A Novel Statistical Framework for Characterizing Mosaic Altered Cells in Single-Cell RNA Data

Author: Amith Saligrama

School: Commonwealth School, Boston MA

MIT PRIMES Project

Mentors: Dr. Giulio Genovese and Prof. Steve McCarroll

McCarroll Lab

Broad Institute of Harvard and MIT, Cambridge, MA

A Novel Statistical Framework for Characterizing Mosaic Altered Cells in Single-Cell RNA Data

Amith Saligrama

Abstract

We introduce a novel statistical framework, to analyze single-cell gene-expression counts in samples with autosomal alterations. Unlike the loss of the Y chromosome—easily detected due to gene de-activation and explored in prior works—identifying cells with autosomal alterations is fundamentally challenging. This complexity arises because, expression for autosomal chromosomes undergoing loss or alteration exhibits significant variability, rendering detection purely based on absolute counts unreliable. Our key insight for detecting chromosomal loss in a cell is based on the idea of normalizing against another chromosome, whose expression is known to be statistically independent of target chromosomal loss/mutation. This leads us to a precise characterization in terms of binomial distributions, and we can perform a hypothesis test for each cell and detect ploidy. We extend this framework for detection of cells with allelic alterations. We then develop a classification algorithm that detects chromosomal loss under control on false positivity rate (FPR). We validate our model by utilizing counts of single RNA molecules from haplotypes affected in a fraction of the cells analyses, and then use the algorithm to identify cells that have lost chromosome 18 in brain cells or carry a 9q CN-LOH alteration in chromosome 9q in induced pluripotent stem cells derived from peripheral blood mononuclear cells. Cell-by-cell identification of chromosomal loss is a critical step for inferring gene expressivity, and we identify a consistent pattern of abnormal trans-chromosomal expression in cells with autosomal loss/alterations. Our study also leads to a rather surprising finding: prior studies associate 9q CN-LOH with diverse detrimental effects, and in contrast our study reveals that the mutated cells behave no differently from non-mutated cells.

Keywords: Drop-seq, RNA sequencing, CN-LOH, Loss of Chromosome, Statistical Model, Trans-chromosomal expression.

Contents

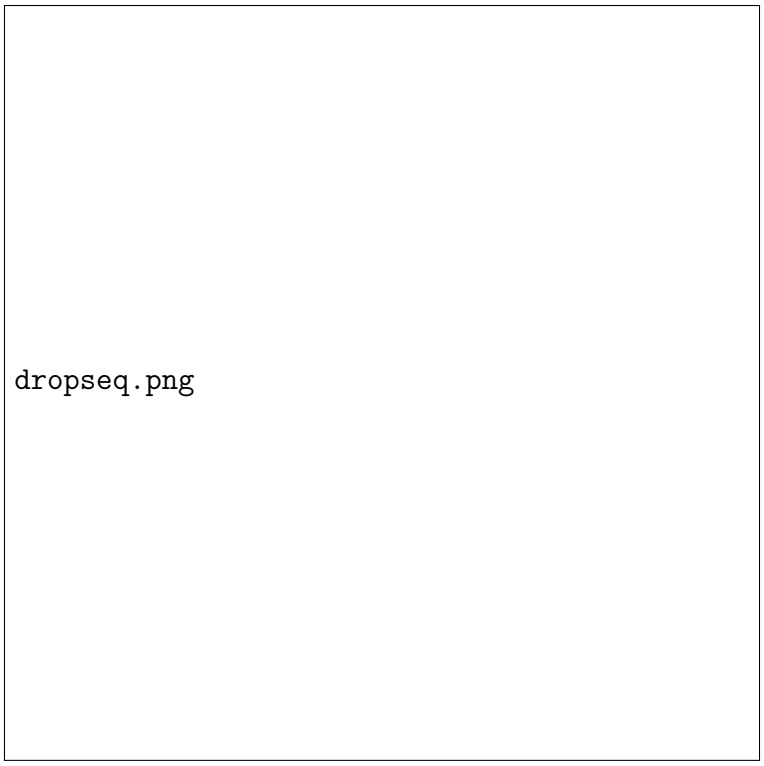

Figure 1: Single cells encapsulate individual cells in nanoliter-sized droplets with beads carrying unique cell barcodes and Unique Molecular Identifiers (UMIs). Inside each droplet, the cell is lysed, and its mRNA binds to the bead, where reverse transcription incorporates the cell barcode and UMI into the cDNA. After breaking the droplets and amplifying the cDNA, a sequencing library is prepared and sequenced. The resulting data is processed by demultiplexing reads based on cell barcodes, assigning them to individual cells. Within each cell's group of reads, UMIs are used to identify and count unique mRNA molecules, with duplicates of the same original mRNA molecule counted as one.

1 Introduction

Our proposed approach, based on Drop-Seq [7] and its extensions [23] (Fig. 1), enables precise quantification of gene expression in individual cells, forming an expression matrix with rows as cells and columns as genes. This high-resolution method has revealed cellular heterogeneity, identified rare cell types, and characterized individual cell states [6]. It has offered insights into developmental biology [16], enhanced understanding of diseases like cancer [11], and paved the way for precision medicine [12]. Additionally, it has led to the discovery of new cell types [8] and expanded our knowledge of cellular interactions and responses [10]. Overall, single-cell analysis has the potential to revolutionize medicine and biology.

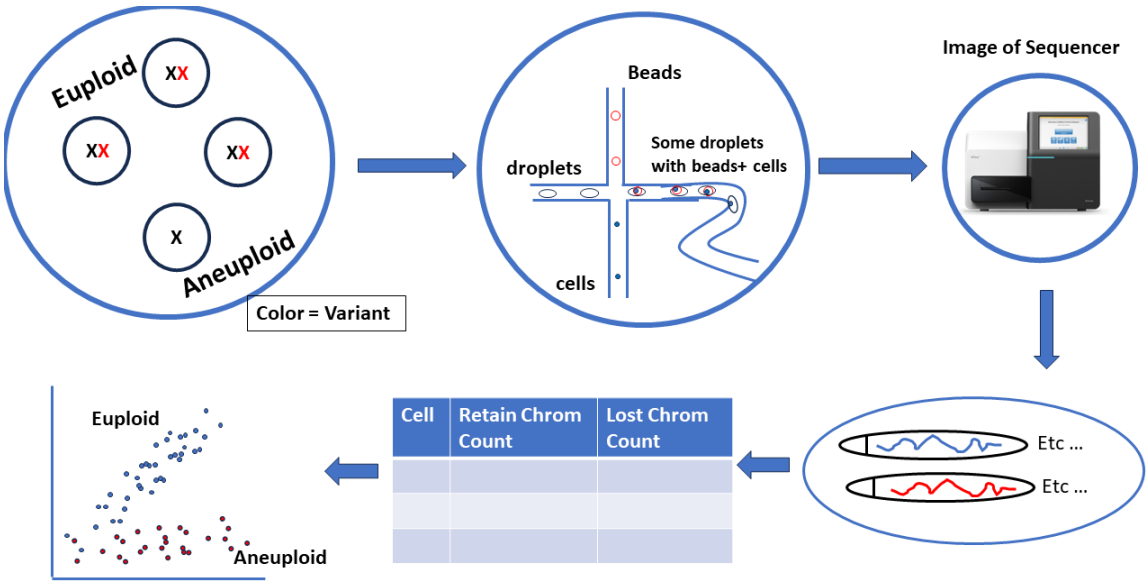

Figure 2: Overall Schematic Diagram of proposed approach. We take data from cells and read them cell by cell using dropseq. Then we use a sequencing machine to create the data for each variant. From there we extract the counts for the "Retained" and "Lost" Chromosomes. We are able to determine which one is which because we can figure out which variant is present more in the cell than the other. That variant is the "Retained" chromosome while the other is "Lost". Finally we can use the logarithmic likelihood model (outlined later) to figure out which cells are euploid and which are aneuploid (mutated). This process leads us to isolating cells with autosomal mutations. We then employ statistical tests to identify significantly abnormal trans-chromosomal expressivity.

Context for Proposed Project. Recent works [17, 14, 25] have explored large-scale analysis of single-cell and single-nuclei RNA to understand mosaic loss of Y (mLOY) chromosomes. [25] presents evidence of mLOY in the microglia and highlights its potential roles in aging and the pathogenesis of neurodegenerative disorders. Further studies found that clonality in blood ascertained from mosaic chromosomal alterations, including mLOY, is strongly associated with aging [14, 18, 25].

In contrast, similar analysis for autosomes is either missing or has not been systematically investigated. Our goal is to develop single-cell analysis methods on populations with mosaic loss in autosomes. Our objectives are twofold:

- Identify mutated cells, namely, those cells in the sample population with chromosomal loss or alteration
- Within the identified mutated cells, identify genes among the trans-chromosomal set

exhibiting consistent abnormal patterns.

We will study datasets with loss of Chromosome 18 (LO18) in brain cells and copy neutral loss of heterozygosity (CN-LOH) in induced pluripotent stem cells (iPSCs), exploring the molecular basis of autosomal alterations and their biological associations. Mosaic chromosomal alterations in brain cells increase with age, as approximately 3% of our 2,500 brain nuclei dataset have LO18. 9q CN-LOH is observed in embryonic stem cells [24] and observed as mosaic in around 5% of iPSCs analyzed by the California Stem Cell Agency (CIRM).

Challenge. Fundamentally methods developed for mLOY such as in [25] cannot be extended to autosomal mutations. This is because mLOY is readily detectable. Indeed if a cell loses the Y chromosome, the counts associated with Y chromosome is essentially zero. In contrast, the situation with autosomal loss or alterations is inherently stochastic. Counts due to loss or alterations is no longer zero since one of the chromosomal copies is retained. Additionally, counts exhibit high variability even among normal cells and as such we can no longer detect alterations purely based on counts corresponding to a chromosome.

Proposed Approach. The schematics of our approach is depicted in Fig. 2, where we utilize single-cell sequencing, and construct test statistics to identify mutated cells at a desired significance level. We present a novel statistical framework by modeling expression counts by means of a binomial likelihood model. This leads us to statistical tests and classification algorithms for cell-by-cell detection of chromosomal alterations under desired false positive rate constraints. We validate our model by utilizing counts of single RNA molecules from haplotypes affected in a fraction of the cells analyses, and then use the algorithm to identify cells that have lost Chromosome 18 in brain cells or carry a 9q CN-LOH alteration in chromosome 9q in iPSCs derived from peripheral blood mononuclear cells. We then test for statistical significance of trans-chromosomal gene expression for the detected mutated cells and study its impact on biological processes.

1.1 Research Contributions

The primary research contributions are:

1. The creation of a novel classifier for identification of mutated and non-mutated cells and validated on brain cells and iPSCs.

2. Method for identification of trans-chromosomal genes that have significance.

3. Observation that specific genes are consistently abnormally expressed across samples of CN-LOH. This suggests severe impact on biological processes such as tumor growth.

2 Methods

In this section we describe methods to analyze single cell RNA count data for two datasets: LO18 on brain cells and 9q CN-LOH for iPSCs. McCarroll lab has access to 2,500 samples of brain tissue among which 3% are known to suffer from LO18. We use SCPred [20] to further classify cells into seven cell types: Astrocyte, Gabaergic, Glutamatergic, Polydendrocyte, Oligodendrocyte, Endothelia, and Microglia, as these are the prevalent cell types observed in brain tissue. In addition, we have data with CN-LOH and we see a significant loss in chromosome 9q in induced pluripotent stem cells (iPSCs). 9q CN-LOH is observed in embryonic stem cells and this is observed as mosaic in around 5% of iPSCs analyzed [24].

2.1 Likelihood Model for Absolute Counts with Chromosome loss

We have one dataset with Chromosomal loss, specifically brain cells with LO18. While we will describe our method in the context of Chromosome 18 loss, it is general and applicable to other cell types or situations with different chromosome loss. In our setting, cells with LO18 are expected to have half the RNA molecules from the aneuploid chromosome. However, detecting ploidy from absolute count data is challenging due to significant variability in both euploid and aneuploid cells, leading to potentially similar counts.

For this reason, we propose to baseline the counts against the counts of another chromosome, which we a priori know to have *suffered no loss*. For each cell barcode we count the number of unique RNA molecules whose sequencing reads align against Chromosome 4 and those that align against Chromosome 18. For our dataset we arbitrarily selected Chromosome 4 as a control chromosome, since we do not expect Chromosome 4 to be found in an aneuploid state in the brain nuclei.

We expect a predefined ratio between the amount of RNA coming from each cell for Chromosome 4 and Chromosome 18 with the ratio for euploid cells being twice the ratio for aneuploid cells. Using this insight we consider the test-statistic, $t(n_4, n_{18}) = \frac{n_4}{n_4+n_{18}}$, for each cell of a cell type, where n_4 is the number of counts for Chromosome 4 and n_{18} is the number of counts for Chromosome 18. This statistic has the nice feature that it is normalized to one, and the ratio $\frac{t(n_4,n_{18})}{1-t(n_4,n_{18})} = 0.5\frac{t(n_4,n_{18}/2)}{1-t(n_4,n_{18}/2)}$. Thus if we have two cells in

euploid state with identical counts, and one of these undergoes LO18, we will expect this statistic to reveal ploidy. However, since no two cells are identical and cells exhibit significant variability, we observe variability in the test statistics as well. Using the insight from this statistic, we develop a novel statistical model based on binomial distribution to account for the variability.

For each cell type, we cluster the list of test statistics for cells of that cell-type in this ratio space, using one of the standard clustering methods [9]. As such we expect to see a nice separation into two clear clusters due to the property of the statistic just described. The median of the cluster corresponding to the larger ratio we denote by p_a (associated with aneuploids) and the median of the cluster with smaller ratios as p_e (associated with euploids). After estimating these two ratios by clustering in the ratio space, we use a binomial likelihood model to estimate likelihoods for the observed counts given according to each model. We then build a classifier based on the ratio of these two likelihoods to infer ploidy.

$$\text{Likelihood}_e = \binom{n_4 + n_{18}}{n_{18}} \cdot p_e^{n_4} \cdot (1 - p_e)^{n_{18}} \tag{1}$$

$$\text{Likelihood}_a = \binom{n_4 + n_{18}}{n_{18}} \cdot p_a^{n_4} \cdot (1 - p_a)^{n_{18}} \tag{2}$$

This is the binomial probability of observing n_4 reads on Chromosome 4 conditional on the total number of reads observed on Chromosome 4 and 18 being $n_{18} + n_4$.

Now, let $\text{Ratio} = \frac{\text{Likelihood}_a}{\text{Likelihood}_e} = \frac{p_a^{n_4} \cdot (1-p_a)^{n_{18}}}{p_e^{n_4} \cdot (1-p_e)^{n_{18}}}$. We then worked off of $\text{Score} = \log_{10}(\text{Ratio})$ to detect the loss if the Score ≥ 1, else if Score ≤ -1 declare as euploid, and otherwise detect the cell as uncertain. The intuition here is that by normalizing with Chromosome 4, the variability is suppressed, while simultaneously allowing for ploidy to be detectable.

2.2 Likelihood Model for Allele-Specific Counts

First, we describe the model in the presence of chromosomal loss and then extend the model in the presence of chromosomal alterations.

Chromosomal Loss. Allele-specific count data in single-cell RNA sequencing (scRNA-seq) in the presence of chromosome loss can be insightful. In a normal euploid cell, genes have two alleles, one from each parent. When there is a loss of a chromosome, one of the alleles may be lost, leading to a monoallelic expression pattern. Allele-Specific counts record the gene expression of each variant of a chromosome in every cell. An analysis on all cells of a cell

type detects which variant appears more than the other. The variant that appears the most we call the *retained* allele and the *loss* allele.

The binomial model for this setting is somewhat simpler than what we needed for absolute count data. Intuitively, suppose a chromosome has two alleles, and if the chromosome is in a euploid state, a random draw would yield either of the two alleles with equal probability. On the other hand, for a aneuploid state, a random draw is essentially deterministic yielding the allele that is preserved. With this in mind, we set the euploid binomial parameter so: $p_e = 0.5$ and the aneuploid parameter so: $p_a = 0.95$. This is because for aneuploid cells, while we expect all RNA molecules to originate from the retained chromosome, due to the inevitability of allele determination errors and ambient RNA contamination, we allow 5% of the RNA molecules count to be from the lost chromosome.

Chromosomal Alterations. In this case while chromosomes are not lost, the alleles lost are replaced predominantly by the alleles that are retained. As a result we can use a similar argument and justify a similar binomial model. More precisely, we set the binomial parameter for the mutated cell to be close to one, and set $p_a \approx 0.95$ since we expect one of the alleles uniformly across both copies. Since in the non-mutated cell both alleles are equally likely we set the non-mutated parameter to be $p_e = 0.5$. Note that we overload notation and use the same symbols as euploid and aneuploid for notational convenience.

The scores with allele-specific counts are computed in a similar fashion - the score is the log-likelihood of the ratio of the probability of mutation to the probability of normality.

2.3 Classification based on the two Likelihoods

We compute likelihoods for each of the two biological models, mutated and non-mutated, corresponding to the two observations, absolute and allele-specific counts. Then we compute scores associated with the two counts. The scores for absolute counts and allele-specific counts are the log-likelihood ratios of the probabilities of observing the absolute count and allele-specific counts respectively under the two biological models, mutated and non-mutated. If Score for the absolute count (resp allele-specific count) ≥ 1, then we labeled the cell as aneuploid; if Score ≤ -1, euploid. Otherwise, the ploidy is uncertain. The total score of a cell is the sum of log-likelihood ratios. We doubled the threshold for detecting aneuploid and euploid, namely, if the sum of the scores is greater than 2, it is euploid, less than -2, aneuploid, and otherwise uncertain. Our intuition is that by doing so we can increase confidence when we declare a cell's ploidy.

2.4 Expression analysis

We also want to understand how gene expression changes with ploidy, so we can use differential gene expression matrices which record every gene's expression in every cell. We can aggregate the columns by their cell type and ploidy, and then create a 2×2 table like below:

	Expression Count mutated cells of a specific cell type	**Expression in non-mutated cells of the same cell type**
Gene A	a	b
Excluding Gene A	c	d

where a, b, c, d are integers. Then, to calculate the expression change between mutated and non-mutated states for each cell type, we can compare the ratios $\frac{a}{b}$ and $\frac{c}{d}$ to get $\frac{a}{b}/\frac{c}{d} = \frac{ad}{bc}$. This ratio is approximately one under null-hypothesis that there is no significant expression change compared to all the other genes. For each gene we use the Mann-Whitney U Test [2] to determine whether gene-expression is statistically significant in the mutated state for all cell types.

2.5 Doublet Detection

Sometimes, the counts have "doublets," or reads that have two cells in the same droplet. It is important to detect these events since the counts corresponding to doublets can mislead us into classifying them as either aneuploid or euploid. We can reduce the problem of doublet detection into a binomial testing problem. Specifically, let (x_i, y_i) be the retain and loss coordinates for a cell, i. Note that we always have $x_i \geq y_i$. Under the hypothesis that we do not have a doublet, we have two possibilities for the cell: (a) the cell is a euploid: to test for significance that the cell is not a euploid, we can compute the p-value with the binomial distribution with parameter 0.5, which yields, $p_i = \Sigma_{k=x_i}^{x_i+y_i} \binom{x_i+y_i}{k} \cdot 0.5^k \cdot (1-0.5)^{x_i+y_i-k}$; (b) the cell is a aneuploid: to test the significance that cell is not a aneuploid, we compute the p-value for the binomial with parameter 0.95, which yields: $q_i = \Sigma_{k=0}^{x_i} \binom{x_i+y_i}{k} \cdot 0.95^k \cdot (1-0.95)^{x_i+y_i-k}$. If both p_i and q_i are less than 0.01, we label that cell as a doublet.

3 Results

In this section, we will describe our data preparation pipeline and the dataset used for our studies. Then we will apply our proposed statistical framework (Section 2) for two studies,

the first on LO18 in brain cells, and the second on 9q CN-LOH in stem cells. Subsequently, we will use the results of these studies to deduce the role of LO18 and 9q CN-LOH on trans-chromosomal gene expression.

3.1 Data Preparation Pipeline.

Using a pipeline (see Fig. 1, 2) developed for the analysis on single cell RNA expression [23], single cell RNA data is generated. Briefly, DNA barcoded beads and brain nuclei are colocalized within droplets and RNA molecules are extracted from the nuclei and barcoded with synthetic oligonucleotides localized on each bead. RNA is the reverse transcribed into DNA and amplified to generate a library that is then analyzed using an Illumina sequencing machine. Reads are grouped into single molecules using UMI barcodes and then grouped into nuclei using bead barcodes.

3.2 Datasets

Brain Cells for LO18 Study. We work on a single sample of a woman who had a low cell fraction, so she was very likely to have LO18. DNA microarray data from brain tissues indicates the presence of a mosaic LO18 and a germline loss of one copy of the tip of 18p, consistent with the presence of a ring Chromosome 18 known to lead to recurrent loss of the whole chromosome [5, 1, 3]. We next use SCPred [20] to further classify cells and extract seven cell types as these are prevalent cell types observed in brain tissue. For these cell types we gather the absolute and allele-specific counts. The dataset details are described on the left of Table 1 and Table 2 for the two different counts. Note that due to machine errors the two cumulative cell counts are not identical. Nevertheless the differences are small, and for a substantial number of cells we observe both counts.

Next, using the data preparation pipeline described above, we generate single cell RNA data for brain nuclei of the donor for each cell-type. Through the data preparation process explained in Sec. 3.1 we obtain over 10,000 brain nuclei. The amount of RNA molecules for each nuclei ranges from a few hundreds to a few thousands and we use these molecules to infer cell type and presence of LO18.

Stem Cell dataset for CN-LOH Study. Here we use a CN-LOH dataset, where we see a significant loss in chromosome 9q in induced pluripotent stem cells (iPSCs). 9q CN-LOH is observed in embryonic stem cells and this event is observed as mosaic in around 5% of iPSCs analyzed [24]. We used four samples of around 10,000 iPSCs to test the framework.

Cell Type	Total Cells
Astrocyte	1452
Gabaergic	1361
Glutamatergic	1620
Polydendrocyte	617
Oligodendrocyte	4138
Endothelia	279
Microglia	806
Total	10373

Euploid	Aneuploid	Uncertain
1312	56	84
1237	54	70
1483	63	74
61	505	51
460	3153	525
180	29	70
578	106	122
5011	3966	996

Table 1: Absolute Counts: Total number of cells, and counts of cells with euploid, aneuploid, and uncertain ploidy for each cell type.

Cell Type	Total Cells
Astrocyte	1480
Gabaergic	1401
Glutamatergic	1629
Polydendrocyte	607
Oligodendrocyte	4034
Endothelia	219
Microglia	784
Total	10154

Euploid	Aneuploid	Uncertain
1241	90	149
1253	36	112
1438	80	111
74	444	89
671	1606	1757
116	18	85
380	66	338
5173	2340	2641

Table 2: Allele-Specific Counts: Total number of cells, and counts of cells with euploid, aneuploid, and uncertain ploidy for each cell type.

Using the same pipeline described for brain data, we generate single cell RNA data for these iPSCs. The details of the number of cells for the different samples is presented in Table 3.

3.3 Cell-by-Cell analysis with LO18 in Brain Cells

In this section we will apply our proposed method described in Section 2 on the LO18 data tabulated in Table 1. We focused mostly on glutamatergic, polydendrocyte and oligodendrocyte cells since endothelia and microglia cells had too few counts and astrocyte and gabaergic ended up acting just like glutamatergic.

Sample	No. of Cells
1	7947
2	9330
3	9431
4	7414
Total	34122

Normal	CNLOH	Uncertain	Doublet
6974	722	68	183
6798	1804	19	709
7664	1252	21	494
6405	695	60	254
27841	4473	168	1640

Table 3: For each sample, we have different counts of cells. The table shows the distribution of normal, non-mutated cells, cells with CN-LOH (Copy-Neutral Loss of Heterozygosity), cells which we are uncertain that they have a 9q CN-LOH mutation, and doublet cells.

Cell Type	No. of Cells
Astrocyte	1223
Gabaergic	1170
Glutamatergic	1392
Polydendrocyte	466
Oligodendrocyte	1865
Endothelia	108
Microglia	373

Agreement	Disagreement
1196	27
1131	39
1352	40
453	13
1698	167
100	8
351	22

Table 4: For each cell type, the agreement and disagreement between the absolute counts and allele-specific count classifiers are shown. We see that there is significant agreement between the two classifiers. Notice that the number of cells for each cell type is different from the numbers in Table 1 and Table 2. This is because not all cells have both counts available.

3.3.1 Absolute Count Based Classification

We analyze LO18 absolute count data using the likelihood for absolute count model described in Sec. 2.1. To get an idea of the variability of the number of counts across cell types for each chromosome we report their numbers. The number of absolute counts for astrocyte: the range of chromosome 4 counts is 1-1458, while for chromosome 18 counts it is 1-580; gabaergic: the range of chromosome 4 counts is 1-2430, while for chromosome 18 counts it is 1-996; glutamatergic: the range of chromosome 4 counts is 1-5705, while for chromosome 18 counts it is 1-1740; polydendrocyte: the range of chromosome 4 counts is 1-1484, while for chromosome 18 counts it is 1-352; oligodendrocyte: the range of chromosome 4 counts is 1-

599, while for chromosome 18 counts it is 1-225; endothelia: the range of chromosome 4 counts is 1-442, while for chromosome 18 counts it is 1-197; microglia: the range of chromosome 4 counts is 1-407, while for chromosome 18 counts it is 1-149. As we remarked it is difficult to identify ploidy directly from absolute counts due to the significant variability in the count data resulting in overlaps between counts of anueploid and euploid cells. Recall that we proposed to use Chromosome 4 as a control since we know that Chromosome 4 is highly likely to be in euploid state for all cells. Clustering into two clusters for each cell type and taking their corresponding medians yields the two Binomial parameters p_a and p_e for the two biological situations, aneuploid and euploid.

Since these clusters are obtained by clustering the list of test statistics, $\frac{n_4}{n_4+n_{18}}$, where n_4, n_{18} are the number of cells of a cell type for Chromosome 4 and Chromosome 18, we expect to see that

$$\frac{p_a}{1-p_a} \approx 2\frac{p_e}{1-p_e}.$$

in our data. We applied our likelihood model of Sec. 2.1 and classified the cells into aneuploid and euploid for each cell type. The overall summary of the classified output is reported in Table 1.

Next, we investigated cells that are classified as aneuploid and euploid. To do so, we consider a 2D plot with Chromosome 4 count on the x-axis and Chromosome 18 count on the y-axis by our model (Sec. 2.1). We only plot three cell types because other cell types follow a similar trend. Our reasoning for the plot is that we should be able to see a clear separation of the ratio. Intuitively, consider two cells with nearly similar count for Chromosome 4. However, suppose one of them is aneuploid and the other euploid, we should see the ratio of Chromosome 18 to Chromosome 4 to be nearly 1/2 when comparing the two cells. From examining the plots in Figure 3, we see that there are two parts of the data, one with a trendline with nearly twice the slope as the other's trendline, just as we theorized, and thus validating our model. The scatter around the higher-ratio trend line are classified as euploid cells - those that have two copies of both chromosomes, while the lower trend line should contain the aneuploid cells - cells with only one copy of Chromosome 18. As seen in Figure 3 our classifier fully separates the aneuploid cells from the euploid as we see manifestation of two clear clusterings.

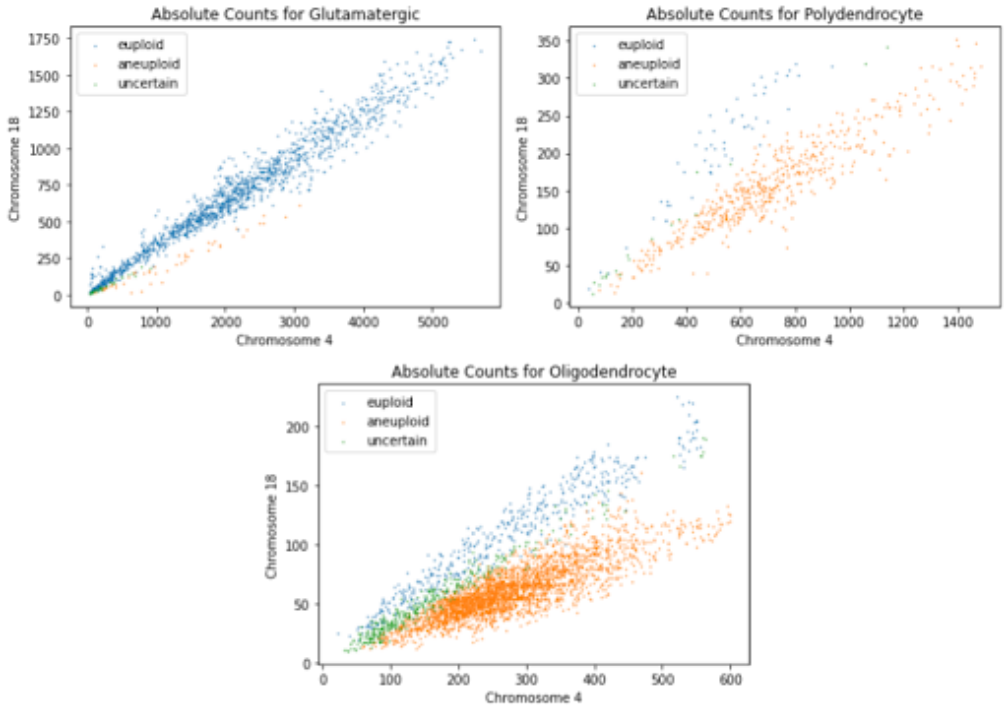

Figure 3: Absolute Counts classifier results for glutamatergic, polydendrocyte, and oligodendrocyte cells. The blue cells represent euploid cells, orange represent aneuploid, and green represent cells with an uncertain ploidy.

3.3.2 Allele-Specific Count Results

Similar to absolute counts, allele specific counts also exhibit significant variability and as such it is difficult to ascertain ploidy purely from these counts. For astrocyte, the retained chromosome count range is 0-47, loss is 0-41; gabaergic, the retained chromosome count range is 0-114, loss is 0-116; glutamatergic, the retained chromosome count range is 0-146, loss is 0-207; polydendrocyte, the retained chromosome count range is 0-59, loss is 0-28; oligodendrocyte, the retained chromosome count range is 0-68, loss is 0-56; endothelia, the retained chromosome count range is 0-16, loss is 0-25; microglia, the retained chromosome count range is 0-20, loss is 0-15.

In the case of LO18, one of the two alleles is lost and the other retained. This insight led us to propose the model in Sec. 2.2 where we set $p_e = 0.5$ for the euploid cell since both alleles are present in equal number, and so a random draw would yield either of the two alleles with equal probability. On the other hand for an aneuploid cell we set $p_a = 0.95$ since the retained allele would be drawn with very high probability. Running our model with these

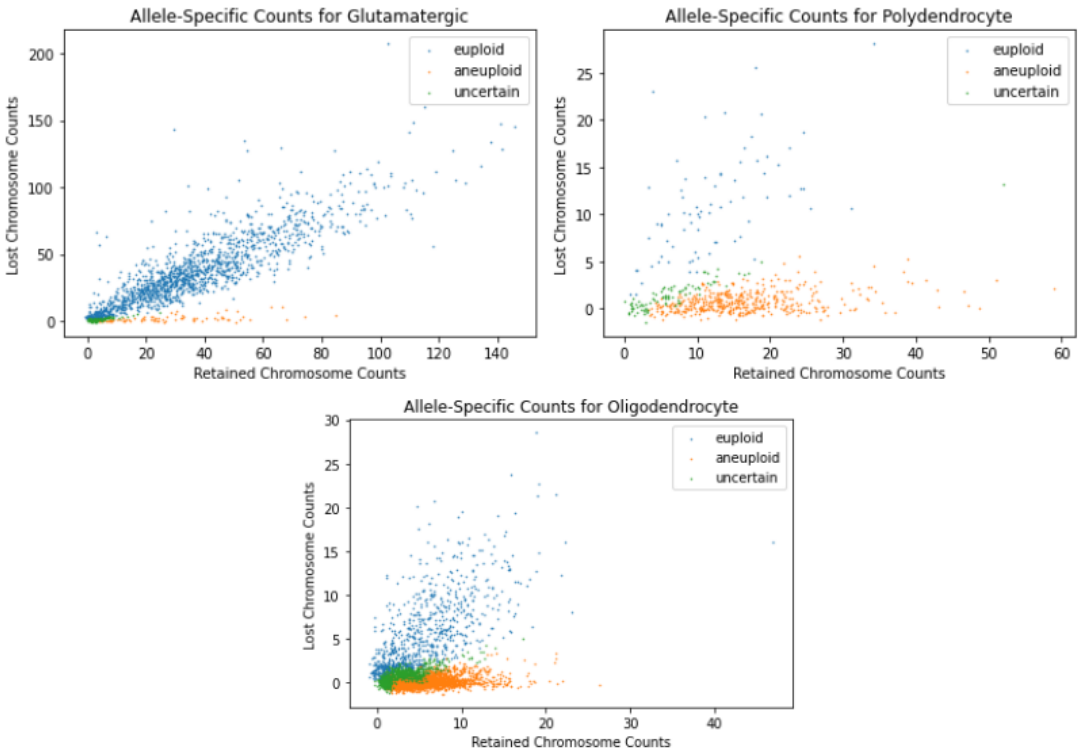

Figure 4: Allele-Specific Count classifier results for glutamatergic, polydendrocyte, and oligodendrocyte cells. The blue cells represent euploid cells, orange represent aneuploid, and green represent cells with an uncertain ploidy.

parameters and thresholding the log-likelihood ratio yields a classification of cells.

Table 2 tabulates results for this model. We see that more cells are declared uncertain here, and this is to be expected from the fact that allele-specific counts are more noisy.

We next construct a 2D plot to visualize the classified cells. The feature space has retained chromosome counts on the x-axis and the lost counts on the y-axis. To see that this makes sense, note that retain counts corresponds to the allele that is predominantly found in all the cells, and lost corresponds to the allele that are lost in aneuploid cells. As a result if a cell is in euploid state we expect equal fractions of retained and lost alleles, and a cell in aneuploid state will consist primarily of the retained allele with no lost allele. Thus the 2D visualization of the classified cells would have aneuploid cells essentially scattered horizontally, while the euploid cells are scattered along the 45° line.

As we can see from Fig. 4 our proposed allele-specific classifier also separates the aneuploid and euploid cells well, as the model has labeled all of the cells with counts only on the retained chromosome as aneuploid, while the cells that have significant counts on both the lost and

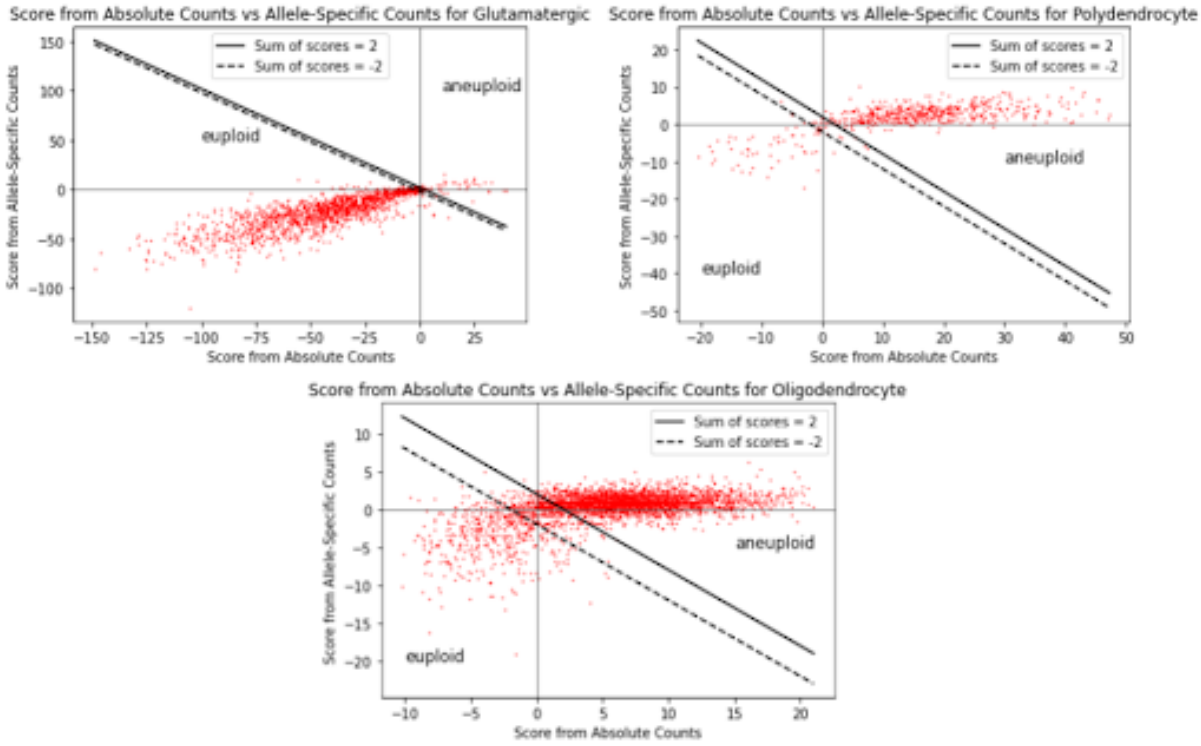

Figure 5: Scores from absolute counts vs scores from allele-specific counts across glutamatergic, polydendrocyte, and oligodendrocyte cells. The lines represent the sum of the scores being 2 and -2, so if it is less than 2 it is euploid and greater than 2 aneuploid. This graph shows that both methods are in tune with one another, as the counts are mostly in quadrants I and III - so the methods either both detect a cell euploid or aneuploid. Very rarely does one method call a cell aneuploid and the other diploid.

the retained chromosomes.

3.3.3 Combined Likelihood Results

We have two scores: one based on absolute count likelihood, and the other based on allele-specific count likelihood. These models are based on different assumptions. The absolute count is based on a test-statistic that baselines against Chromosome 4, which we assume to be in euploid state. The allele-specific count is based on a different statistic and assumes that nature chooses one of the alleles and retains it across all cells, and the other allele is predominantly lost in all the aneuploid cells. As such the absolute count based classifier tends to be less noisy (since our absolute-count model requires weaker assumptions to hold). This is seen in Tables 1, 2 where we see more cells declared as uncertain for the latter. Nevertheless, both tables demonstrate a consistent pattern across all cell types. We investigate this further

in Table 4, which depicts the number of cells that both agree on. Note that as seen from the left side of Table 4 the number of cells for which we have both counts is smaller than their individual counts, which we attribute to machine read errors. Nevertheless, there are sufficiently large number of cells for which we have both counts, and we see consistent agreement between the two classifiers. Indeed, it is surprising to see that we have only a small number of cells which are differentially classified. To visualize the table we construct a 2D plot with absolute score against the allele-specific score on the x and y axis and the results are depicted in Figure 5. Note that cells belonging to the first and third quadrant imply that the two classifiers agree on a cell being aneuploid or euploid, and it is clear that both classifiers predominantly agree.

3.4 Cell-by-Cell analysis for CN-LOH Mutations

9q CN-LOH mutations are chromosomal alterations. We have 4 samples of iPSCs as described in Sec. 3.2 and the number of cells across the different samples is reported in Table 3. In this dataset we only have allele-specific count data. Instead of euploid and aneuploid, our two states are normal and 9q CN-LOH.

Recall from our model in Sec. 2.2 we expect that with chromosomal alterations, in mutated cells one of the two alleles is essentially lost and replaced with the other allele. This is what happens with CN-LOH mutation, namely, a portion of the "lost" 9q is replaced by alleles from the "retained" 9q. Thus we will again expect to see that in cells with the 9q CN-LOH mutation, the lost counts will be close to 0, while in normal cells, the counts will be around the same. This leads to setting the binomial parameter $p_a = 0.95$ for mutated cells and $p_e = 0.5$ for normal cells. We then run our classification algorithm on the CN-LOH dataset. In this case we also encounter a number of doublets, which we remove using the method described in Sec. 2.5. The summary of our results is reported in Table 3. About 12% of the cells are CN-LOH cells across all the samples. Since our model is based on the assumption that the lost alleles are replaced by the retained alleles, a 2D plot similar to that in LO18 for allele-specific count classification makes sense. The plots here have more doublets than the Chromosome 18 data, which may skew the data. We see that many points between the critical masses of the normal and CN-LOH clusters are labeled as one or the other, but they do not conform to either model. Using the method of detecting them as detailed in the methods, we can point out the doublets in these graphs (See Figure 6. Removing the doublets yields the results depicted in Figure 7.

Now the picture is much more clear. Just as in Chromosome 18 loss, we see that there

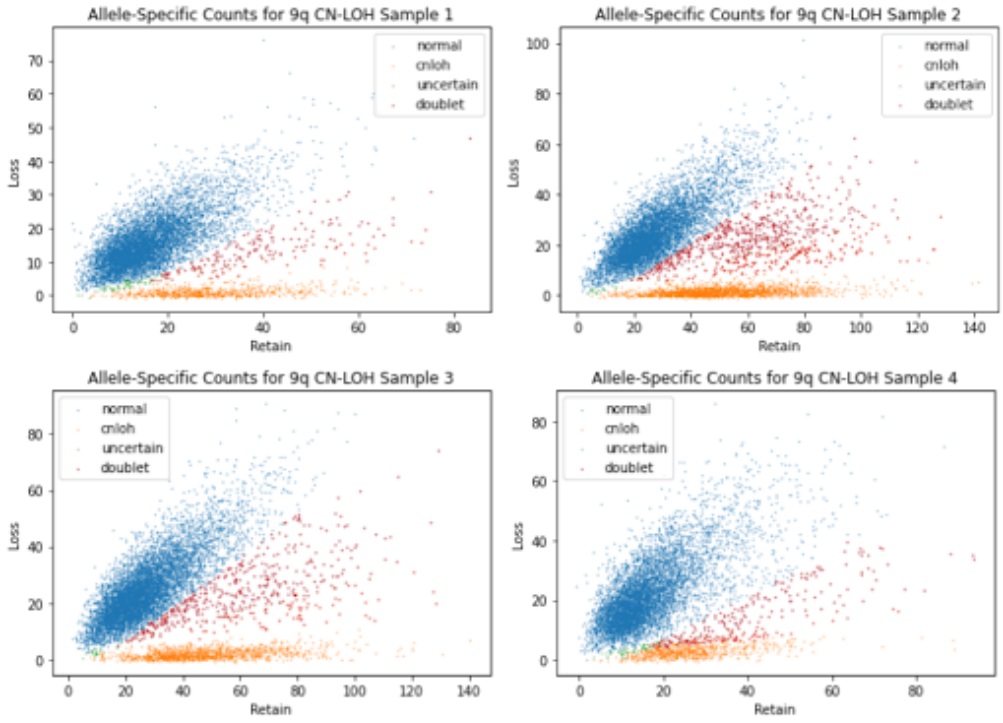

Figure 6: Allele-Specific counts for the four 9q CN-LOH samples, with doublets in red. The blue points represent euploid cells, orange points represent aneuploid, green points represent cells with an uncertain ploidy, and the red points represent doublets.

is a critical mass of points with almost no counts in the lost chromosome and another mass with roughly equal counts in both chromosomes. These two clusters are the 9q CN-LOH and normal cells, and the method correctly identifies these two clusters. This method is not just restricted to 18 loss.

3.5 Gene-Expression Analysis

Here we will apply our Gene Analysis model (Sec. 2.4) to both the LO18 and CN-LOH datasets. Our goal is to classify genes that exhibit significant differences from their normal behavior. Volcano plots [4] are particularly valuable in this context. In a volcano plot, the X-axis represents the log2 fold change in gene expression, while the Y-axis represents the negative logarithm of the p-value, reflecting statistical significance. Genes that are downregulated appear on the left-side of the plot, and in particular we should expect genes associated with the chromosomal loss to be in this category. These genes provide additional evidence that there is a potential chromosomal loss. On the right side, potentially in response to chro-

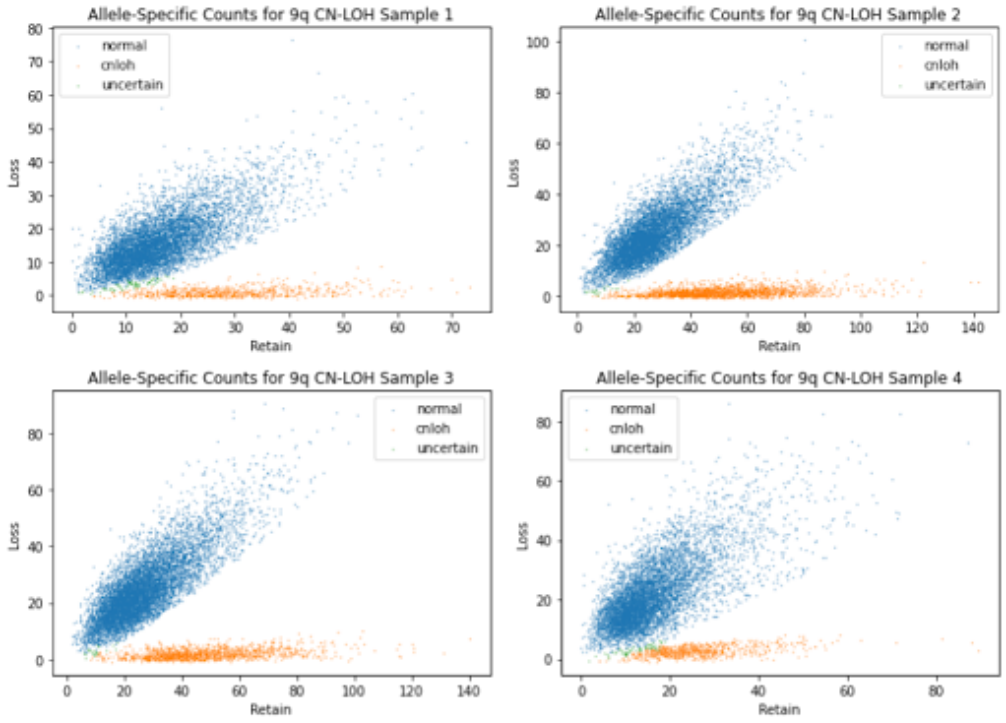

Figure 7: Allele-Specific counts for the four 9q CN-LOH samples, with doublets removed.The blue cells represent euploid cells, orange represent aneuploid, and green represent cells with an uncertain ploidy.

mosomal loss, genes on other chromosomes might be upregulated as part of compensatory mechanisms or other cellular responses. These related alterations would be reflected on the right side of the plot.

3.5.1 LO18 Analysis

With this in mind, we ran our gene analysis model Sec. 2.4 and plot the genes in Figure 8 that were in the germline deleted region and mosaically deleted region of 18 as well as those that were not in Chromosome 18:

We expect the Chromosome 18 genes' expression change to be around 0.5 as Chromosome 18 went from 2 copies to 1. This can be seen in the above plots. Interestingly, we notice that a significant alteration in expression of trans-chromosomal genes. Thus it suggests that biological processes and molecular functions changed too.

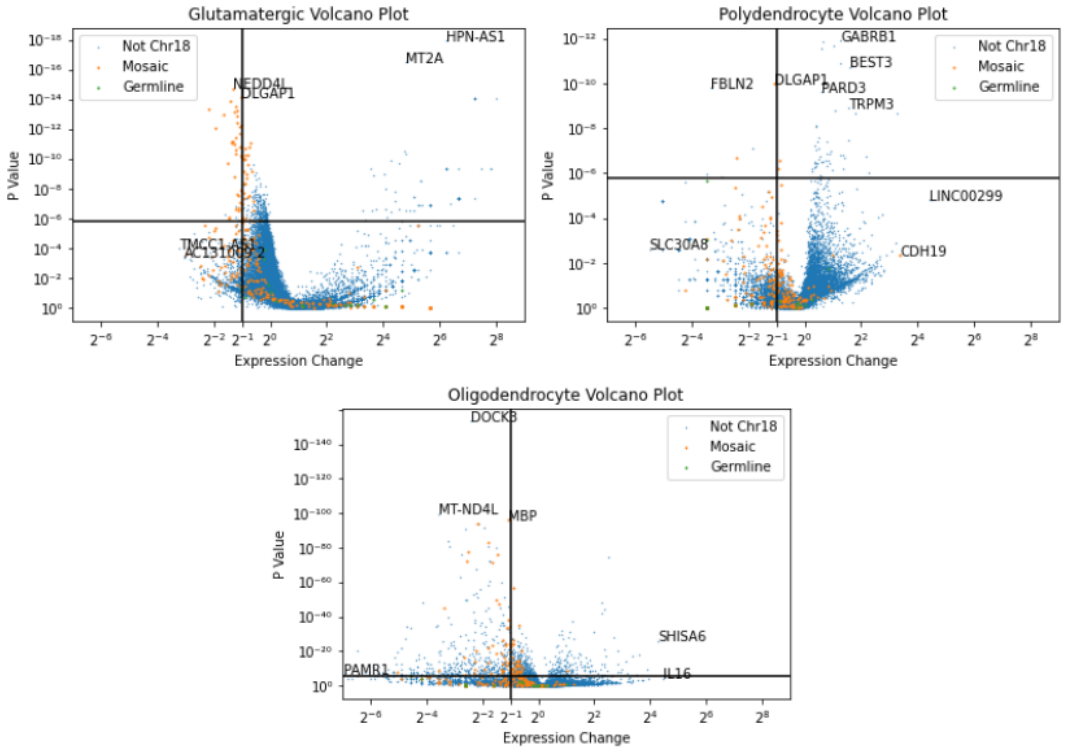

Figure 8: Volcano plots for gene expression change and p value in glutamatergic, polydendrocyte, and oligodendrocyte cells. The blue points represent cells not in Chromosome 18, the orange ones represent cells in the mosaically deleted region of Chromosome 18, and the green ones represent cells in the germline deleted region of Chromosome 18. The horizontal line is the bonferroni correction of the data ($p = \frac{0.05}{n} \approx 10^{-6}$) while the vertical line is $x = 0.5$. The Chromosome 18 genes straddle this line, as we expect since it lost half of the copies of the chromosomes. However we see many non-Chromosome 18 genes having significant expression change, which could mean that biological processes and molecular functions are severely impacted by this loss.

3.5.2 CN-LOH Analysis

We can perform a similar gene expression analysis on CN-LOH samples. The results are displayed in Figure 9. Just like with LO18, we can see a large shift in expression change among the genes not in chromosome 9q. Consequently, CN-LOH mutation has the potential to drastically alter biological processes and molecular functions in their cells. To understand this we further investigate the specific genes that consistently differentially expressed next.

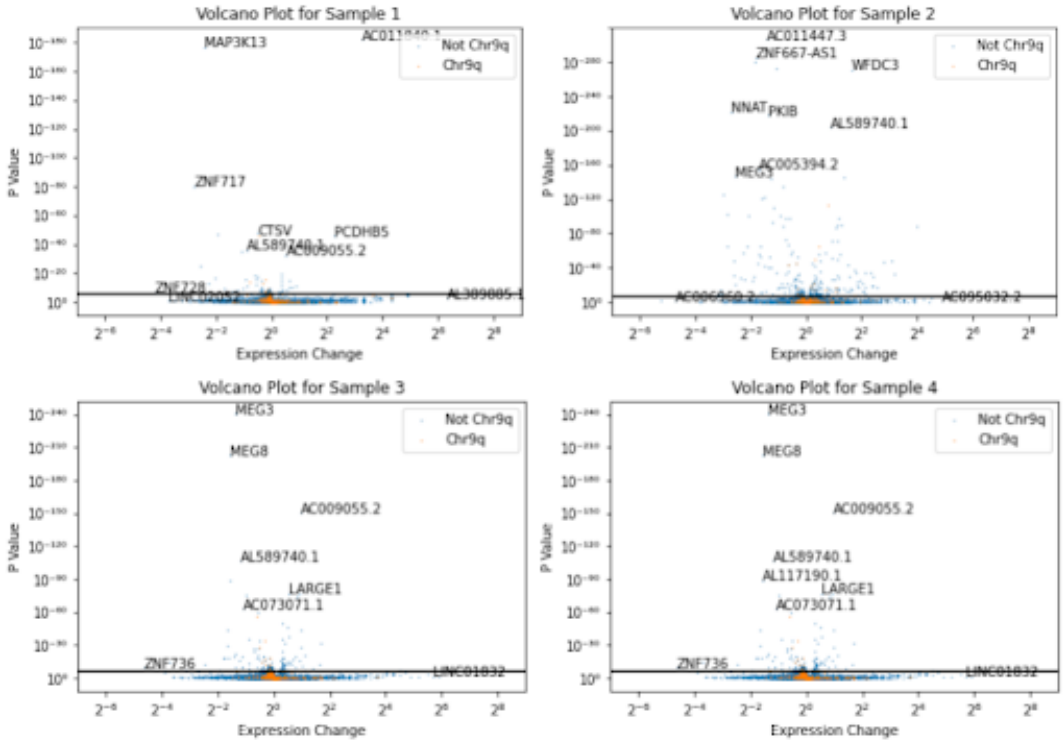

Figure 9: Volcano plots for the four 9q CN-LOH samples' gene expression change vs p value. The blue points represent genes not in chromosome 9q and the orange ones represent genes in chromosome 9q. The horizontal line is the bonferroni correction of the data ($p = \frac{0.05}{n} \approx 10^{-6}$).

3.5.3 Impact of Trans-Chromosomal Regulation under CN-LOH

We used [26] database to locate the genes and [**snyder2015gene**] for researching biological processes controlled by these genes. Specifically, our Figure 9 shows consistent significance for expression of MEG3 located on chromosome 14 and ZNF728, which is located on chromosome 19.

Abnormal expression of MEG3 gene can lead to a variety of problems [19] such as overexpression leading to increased cell proliferation, decreased apoptosis, and enhanced epithelial-mesenchymal transition (EMT). Abnormal expression of the ZNF728 gene in CN-LOH for iPSCs can have a number of biological impacts, including promoting tumor growth and metastasis [21]. Interestingly using a different method [21] also found that The paper also found that ZNF728 overexpression was associated with increased CN-LOH at the ZNF728 locus.

4 Future Work

In blood we now know that clonality in blood is a risk factor for developing blood cancer and, more recently, a connection with inflammatory related diseases was established [13]. Similarly we are interested in exploring whether clonality in brains, specifically Chromosome 18 loss, can be linked to neurological diseases that develop with aging. We also know that the genes EXD3, FANCC, LHX6, SLC46A2, SPTLC1, and TMOD1 are imprinted in 9q [22, 15], so we wish to see how they are affected in these CN-LOH mutations. We also want to understand why this clonality occurred in the first place. Finally, we plan on using gene ontology to figure out what biological processes and molecular functions are affected.

5 Acknowledgements

First and foremost, I would like to thank my direct mentor, Dr. Giulio Genovese, for teaching me about single cell RNA sequencing and patiently guiding me through my research. He helped me understand many of the complex statistical and biological concepts that were critical to the project.

Further, I would like to thank Prof. Steve McCarroll who provided resources, guidance, and encouragement. His weekly lab meetings were very helpful as everyone eagerly chimed in to add their insights on how to push this project further.

I would like to thank Dr. Slava Gerovitch, the head of the MIT PRIMES program, for matching me up with such an interesting solo project at the Broad Institute.

Bibliography

[1] R. Eiben and S. Hahn. "Prenatal diagnosis of monosomy 18 and ring chromosome 18 mosaicism". In: *Prenatal Diagnosis* 12.9 (1992), pp. 687–690. DOI: 10.1002/pd.1340120902.

[2] David J. Sheskin and Brian S. Everitt. *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley, 2000.

[3] C. Yardin et al. "First familial case of ring chromosome 18 and monosomy 18 mosaicism". In: *American Journal of Medical Genetics* 104.3 (2001), pp. 257–259. DOI: 10.1002/ajmg.1040257.

[4] Wenhui Li. "Volcano plots in analyzing differential expressions with mRNA microarrays". In: *BMC Bioinformatics* 13.1 (2012), p. 155. DOI: 10.1186/1471-2105-13-155.

[5] L.M. Zhang et al. "Mosaic loss of chromosome 18 in a patient with ring chromosome 18 and intellectual disability". In: *American Journal of Medical Genetics. Part A* 158A.1 (2012), pp. 151–155.

[6] Cole Trapnell et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature Biotechnology* 32.4 (2014), pp. 381–386. DOI: 10.1038/nbt.2859.

[7] Evan Z. Macosko et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5 (2015), pp. 1202–1214. DOI: 10.1016/j.cell.2015.05.002. URL: https://www.cell.com/cell/pdf/S0092-8674(15)00549-8.pdf.

[8] Amit Zeisel et al. "Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". In: *Science* 347.6226 (2015), pp. 1138–1142. DOI: 10.1126/science.aaa1934.

[9] J. Oyelade et al. "Clustering Algorithms: Their Application to Gene Expression Data". In: *Bioinformatics and Biology Insights* 10 (2016), pp. 237–253. DOI: 10.4137/BBI.S38316.

[10] Patrik L. Ståhl et al. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". In: *Science* 353.6294 (2016), pp. 78–82. DOI: 10.1126/science.aaf2403.

[11] Itay Tirosh et al. "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". In: *Science* 352.6282 (2016), pp. 189–196. DOI: 10.1126/science.aad0501.

[12] Allon Wagner, Aviv Regev, and Nir Yosef. "Revealing the vectors of cellular identity with single-cell genomics". In: *Nature Biotechnology* 34.11 (2016), pp. 1145–1160. DOI: 10.1038/nbt.3711.

[13] S. Jaiswal et al. "Clonal hematopoiesis of indeterminate potential and risk of cardiovascular disease". In: *Nature Medicine* 23.1 (2017), pp. 119–127.

[14] M.L. Loh et al. "Clonal hematopoiesis and age-related diseases". In: *Nature Reviews. Disease Primers* 4.1 (2018), p. 18008.

[15] D.M. Zink et al. "Epigenetic regulation of imprinted genes at 9q34". In: *Nature Reviews. Genetics* 19.6 (2018), pp. 341–355.

[16] Junyue Cao et al. "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566.7745 (2019), pp. 496–502. DOI: 10.1038/s41586-019-0969-x.

[17] R. Liu et al. "Single-cell genomics identifies mosaic loss of Y chromosomes in aging males". In: *Nature* 565.7736 (2019), pp. 358–362.

[18] C. Thompson et al. "Clonal hematopoiesis as a risk factor for age-related diseases". In: *Nature Medicine* 25.10 (2019), pp. 1429–1438.

[19] Qi Li et al. "MEG3 promotes tumor growth by regulating cell cycle progression, apoptosis, and epithelial-mesenchymal transition in human glioma cells". In: *Cell Death and Disease* 11.11 (2020), p. 918. DOI: 10.1038/s41419-020-2932-x.

[20] Steven McCarroll et al. "SCPred: A deep learning-based algorithm for cell type classification in brain tissue". In: *Nature Methods* 17.2 (2020), pp. 195–202.

[21] Y. Ma et al. "ZNF728 promotes tumor growth and metastasis by regulating the expression of EMT-related genes in iPSC-derived breast cancer cells". In: *Cell Death and Disease* 12.3 (2021), p. 364. DOI: 10.1038/s41419-021-02702-3.

[22] N. Akbari, S. Moghaddam, and A. Ghaffari. "Imprinted genes at 9q34 and their role in human diseases". In: *Frontiers in Genetics* 13 (2022), p. 801768.

[23] Y. Chen et al. "Single-cell RNA-seq of frozen brain nuclei using droplet digital PCR". In: *Nature Methods* 19.2 (2022), pp. 175–183. DOI: 10.1038/s41592-021-01539-w.

[24] T.F. Merkle et al. “Whole-genome analysis of human embryonic stem cells enables rational line selection based on genetic variation”. In: *Cell Stem Cell* 29.3 (2022), 472–486.e7.

[25] M. C. Vermeulen et al. “Mosaic loss of Chromosome Y in aged human microglia”. In: *Genome Research* 32.10 (2022), pp. 1795–1807.

[26] *The National Center for Biotechnology Information (NCBI)*. `https://www.ncbi.nlm.nih.gov/`.