P028

Doublets Detection in Single-Cell DNA-Sequencing using Deep Learning

Sombeet Sahu[1], Manimozhi Manivannan[1] ,Shu Wang[1], Dong Kim[1], Saurabh Gulati[1], Nianzhen Li[1], Adam Sciambi[1], Christine Scherp[1] and Nigel Beard[1]
[1] Mission Bio, South San Francisco, CA, USA
Conflicts of interest: S.S., M.M., S.W., D.K., S.G, A.S., N.L., C.S, N.B. are employees and shareholders of Mission Bio, Inc.

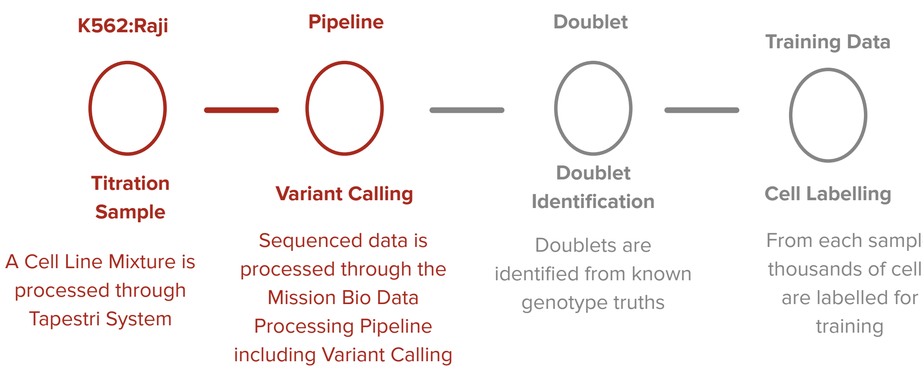
Abstract

Background
Tapestri is a high-throughput single-cell DNA analysis platform that leverages droplet microfluidics and multiplex-PCR based targeted sequencing approach. Often times in droplet-based protocols, cell doublets are an issue which limits cell throughput and results in spurious signals which contribute to false positives in genotyping. To address this issue, we here present a doublet identification tool based on deep neural networks.

Methods
We processed 60 samples of Raji : K562 cell line mixture (50%:50%) each via Tapestri. Four known loci that are genotypically distinct between the two cell lines assigned each cell to K562, RAJI or a doublet. We also used fluorescent images of cells to confirm the doublet rate. These data are used as ground truth for our classifier.

We next train neural networks on a binary classification label using the ground truth on the amplicon-cell read matrix. We set up a densely connected neural network classifier using tensorflow. The number of hidden layers in the classifier is equal to the number of amplicons in the targeted panel. We apportion data from the 60 samples into training and test datasets. We train and test separately on multiple replicates of experiments. The hyperparameters were further optimized for low and high performing amplicons since their distribution of reads arenoisier. To improve accuracy, we introduced artificial doublet by randomly sampling barcodes without replacement from the pool and adding to the pool. Results show detection of ~50% of doublets confirmed by ground truth. Further training with more cells and hyper-parameter optimization will further improve accuracy.

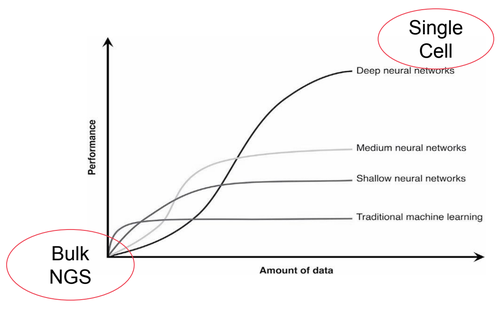
Experimental Steps and Truth Data



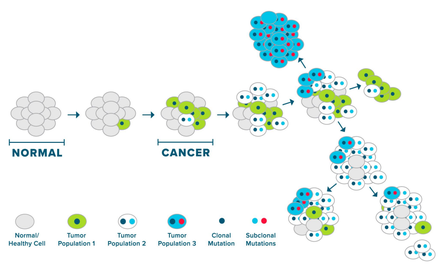
Locus	K562	RAJI	REF	ALT
chr4_55599436	2	0	T	C
chr17_7578523	2	0	T	TG
chr16_55770629	2	0	C	T
chr14_56969005	2	0	C	T
chr7_148504818	0	1	A	G
chr17_7577581	0	1	A	G
chr17_7578211	0	1	C	T
chr6_17076917	0	1	ATAAG	A

Distinguishing mutations to identify double truth

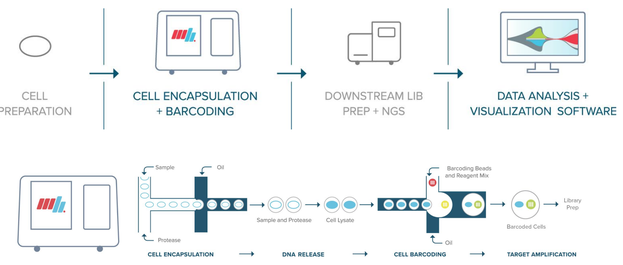
Why Deep Learning for Single-Cell ?



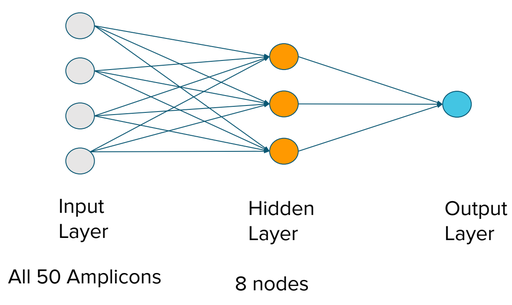
Complexity of Cancer Evolution



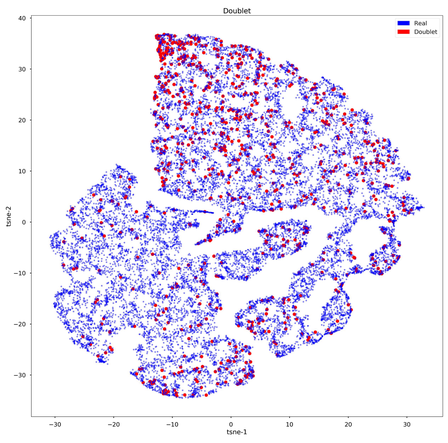
Mission Bio Tapestri Workflow



Deep Learning Architecture

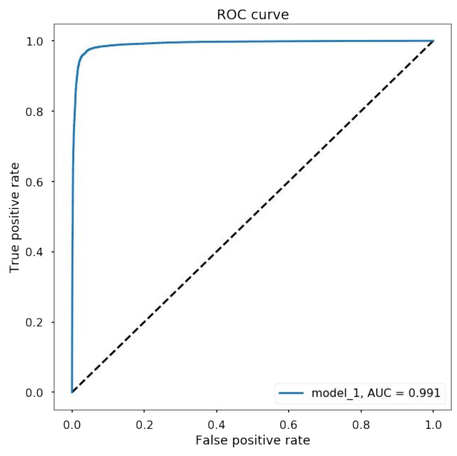


Each cell has multiple amplicon read count data as features. Input layer contains all amplicons. A single hidden layer with 8 nodes were chosen.

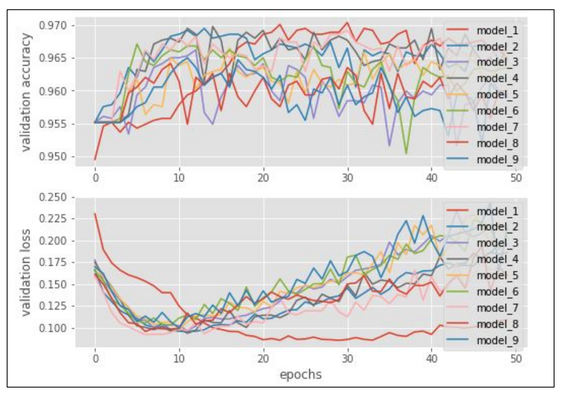


T-SNE plot of amplicon read count. Doublet truth is generated by genotyping. Doublets are not localized to any cluster

Accuracy

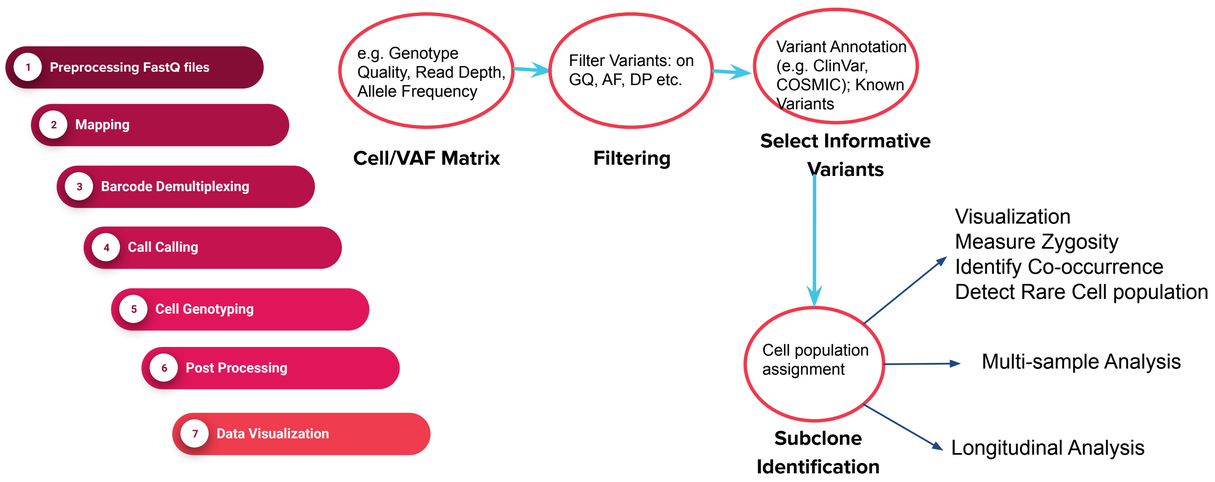


AUC-ROC curve shows the high performance of our model.



Evaluation of different models with increasing number of hidden layers. We evaluated multiple models with increasing number of layers. model_1 has a single layer and model_9 has 9 layers. The accuracy and validation loss data suggests model_1 has better performance

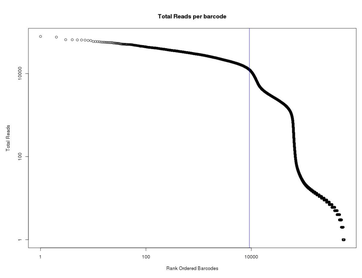
Primary Analysis and Visualization



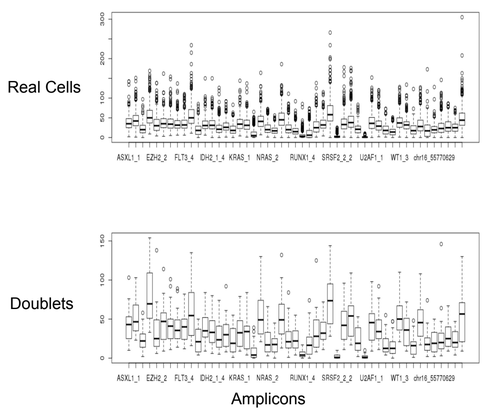
Results and Conclusions

With cross validation we are able to achieve ~99% AUC. Our cell mixing data of ~400k cell show significant performance in terms of accuracy. Further optimization in network architecture and dropout would improve the accuracy. Doublet detection from read count matrix without knowing genotype would significantly improve the throughput of Tapestri System.

Single Cell Cell Finding and Doublets



- Mission Bio Cell Finder
- Doublets still have high reads and hard to remove



Learn more about Mission Bio at our other posters at SCG

Poster	Session	Title
P028	24th	Doublets Detection in Single Cell DNA-Sequencing using Deep Learning
P031	24th	Co-detection of mutations and copy number variations in thousands of single-cells using an automated platform
P034	24th	Using machine learning to optimize assays for single cell targeted DNA sequencing
P093	24th	Single-cell Simultaneous Detection of DNA Genotype and Protein Expression
P099	24th	Error Correction in single-cell DNA sequencing: Finding rare allele for MRD clone
P109	25th	A high throughput single cell workflow for paired genomic and phenotypic analysis
P112	25th	A triomic single-cell high-throughput microfluidic workflow for resolution of genotype-to-phenotype modalities: parallel analysis of DNA, RNA and protein