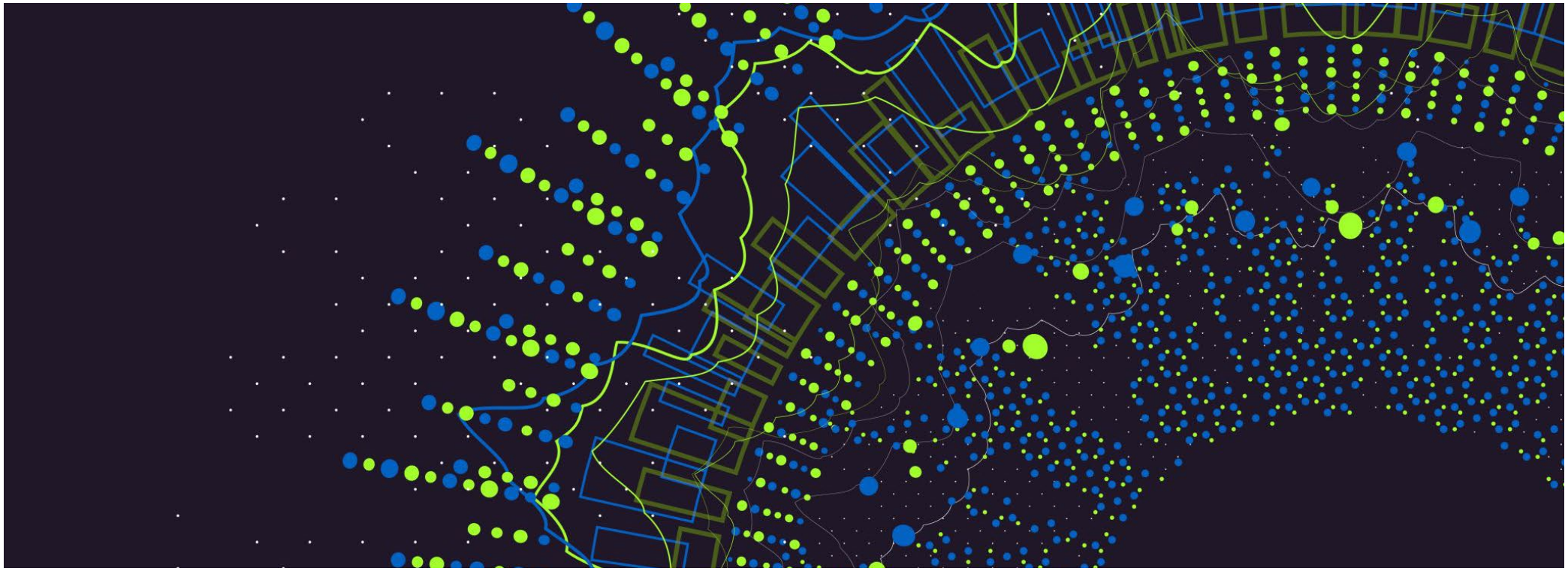


Title:

A Novel Statistical Method for Identifying Cells with Mosaic Alterations/Loss in Single-cell Sequenced Data.

Amith Saligrama

Commonwealth School, Boston MA 02215

MIT PRIMES PROJECT

Mentors : Dr. Giulio Genovese & Prof. Steve McCarroll

McCarroll Lab

Introduction

- **Questions motivating my Project:**
  1. Can we develop statistical tools for pinpointing which cells in a sample (e.g. brain cells of a person) has chromosomal mutations?
  2. Can we uncover gene expression patterns that are unique to mutated cells?

- **Why study Chromosomal Mutations?**
  - Are chromosomal alterations in neurotypical individuals the first steps to the development of overt brain cancer"?

- **Why cell-level classification?**
  - We can understand the heterogeneity of mutations, which can be crucial for diagnosis, treatment, predicting disease progression, and personalized treatment.

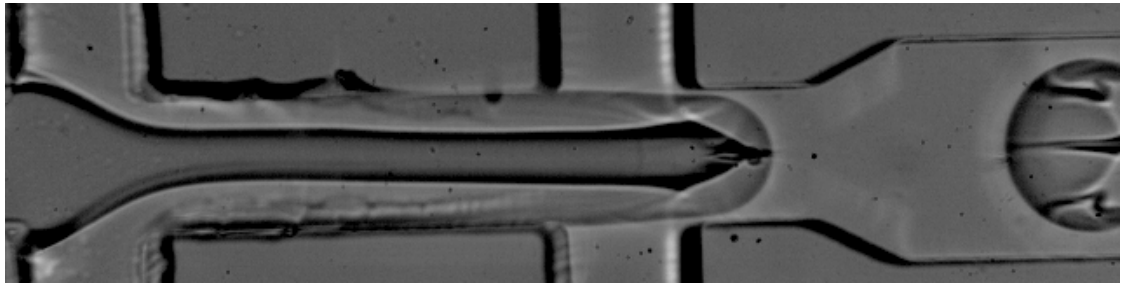
[Macosko et. al. 2015]

Drop-Seq: A Single-cell-RNA-sequencing Method

- Droplets isolate individual cells
  - Barcodes for each cell.
- Digital Gene-Expression for each cell
  - Cell lysed → mRNA molecules associated with each gene identified.
  - Gene expression "essentially" #mRNA molecules detected
  - Chromosomal counts = Group gene counts.

Cell:	1	2	...	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

- Contrast with bulk-analysis
  - "Fruit salad vs. fruit smoothie"



Datasets

Loss-of-18 - Brain Cells Dataset

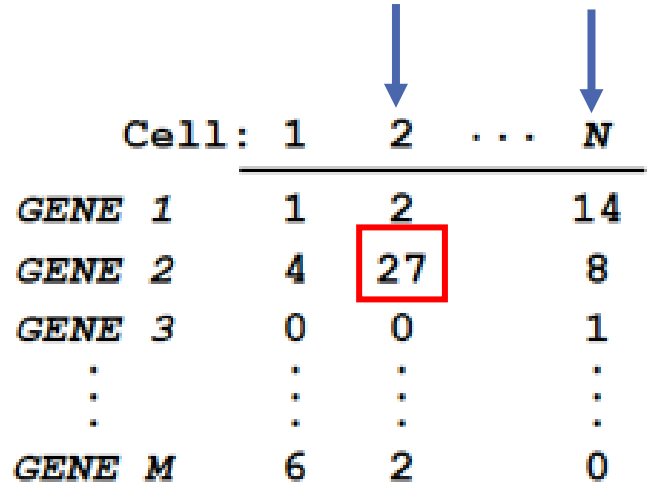
- 7-Cell-types: Astrocyte, Gabaergic, Glutamatergic, Polydendrocyte, Oligodendrocyte, Endothelia, and Microglia
- Raw Data: DGE Matrix for each type.
- (i, j) component: counts for gene i and cell j.

Cell:	1	2	...	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

Context: **The dataset is from a person with known ring 18 chromosome. Evidently cells recurrently lose chromosome 18 as a result.**
[Yardin et.al 2001, https://pubmed.ncbi.nlm.nih.gov/11754054/]
This data serves as a ground-truth for testing and validation for our approach

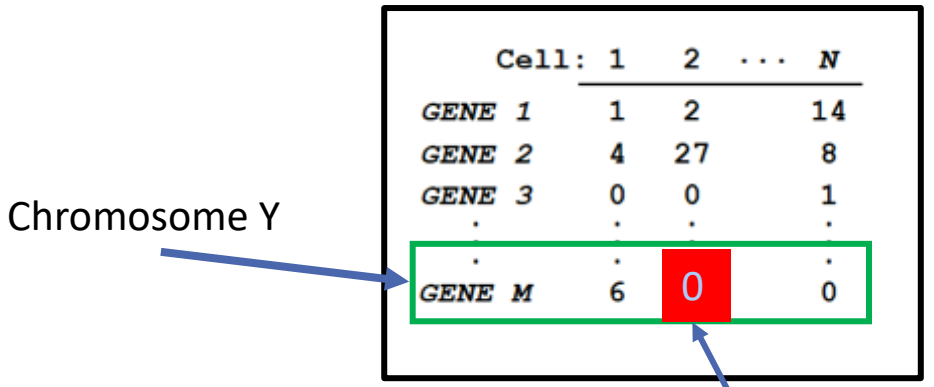
Problem Statement: Mutated Cell-Identification

- Normal Cell:
  - Cell j is normal (jth column) expression is statistically consistent with normal cells.
- Mutated Cell:
  - Column j has subset of rows, (e.g. genes k, l, m in chromosome xx) that are statistically abnormal.
- Problem:
  - Identify cells (columns) that are mutated.



Prior Works

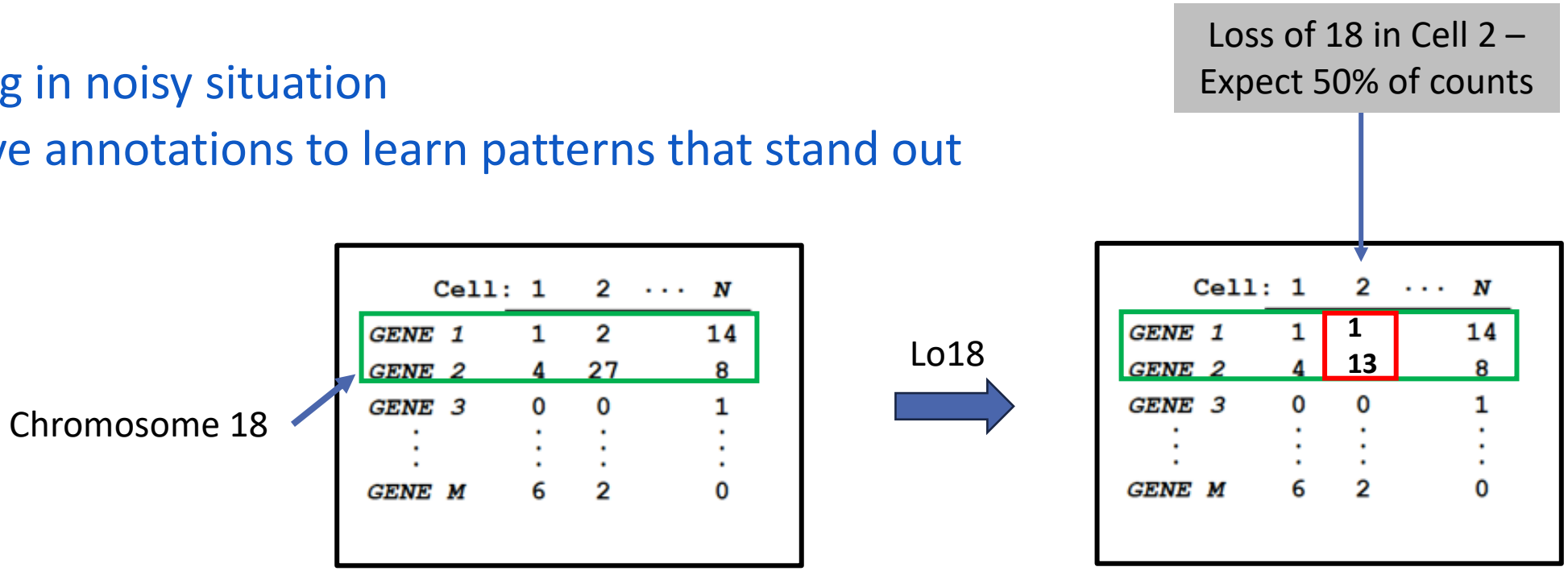
- Large-scale Bulk RNA [Anders 2013]
  - *average* analysis



- Cell-by-cell identification ([Vermeulen et.al 2022])
  - Loss of (Sex) Chromosome Y (LoY)

Non-Sex Chromosome

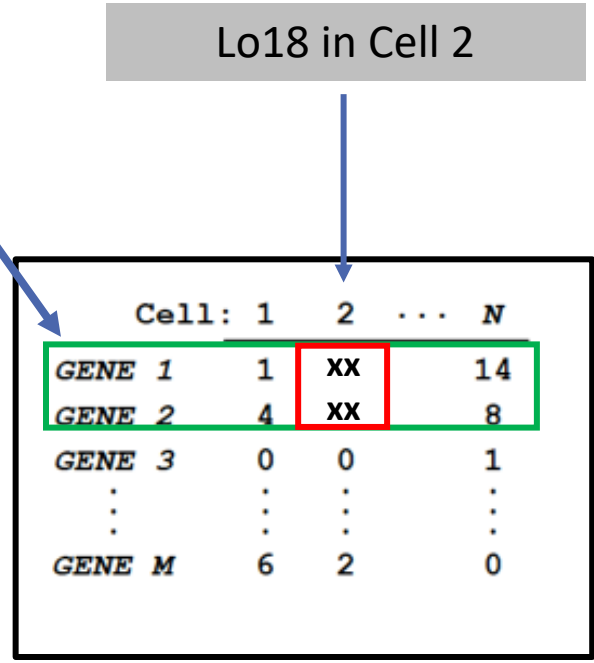
- Chromosome pair – Loss of one expected to change count.
  - Yet - in a Perfect World (if no Noise)
    - Count reduces by one-half!!
  - Challenging in noisy situation
  - Do not have annotations to learn patterns that stand out



Non-Sex Chromosome Loss Detection

- Chromosome pair – Loss expected to change count.
  - Loss difficult to "predict" purely from counts.

- Sampling Noise in DGE:
  - Technical Variations
    - # Reads/Cell, Amplification Noise, Read Efficiency etc.
  - Biological Variations
    - Cell diversity – not all cells are identical



Probabilistic Model

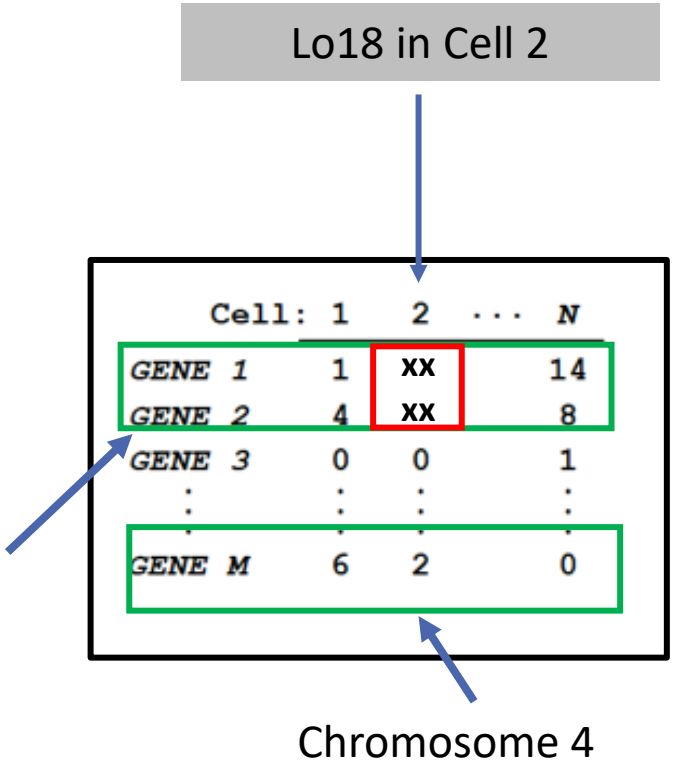
Key Idea (Control Chr): *statistically independent* of target chromosome.

- Validated with Bulk MoChA analysis "Chromosome 4" independent of Chromosome 18.

- Property 1:

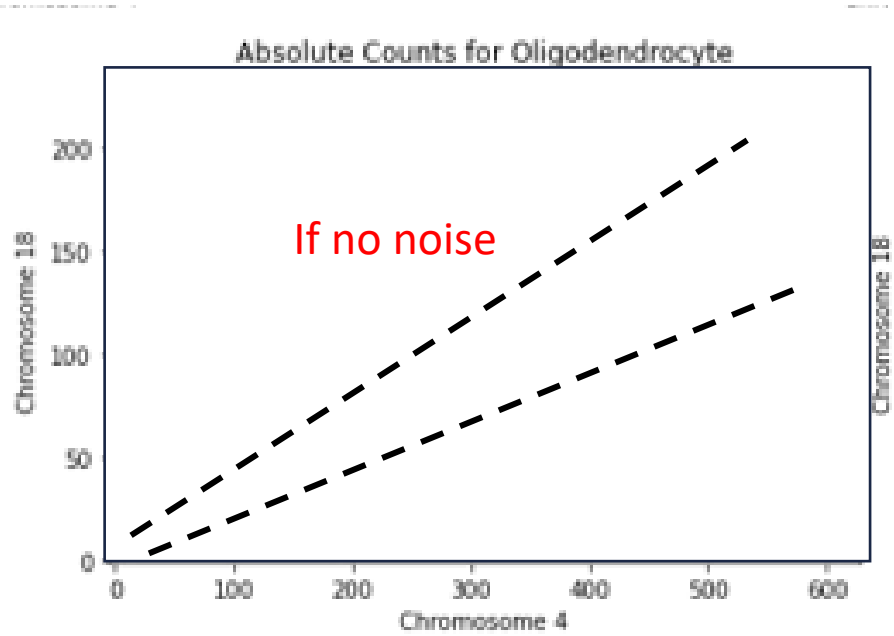
$$Prob[Count(\text{Gene } M) = c \mid \text{Loss of 18}]$$
$$= Prob[Count(\text{Gene } M) = c \mid \text{any Cell}]$$

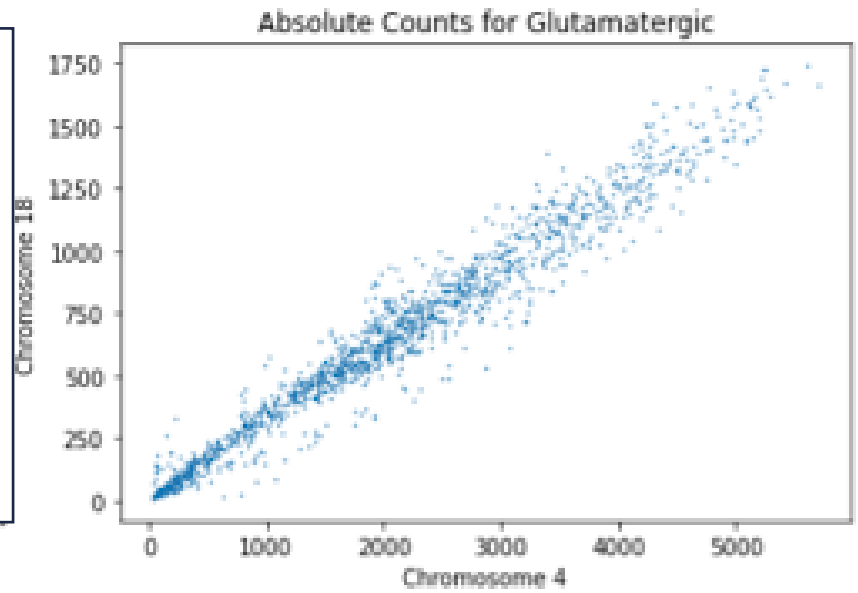
- Property 2: On average, count in Loss of 18 cell (e.g. cell 2 and gene 1) is ½ of average counts of normal cell.

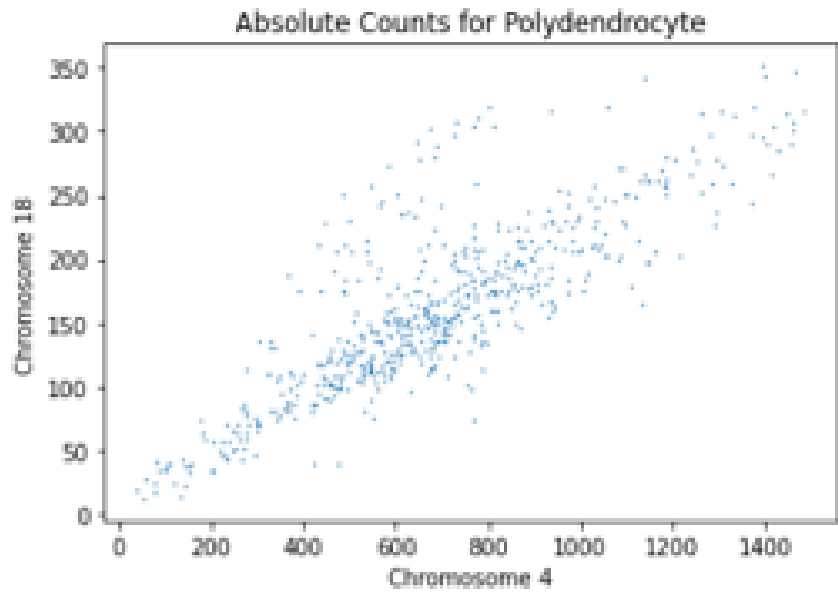


https://software.broadinstitute.org/software/mocha/

Scatter Plots: Diverse Cell types with Lo18 (Brain Cells)







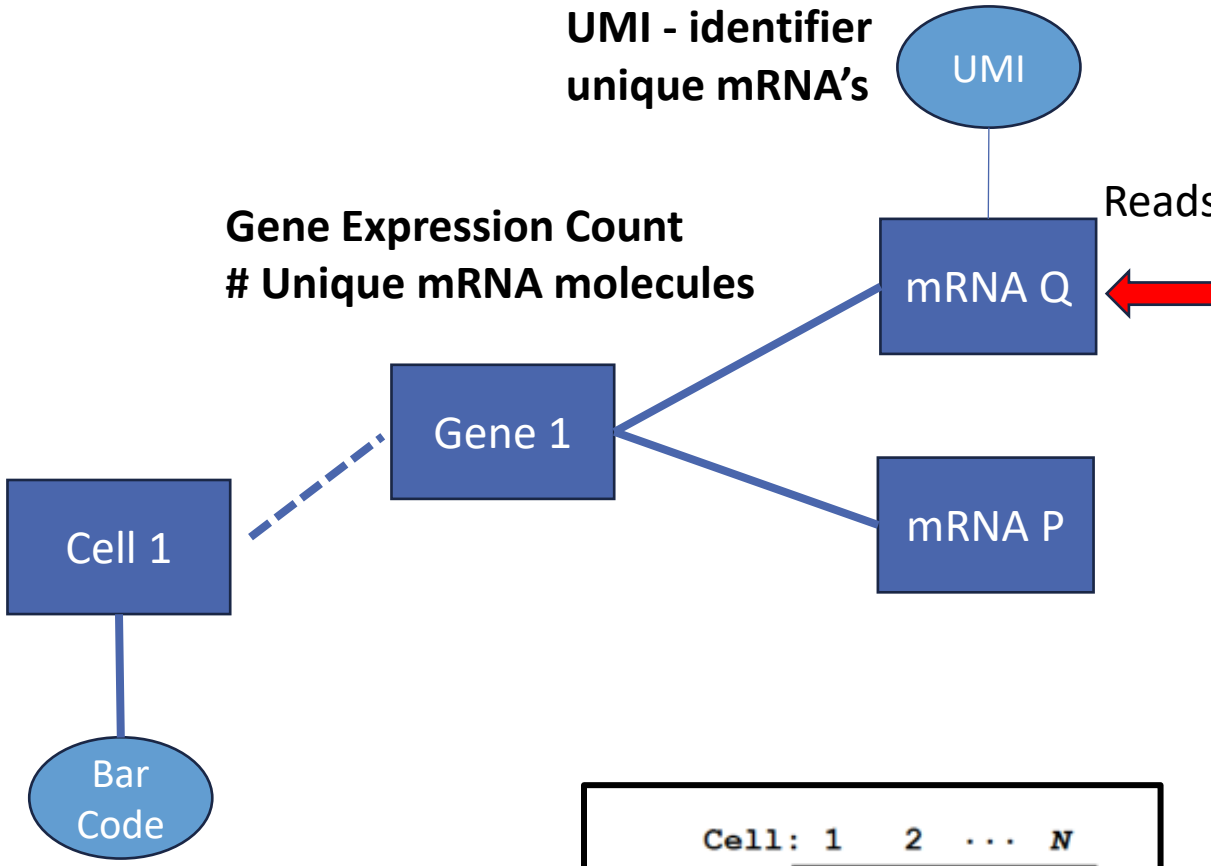
- Validation
  - MoChA study – No impact of Lo18 on CH4
- Probabilistic Framework – Scatter plots depict ploidy of Lo18 cells
  - Evident separation into different clusters

Binomial Model

- Assumptions:
  - Detection prob constant across molecules
  - Reads uniform distributed across molecules
  - (recall) CH4 independent of CH18

- Prob Count(CH18) given CH4+CH18:
  - Each count a coin toss;
  - p: success prob of CH4 count. So,

$$\text{CH4} \sim Bin(N, p)$$
$$\text{CH18} \sim Bin(N, 1-p)$$



Cell:	1	2	...	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
:	:	:		:
GENE M	6	2		0

Binomial Model

- Assumptions:
  - (recall) E[Counts(CH18) | Loss]=0.5 E[Counts(CH18)|No Loss]
- Two Cases:
  - No Loss Cell

$$\begin{aligned} \text{CH4} &\sim Bin(N, p) \\ \text{CH18} &\sim Bin(N, 1-p) \end{aligned}$$

  - Lossy Cell

$$\begin{aligned} \text{CH4} &\sim Bin(N, q) \\ \text{CH18} &\sim Bin(N, 1-q) \end{aligned}$$

Odds ratio (success vs. failure) for Lossy CH18 cell is twice as likely!!

Think of a casino with two tables
Table 1: CH4 against lossy CH18,
Table 2: CH4 against normal CH18.
Ist table odds 1:2 means 2nd is 2:2)

- How would p and q be related?

$$\frac{q}{1-q} = 2\frac{p}{1-p}$$

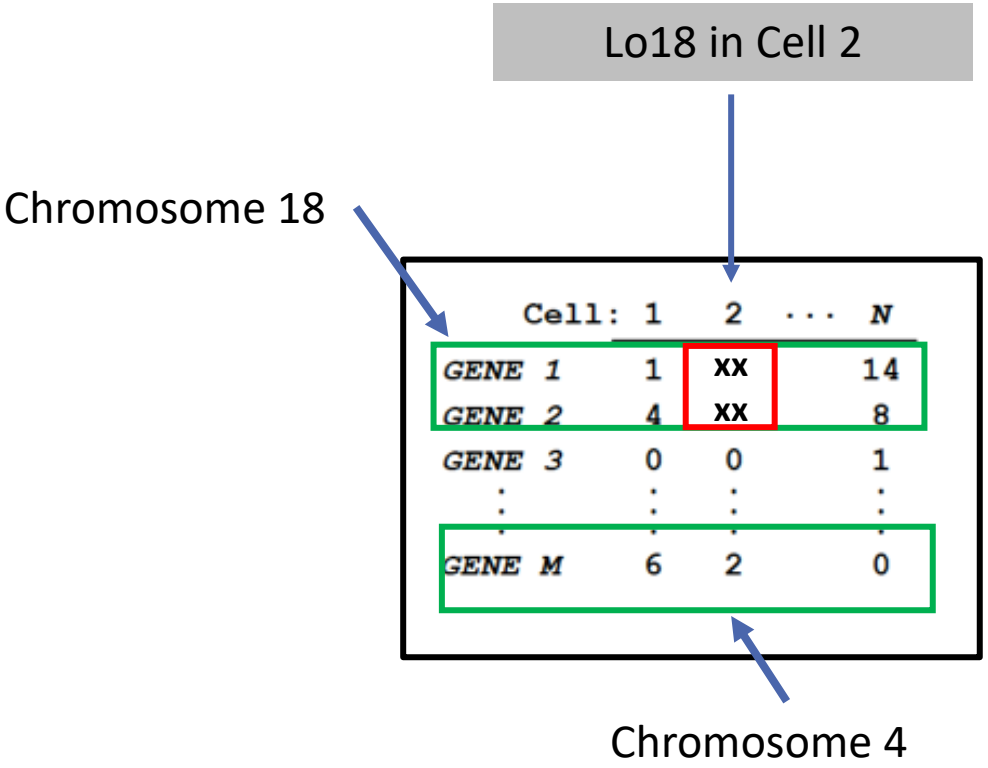
Validation

- Cluster ratios, i.e., for each cell i

$$\rho_i = \frac{\#[G_{ij}]_{j \in CH4}}{\#[G_{ij}]_{j \in CH4} + \#[G_{ij}]_{j \in CH18}}$$

- Validate cluster median, $\hat{p}, \hat{q}$ satisfy:

$$\frac{\hat{q}}{1-\hat{q}} \approx 2\frac{\hat{p}}{1-\hat{p}}$$



DGE Matrix **G**

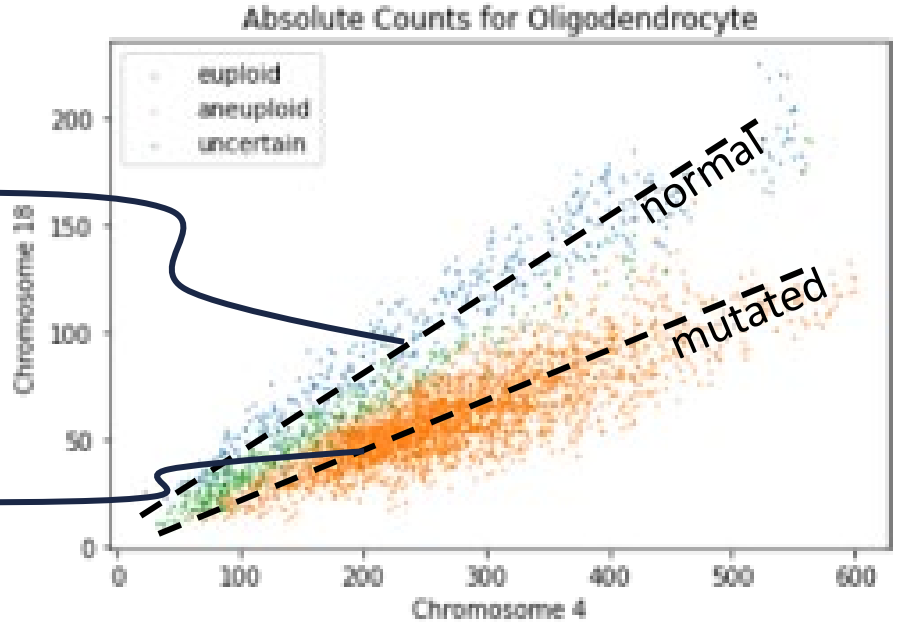
Detected Cells with Binomial Model

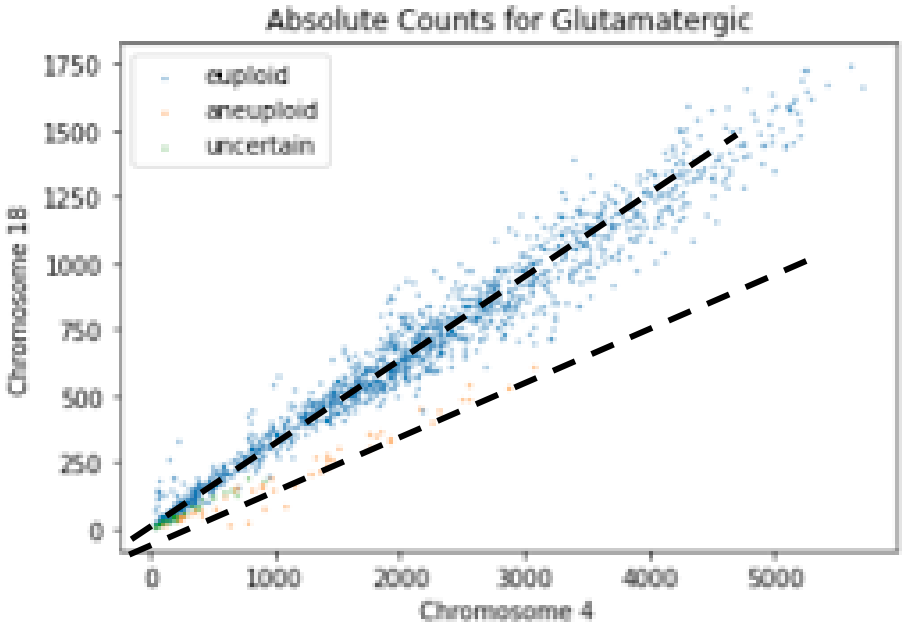
slope ratio=2

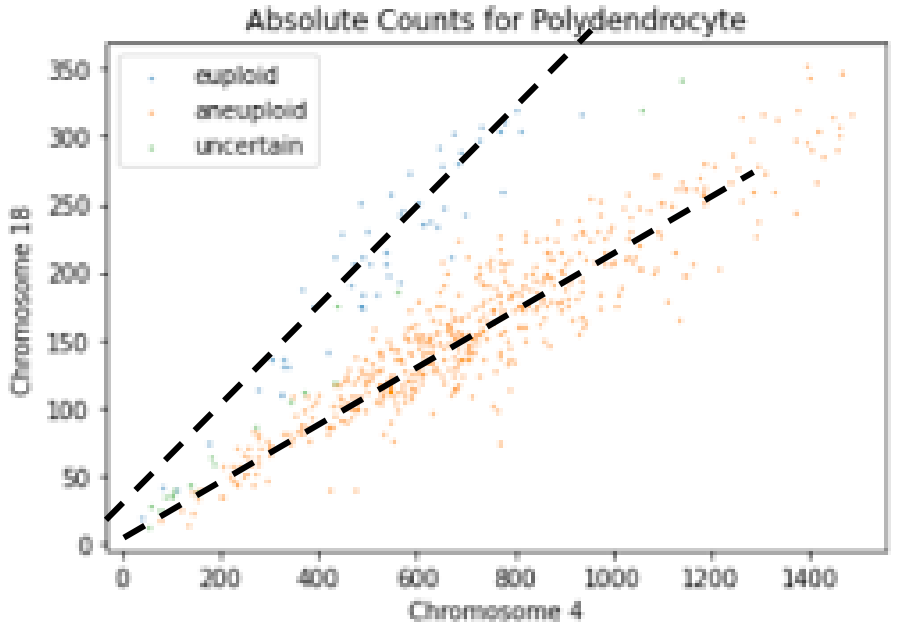
$$\frac{\hat{q}}{1-\hat{q}} \approx 2\frac{\hat{p}}{1-\hat{p}}$$

$$\text{Slope} = \frac{1-\hat{p}}{\hat{p}}$$

$$\text{Slope} = \frac{1-\hat{q}}{\hat{q}}$$

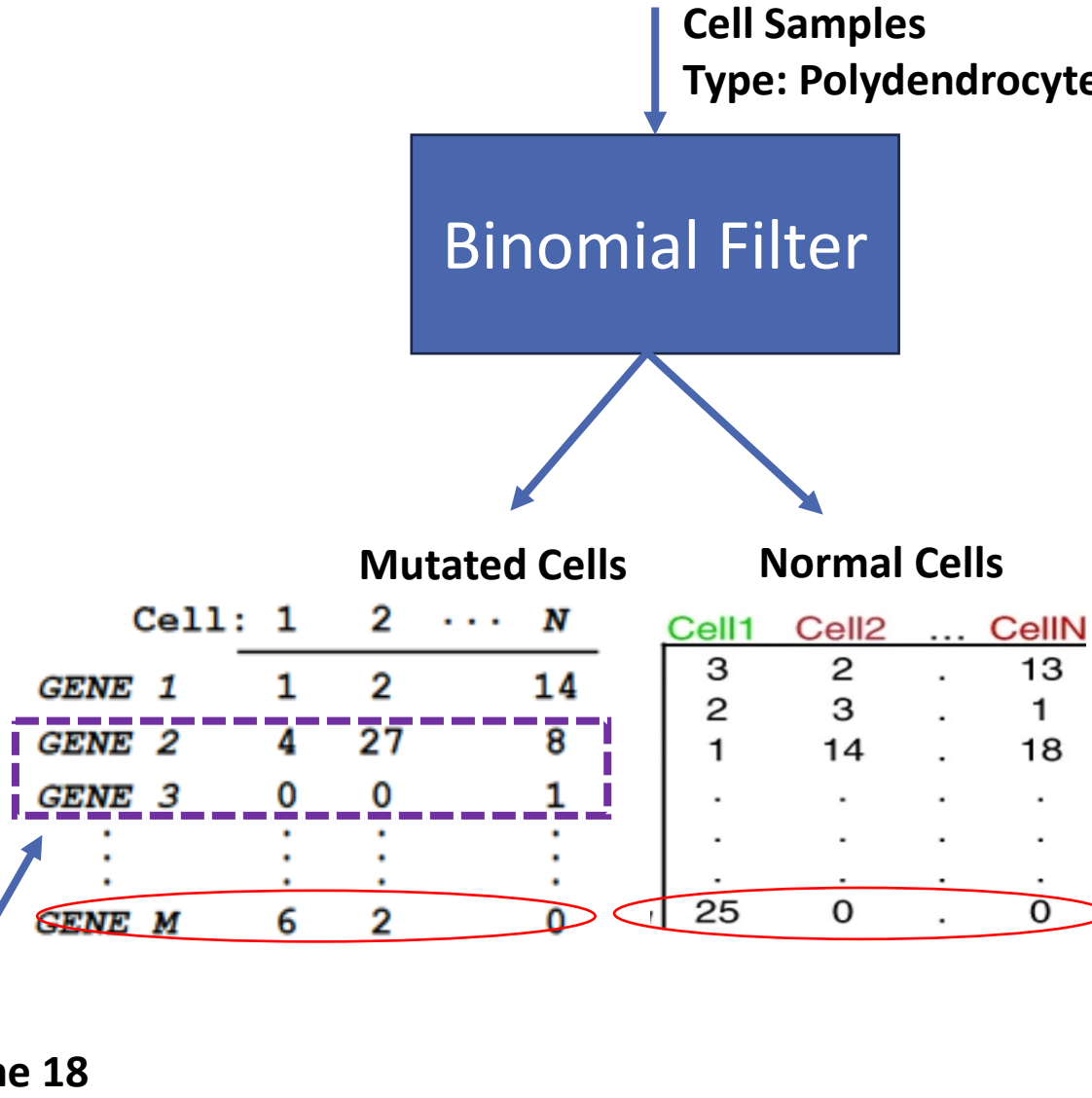




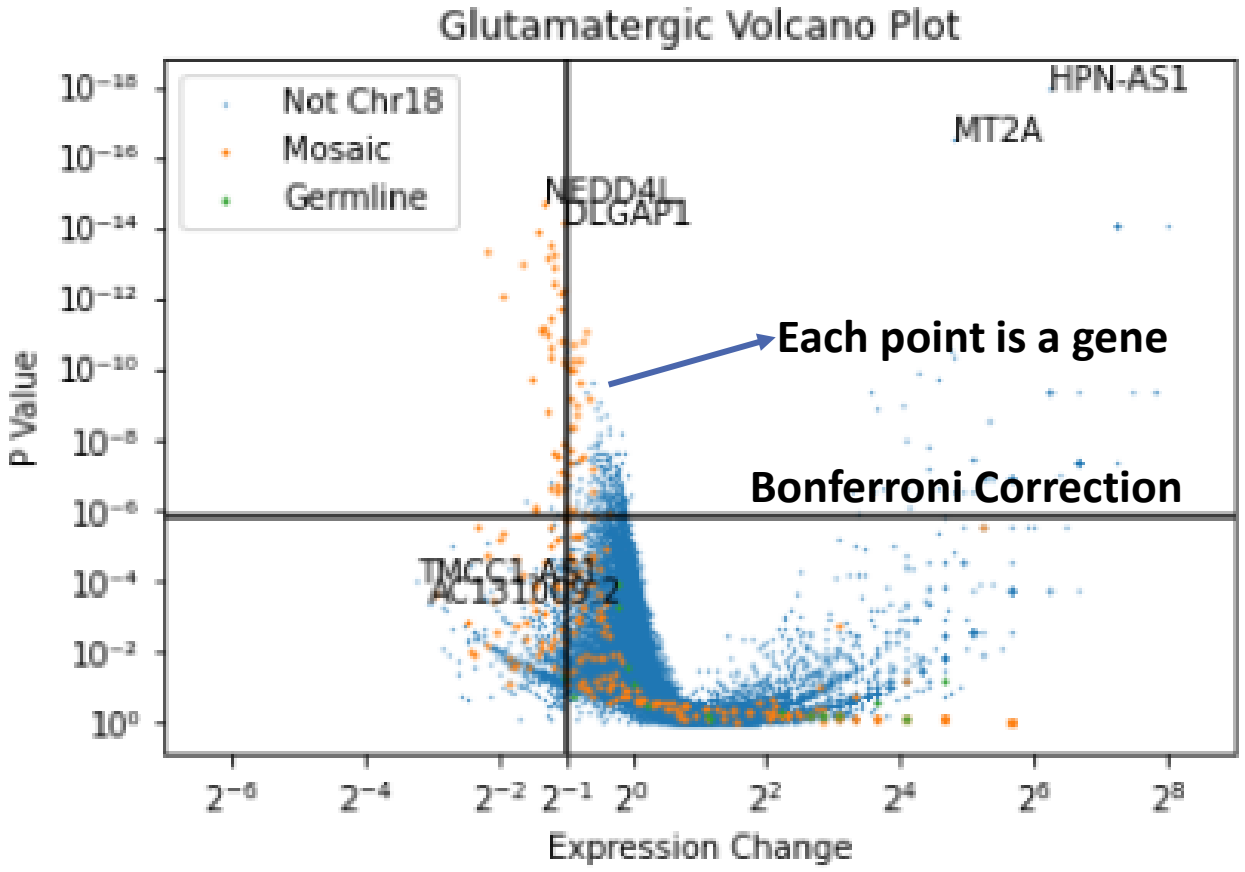


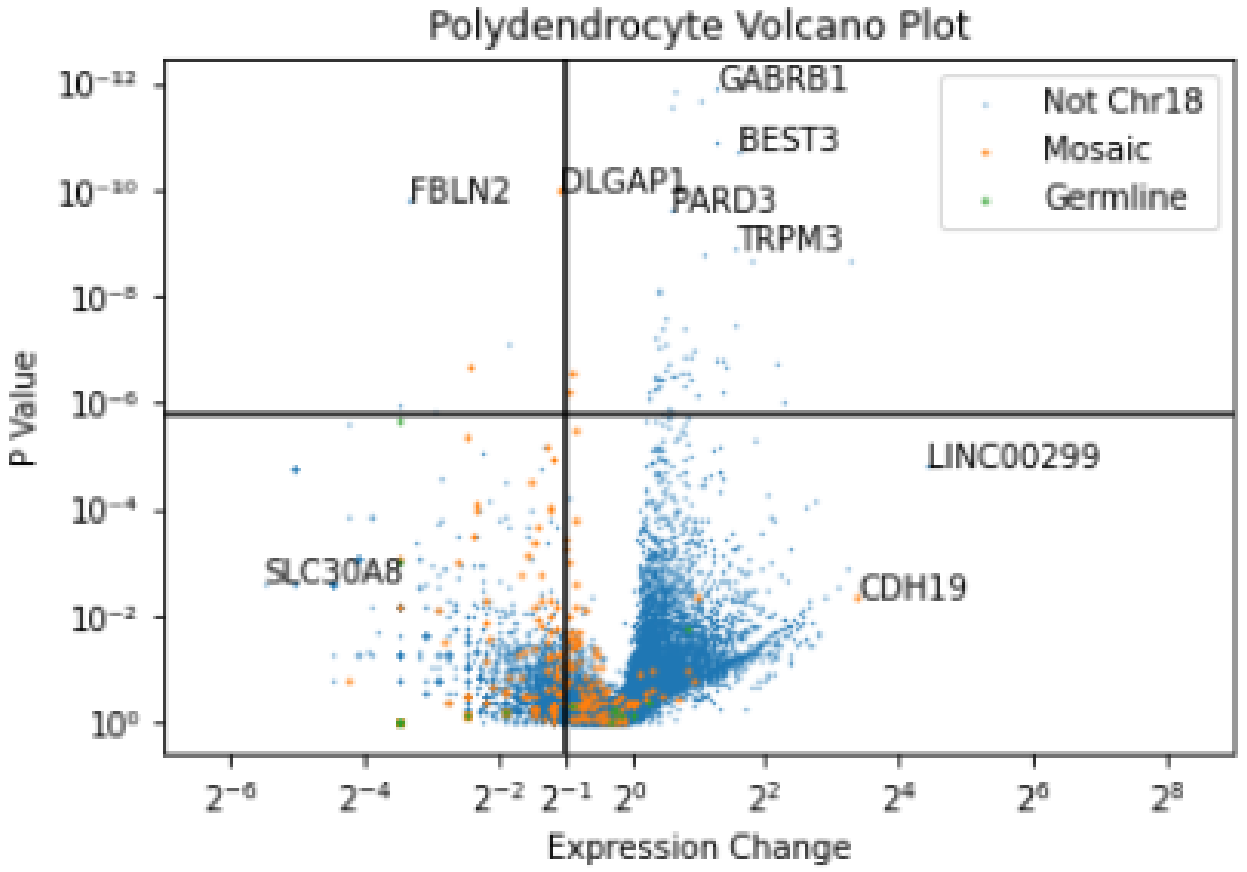
Trans-Chromosomal Expression in Mutated Cells

- Mutated vs. normal
  - Gene A expression different in mutated
  - Null: No statistical difference
- Wilcoxon Rank-Sum Test (p-value)
  - Non-parametric test - independent populations
  - Works well with small counts.
- Multiple-comparisons
  - Bonferroni Correction
  - Burden of simultaneous gene comparisons



Volcano Plots – Expression Change vs. Significance





- Volcano Plot: P-Value vs. Gene k Expression Change (odds-ratio - gene k vs. control)
- After Bonferroni correction – for many non-Chromosome 18 genes
  - expression change statistically significant (adj p-value (0.05))

Conclusions

- We show that it is possible to classify single brain nuclei from post-mortem samples as whether they harbor Chromosome 18 loss or not
- We show that Loss of Chromosome 18 can affect the majority of oligodendrocytes and polydendrocytes of a normal person (no specific neurological phenotype at the time of death).
- We show that we can identify gene expression differences beyond chromosome 18 within each cell type mosaic for Loss of chromosome 18
- Future directions:
- We have preliminary extensions of our method for analyzing 9q Copy Neutral-loss of heterozygosity in Induced Pluripotent Stem Cells.
- Extend work to other samples to identify gene expression differences consistent across multiple individuals

Acknowledgements

- I would like to thank my mentor Dr. Giulio Genovese for his mentorship and advice over the last 10 months.
- Prof. Steve McCarroll was a sounding board and generously offered advice during the project.
- Dr. Nicole Rockweiler and Bob Handsaker at the McCarroll lab for making me feel at home and patiently answering my many questions.