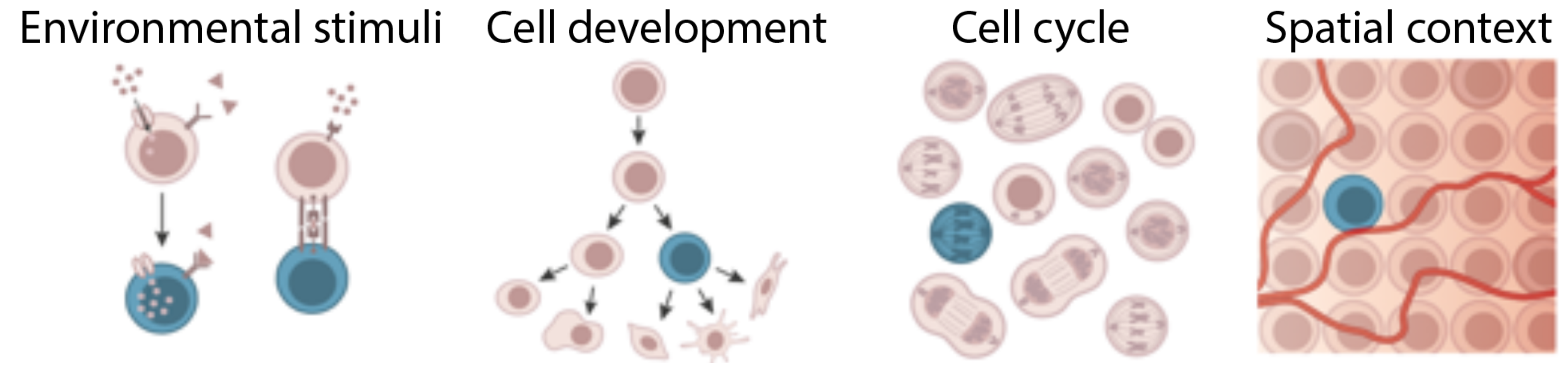
UNIVERSITY OF VIRGINIA
DATA SCIENCE INSTITUTE

Gene Set Creation Algorithm for Microarray Studies with Low Sample Size

Erik Langenborg, Kevin Sun, Lingfeng Cao, Christopher Overall, and Abigail Flower

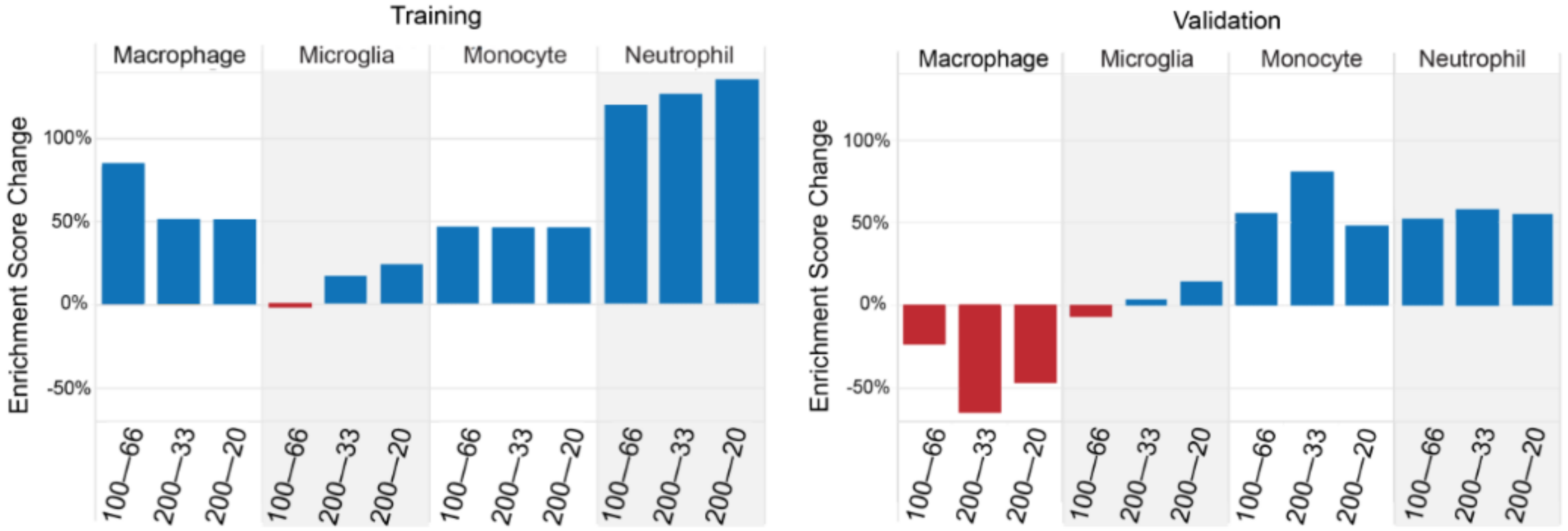
Problem Statement

- Classify cell types based on high-dimensional microarray gene expressions
- Data contained 137 observations and 20,270 predictors
- Identifying discriminatory genetic markers for cell types is challenging due to many factors



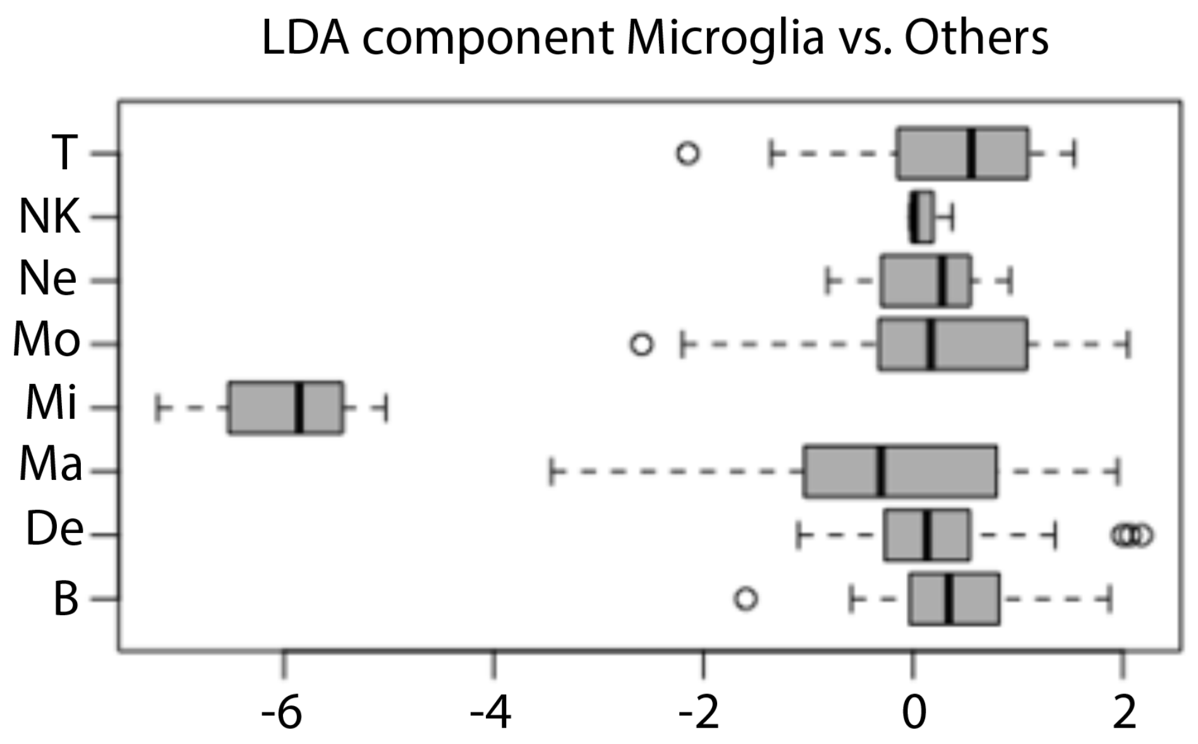
Pipeline Results

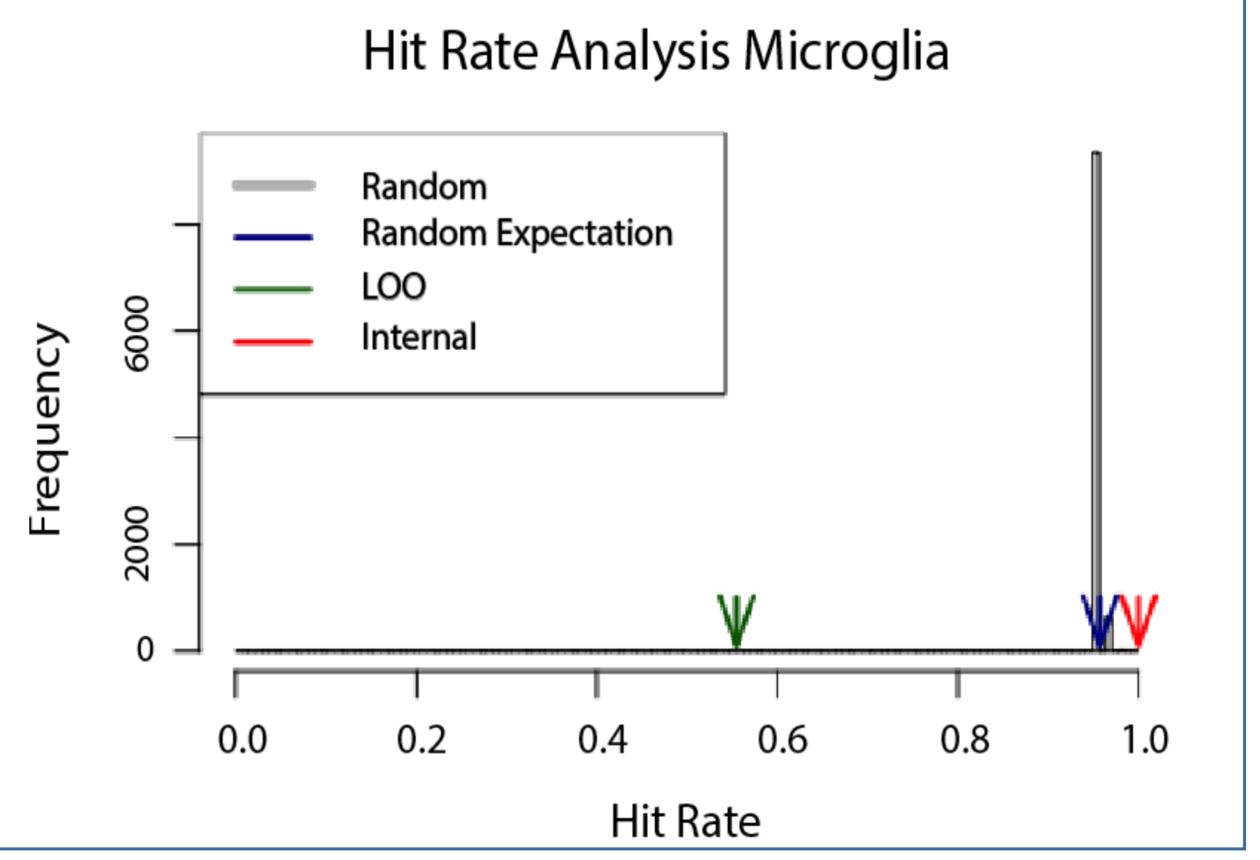
- Performance is robust for training data
- Poor performance on macrophages for validation set
- Strong performance on monocytes and neutrophils on validation set
- In the figures, the first number represents the number of bootstraps, the second number represents the percentage of genes used for each bootstrap



Methodology

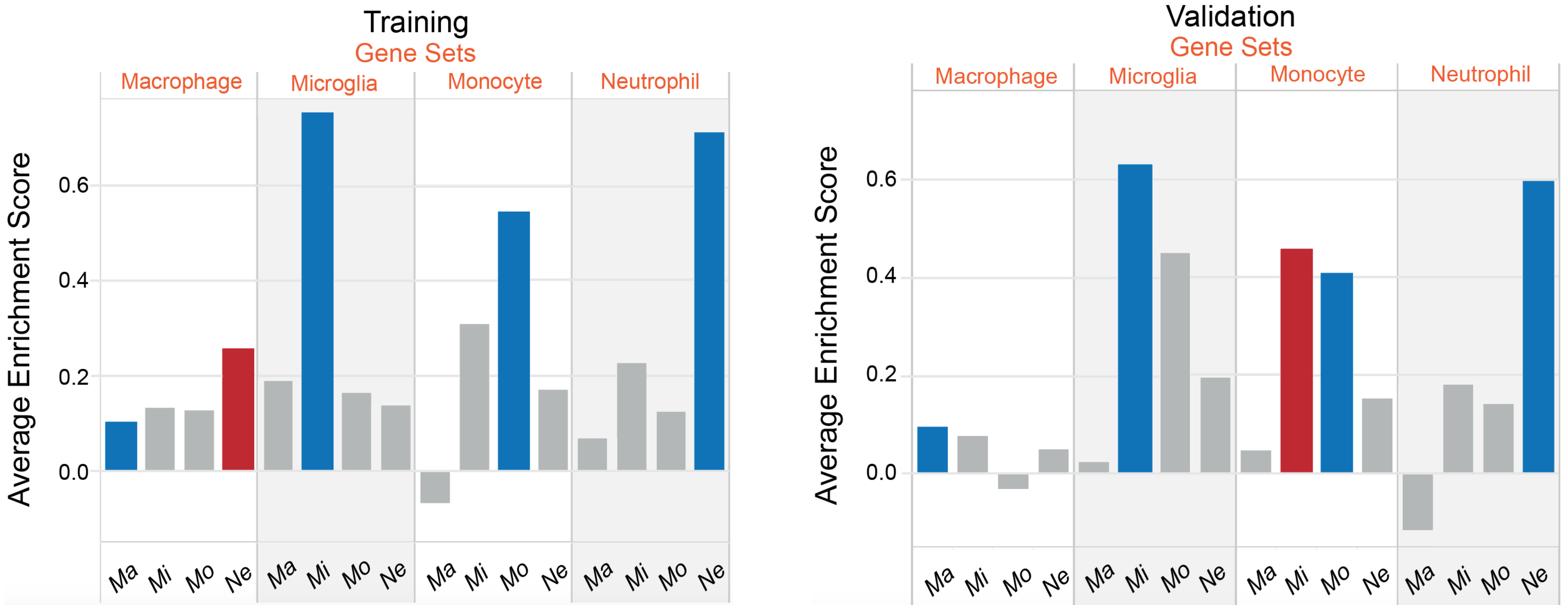
- Used LDA for dimension reduction
- Unlike PCA, LDA preserves class discrimination information
- Preliminary results show good class separation, but may be prone to overffitting





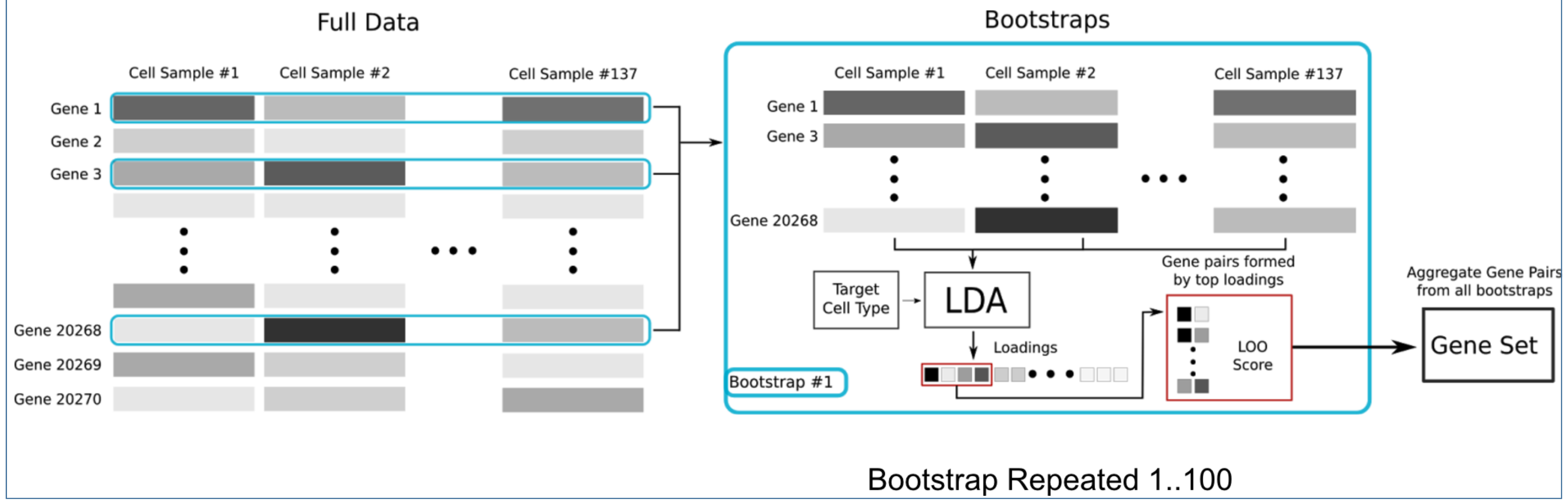
Gene Set Evaluation

- Neutrophil gene set had strong differential enrichment
- Microglia gene set had relatively strong differential enrichment
- Macrophage and Monocyte gene sets were differentially enriched the greatest in other cell types



Pipeline

- Bootstrapped, LOOCV, down-sampled variant of LDA
- For each cell type in each bootstrap, a subset of genes are extracted, and an LDA component was formed such that the top X genes were paired
- F-score ascertained by LDA prediction with LOOCV for a gene pair and all gene pairs were merged with a final score
- Gene sets were created by linking significant pairs with common genes, and evaluated using GSVA

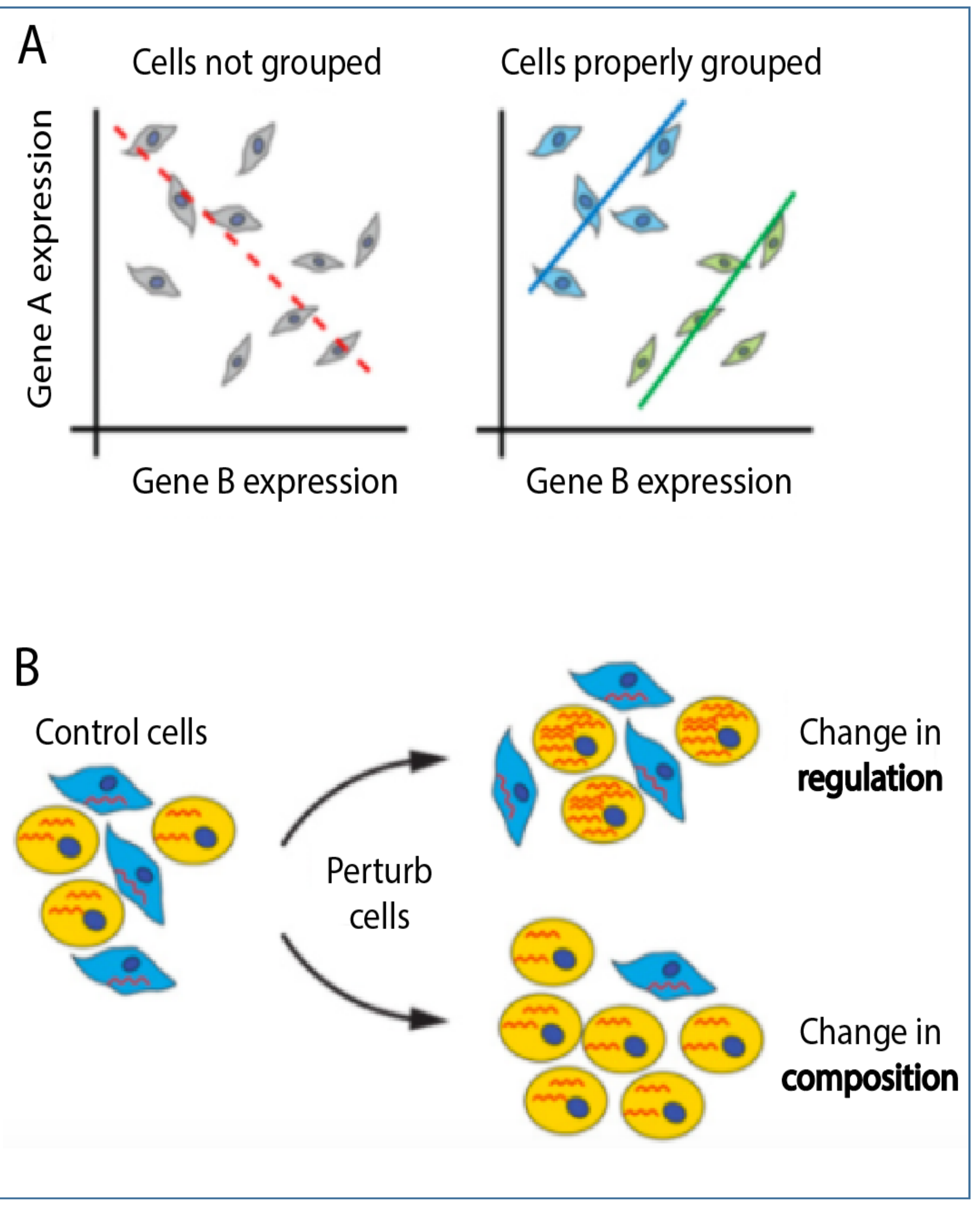


Future Work

- Variance may be important
  - Combine PCA and LDA
- Data may not be linear
  - CDA/QDA
- Bulk microarray array averages signals:

1. Simpson's Paradox (A): Failing to properly subgroup the data by cell type can lead to incorrect correlation analysis
2. Causation of gene expression change (B): Bulk averaged measurements cannot distinguish whether the cause of changes in gene expression was due to a changes in cell compositions or changes in regulation
   - Attain Single-cell data

- Further hyper-parameter tuning
  - The research is currently creating bigger gene sets


A
Cells not grouped
Cells properly grouped
Gene A expression
Gene B expression
Gene B expression
B
Control cells
Perturb cells
Change in regulation
Change in composition