 0 | TRUE | 0.014846 | 0.035806 | 0.02096 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_0', 'chaotic_cnv')* | 0 | TRUE | 0.014846 | 0.043837 | 0.02899 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_2', 'normal_cnv_bipolar')* | 0 | TRUE | 0.014846 | 0.033296 | 0.01845 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_2', 'normal_cnv_tripolar')* | 0 | TRUE | 0.014846 | 0.035001 | 0.020155 |
| *('Seurat_1', 'Seurat_0') vs ('Seurat_2', 'chaotic_cnv')* | 0 | TRUE | 0.014846 | 0.0424 | 0.027554 |
| *('Seurat_1', 'Seurat_0') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 0 | TRUE | 0.014846 | 0.030537 | 0.015691 |
| *('Seurat_1', 'Seurat_0') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 0 | TRUE | 0.014846 | 0.026418 | 0.011572 |
| *('Seurat_1', 'Seurat_0') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 0 | TRUE | 0.014846 | 0.033117 | 0.018271 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_1', 'normal_cnv_bipolar')* | 0 | TRUE | 0.01527 | 0.032338 | 0.017068 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_1', 'normal_cnv_tripolar')* | 5.95E-256 | TRUE | 0.01527 | 0.03421 | 0.01894 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_1', 'chaotic_cnv')* | 0 | TRUE | 0.01527 | 0.041218 | 0.025948 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_0', 'Seurat_2')* | 0.738309 | FALSE | 0.01527 | 0.015289 | 1.98E-05 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_0', 'normal_cnv_bipolar')* | 0 | TRUE | 0.01527 | 0.034638 | 0.019368 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_0', 'normal_cnv_tripolar')* | 0 | TRUE | 0.01527 | 0.035806 | 0.020536 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_0', 'chaotic_cnv')* | 0 | TRUE | 0.01527 | 0.043837 | 0.028567 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_2', 'normal_cnv_bipolar')* | 0 | TRUE | 0.01527 | 0.033296 | 0.018027 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_2', 'normal_cnv_tripolar')* | 7.19E-199 | TRUE | 0.01527 | 0.035001 | 0.019732 |
| *('Seurat_1', 'Seurat_2') vs ('Seurat_2', 'chaotic_cnv')* | 1.12E-303 | TRUE | 0.01527 | 0.0424 | 0.02713 |
| *('Seurat_1', 'Seurat_2') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 0 | TRUE | 0.01527 | 0.030537 | 0.015267 |
| *('Seurat_1', 'Seurat_2') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 3.39E-192 | TRUE | 0.01527 | 0.026418 | 0.011149 |
| *('Seurat_1', 'Seurat_2') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 2.06E-261 | TRUE | 0.01527 | 0.033117 | 0.017848 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_1', 'normal_cnv_tripolar')* | 0.00404 | TRUE | 0.032338 | 0.03421 | 0.001872 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_1', 'chaotic_cnv')* | 1.62E-84 | TRUE | 0.032338 | 0.041218 | 0.00888 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_0', 'Seurat_2')* | 0 | TRUE | 0.032338 | 0.015289 | 0.017049 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_0', 'normal_cnv_bipolar')* | 5.22E-41 | TRUE | 0.032338 | 0.034638 | 0.0023 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_0', 'normal_cnv_tripolar')* | 0.102778 | FALSE | 0.032338 | 0.035806 | 0.003468 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_0', 'chaotic_cnv')* | 0 | TRUE | 0.032338 | 0.043837 | 0.011499 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_2', 'normal_cnv_bipolar')* | 0.093923 | FALSE | 0.032338 | 0.033296 | 0.000958 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_2', 'normal_cnv_tripolar')* | 0.121693 | FALSE | 0.032338 | 0.035001 | 0.002663 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('Seurat_2', 'chaotic_cnv')* | 3.03E-57 | TRUE | 0.032338 | 0.0424 | 0.010062 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 1.16E-36 | TRUE | 0.032338 | 0.030537 | 0.001801 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 2.35E-123 | TRUE | 0.032338 | 0.026418 | 0.00592 |
| *('Seurat_1', 'normal_cnv_bipolar') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 0.38298 | FALSE | 0.032338 | 0.033117 | 0.000779 |

| | | | | | |
|---|---|---|---|---|---|
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_1', 'chaotic_cnv')* | 8.16E-63 | TRUE | 0.03421 | 0.041218 | 0.007008 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_0', 'Seurat_2')* | 0 | TRUE | 0.03421 | 0.015289 | 0.01892 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_0', 'normal_cnv_bipolar')* | 2.18E-26 | TRUE | 0.03421 | 0.034638 | 0.000428 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_0', 'normal_cnv_tripolar')* | 0.000767 | TRUE | 0.03421 | 0.035806 | 0.001596 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_0', 'chaotic_cnv')* | 1.81E-189 | TRUE | 0.03421 | 0.043837 | 0.009627 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_2', 'normal_cnv_bipolar')* | 0.000193 | TRUE | 0.03421 | 0.033296 | 0.000913 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_2', 'normal_cnv_tripolar')* | 0.804658 | FALSE | 0.03421 | 0.035001 | 0.000792 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('Seurat_2', 'chaotic_cnv')* | 2.27E-46 | TRUE | 0.03421 | 0.0424 | 0.00819 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 1.84E-06 | TRUE | 0.03421 | 0.030537 | 0.003672 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 2.22E-42 | TRUE | 0.03421 | 0.026418 | 0.007791 |
| *('Seurat_1', 'normal_cnv_tripolar') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 0.016588 | TRUE | 0.03421 | 0.033117 | 0.001092 |
| *('Seurat_1', 'chaotic_cnv') vs ('Seurat_0', 'Seurat_2')* | 0 | TRUE | 0.041218 | 0.015289 | 0.025928 |
| *('Seurat_1', 'chaotic_cnv') vs ('Seurat_0', 'normal_cnv_bipolar')* | 4.63E-52 | TRUE | 0.041218 | 0.034638 | 0.00658 |
| *('Seurat_1', 'chaotic_cnv') vs ('Seurat_0', 'normal_cnv_tripolar')* | 3.48E-56 | TRUE | 0.041218 | 0.035806 | 0.005412 |
| *('Seurat_1', 'chaotic_cnv') vs ('Seurat_0', 'chaotic_cnv')* | 6.96E-07 | TRUE | 0.041218 | 0.043837 | 0.002619 |
| *('Seurat_1', 'chaotic_cnv') vs ('Seurat_2', 'normal_cnv_bipolar')* | 4.93E-66 | TRUE | 0.041218 | 0.033296 | 0.007921 |
| *('Seurat_1', 'chaotic_cnv') vs ('Seurat_2', 'normal_cnv_tripolar')* | 5.61E-45 | TRUE | 0.041218 | 0.035001 | 0.006216 |
| *('Seurat_1', 'chaotic_cnv') vs ('Seurat_2', 'chaotic_cnv')* | 0.33885 | FALSE | 0.041218 | 0.0424 | 0.001182 |
| *('Seurat_1', 'chaotic_cnv') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 1.20E-132 | TRUE | 0.041218 | 0.030537 | 0.010681 |
| *('Seurat_1', 'chaotic_cnv') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 1.97E-207 | TRUE | 0.041218 | 0.026418 | 0.014799 |
| *('Seurat_1', 'chaotic_cnv') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 4.18E-47 | TRUE | 0.041218 | 0.033117 | 0.008101 |
| *('Seurat_0', 'Seurat_2') vs ('Seurat_0', 'normal_cnv_bipolar')* | 0 | TRUE | 0.015289 | 0.034638 | 0.019348 |
| *('Seurat_0', 'Seurat_2') vs ('Seurat_0', 'normal_cnv_tripolar')* | 0 | TRUE | 0.015289 | 0.035806 | 0.020517 |
| *('Seurat_0', 'Seurat_2') vs ('Seurat_0', 'chaotic_cnv')* | 0 | TRUE | 0.015289 | 0.043837 | 0.028547 |
| *('Seurat_0', 'Seurat_2') vs ('Seurat_2', 'normal_cnv_bipolar')* | 0 | TRUE | 0.015289 | 0.033296 | 0.018007 |
| *('Seurat_0', 'Seurat_2') vs ('Seurat_2', 'normal_cnv_tripolar')* | 0 | TRUE | 0.015289 | 0.035001 | 0.019712 |
| *('Seurat_0', 'Seurat_2') vs ('Seurat_2', 'chaotic_cnv')* | 0 | TRUE | 0.015289 | 0.0424 | 0.02711 |
| *('Seurat_0', 'Seurat_2') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 0 | TRUE | 0.015289 | 0.030537 | 0.015248 |
| *('Seurat_0', 'Seurat_2') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 0 | TRUE | 0.015289 | 0.026418 | 0.011129 |
| *('Seurat_0', 'Seurat_2') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 0 | TRUE | 0.015289 | 0.033117 | 0.017828 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('Seurat_0', 'normal_cnv_tripolar')* | 8.46E-24 | TRUE | 0.034638 | 0.035806 | 0.001168 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('Seurat_0', 'chaotic_cnv')* | 9.06E-298 | TRUE | 0.034638 | 0.043837 | 0.009199 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('Seurat_2', 'normal_cnv_bipolar')* | 1.19E-16 | TRUE | 0.034638 | 0.033296 | 0.001341 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('Seurat_2', 'normal_cnv_tripolar')* | 1.19E-12 | TRUE | 0.034638 | 0.035001 | 0.000364 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('Seurat_2', 'chaotic_cnv')* | 6.22E-36 | TRUE | 0.034638 | 0.0424 | 0.007762 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 6.53E-206 | TRUE | 0.034638 | 0.030537 | 0.004101 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 0 | TRUE | 0.034638 | 0.026418 | 0.008219 |
| *('Seurat_0', 'normal_cnv_bipolar') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 3.47E-07 | TRUE | 0.034638 | 0.033117 | 0.00152 |
| *('Seurat_0', 'normal_cnv_tripolar') vs ('Seurat_0', 'chaotic_cnv')* | 5.14E-247 | TRUE | 0.035806 | 0.043837 | 0.008031 |
| *('Seurat_0', 'normal_cnv_tripolar') vs ('Seurat_2', 'normal_cnv_bipolar')* | 0.864337 | FALSE | 0.035806 | 0.033296 | 0.00251 |
| *('Seurat_0', 'normal_cnv_tripolar') vs ('Seurat_2', 'normal_cnv_tripolar')* | 0.029116 | TRUE | 0.035806 | 0.035001 | 0.000805 |

| | | | | | |
|---|---|---|---|---|---|
| *('Seurat_0', 'normal_cnv_tripolar') vs ('Seurat_2', 'chaotic_cnv')* | 4.24E-39 | TRUE | 0.035806 | 0.0424 | 0.006594 |
| *('Seurat_0', 'normal_cnv_tripolar') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 2.15E-39 | TRUE | 0.035806 | 0.030537 | 0.005269 |
| *('Seurat_0', 'normal_cnv_tripolar') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 2.77E-115 | TRUE | 0.035806 | 0.026418 | 0.009388 |
| *('Seurat_0', 'normal_cnv_tripolar') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 0.804658 | FALSE | 0.035806 | 0.033117 | 0.002689 |
| *('Seurat_0', 'chaotic_cnv') vs ('Seurat_2', 'normal_cnv_bipolar')* | 4.67E-247 | TRUE | 0.043837 | 0.033296 | 0.01054 |
| *('Seurat_0', 'chaotic_cnv') vs ('Seurat_2', 'normal_cnv_tripolar')* | 2.00E-115 | TRUE | 0.043837 | 0.035001 | 0.008835 |
| *('Seurat_0', 'chaotic_cnv') vs ('Seurat_2', 'chaotic_cnv')* | 0.006271 | TRUE | 0.043837 | 0.0424 | 0.001437 |
| *('Seurat_0', 'chaotic_cnv') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 0 | TRUE | 0.043837 | 0.030537 | 0.013299 |
| *('Seurat_0', 'chaotic_cnv') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 0 | TRUE | 0.043837 | 0.026418 | 0.017418 |
| *('Seurat_0', 'chaotic_cnv') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 1.27E-126 | TRUE | 0.043837 | 0.033117 | 0.010719 |
| *('Seurat_2', 'normal_cnv_bipolar') vs ('Seurat_2', 'normal_cnv_tripolar')* | 0.018858 | TRUE | 0.033296 | 0.035001 | 0.001705 |
| *('Seurat_2', 'normal_cnv_bipolar') vs ('Seurat_2', 'chaotic_cnv')* | 5.27E-47 | TRUE | 0.033296 | 0.0424 | 0.009103 |
| *('Seurat_2', 'normal_cnv_bipolar') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 2.76E-33 | TRUE | 0.033296 | 0.030537 | 0.002759 |
| *('Seurat_2', 'normal_cnv_bipolar') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 1.57E-106 | TRUE | 0.033296 | 0.026418 | 0.006878 |
| *('Seurat_2', 'normal_cnv_bipolar') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 0.814271 | FALSE | 0.033296 | 0.033117 | 0.000179 |
| *('Seurat_2', 'normal_cnv_tripolar') vs ('Seurat_2', 'chaotic_cnv')* | 4.76E-36 | TRUE | 0.035001 | 0.0424 | 0.007398 |
| *('Seurat_2', 'normal_cnv_tripolar') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 2.72E-05 | TRUE | 0.035001 | 0.030537 | 0.004464 |
| *('Seurat_2', 'normal_cnv_tripolar') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 3.07E-30 | TRUE | 0.035001 | 0.026418 | 0.008583 |
| *('Seurat_2', 'normal_cnv_tripolar') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 0.111946 | FALSE | 0.035001 | 0.033117 | 0.001884 |
| *('Seurat_2', 'chaotic_cnv') vs ('normal_cnv_bipolar', 'normal_cnv_tripolar')* | 5.49E-86 | TRUE | 0.0424 | 0.030537 | 0.011863 |
| *('Seurat_2', 'chaotic_cnv') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 1.39E-136 | TRUE | 0.0424 | 0.026418 | 0.015981 |
| *('Seurat_2', 'chaotic_cnv') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 6.77E-38 | TRUE | 0.0424 | 0.033117 | 0.009283 |
| *('normal_cnv_bipolar', 'normal_cnv_tripolar') vs ('normal_cnv_bipolar', 'chaotic_cnv')* | 1.29E-41 | TRUE | 0.030537 | 0.026418 | 0.004119 |
| *('normal_cnv_bipolar', 'normal_cnv_tripolar') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 3.25E-13 | TRUE | 0.030537 | 0.033117 | 0.00258 |
| *('normal_cnv_bipolar', 'chaotic_cnv') vs ('normal_cnv_tripolar', 'chaotic_cnv')* | 9.47E-51 | TRUE | 0.026418 | 0.033117 | 0.006699 |