Poster 861

mission bio

Improvements in variant calling sensitivity and specificity in single-cell DNA sequencing using deep learning

Manimozhi Manivannan[1], Dong Kim[1], Sombeet Sahu[1], Saurabh Gulati[1], Shu Wang[1], Saurabh Parikh[1] and Nigel Beard[1]

[1] Mission Bio, South San Francisco, CA, USA

Conflicts of interest: D.K., M.M., S.S., S.W., S.G., S.P., N.B. are employees and/or shareholders of Mission Bio, Inc.

Abstract

Background

It is now possible to interrogate thousands of cells in a single experiment for studying genetic variability with the advancements in single-cell sequencing technologies. Single-Cell DNA platforms like Tapestri is still susceptible to errors from polymerase incorporations, structure induced template switching, PCR mediated recombination in Tapestri workflow or DNA-damage. Errors from sequencing could propagate from cluster amplification, cycle sequencing or image analysis. All together these errors can be divided into substitutions, insertions and deletion errors and can range from 0.5% to 2% depending on the sequencer. This makes rare variant and minimal residual disease detection challenging. To address these challenges, we developed deep learning models for correcting the errors, reduce false-positive rates and predict true variants.

Methods

First we build a consensus sequence from several reads to predict the correct sequence. The initial layers learn the motifs and local sequence contexts in classifying the patterns. The output of this network is a probability distribution over possible bases and the prediction is the base with highest probability. The bases in the reads are subsequently corrected to the predicted base from the first step model. After error correcting the reads, we used the variants called by Genome Analysis Toolkit to feed into a multi-class classifier network. Our features consists of percent of cells mutated, and the different genotype features including depth, AF and quality of each variant in these cells. The truth labels are generated using tapestri instrument from multiple experiments with known bulk truth. We trained the network on over 200k cells from 13 samples and tested on a larger set of samples. Class imbalance was handled using upsampling the truth data. Our training samples include diverse samples from cell mixtures at various dilution uptill 0.1% and clinical samples processed through tapestri instrument and sequenced on a diverse set of sequencers including miseq and novaseq.

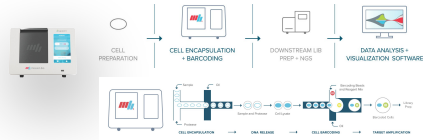
To validate this method, we used two different targeted panels on a Latin square model system with known truth mutations. With our 2 step workflow using error correction and variant prediction model, we significantly improved our median PPV 2-3 fold at 0.5% LOD.

Why Single-Cell ?

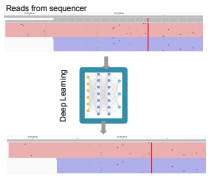
Complexity of Cancer Evolution

Understand Disease at a Single-Cell Level

Mission Bio Tapestri Workflow Overview



Error correction workflow using DNNs

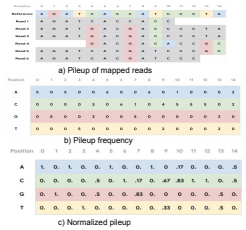


Accuracy

	# of examples	accuracy (%)
Training set1	1,000	82.0
Training set 2	11,113	98.9

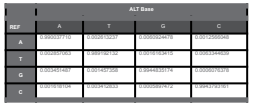
	# of examples	accuracy (%)
Prediction set	29,935	96.5

Preparing input data for CNN

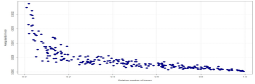


a) For each aligned read in a BAM file, a pileup is generated around a mismatch position with n flanking bases to the left and right. b) Frequency of A,C,G, and T for each location in a window. c) Frequencies of each base are normalized by the total number of bases appearing on a given position.

Generating substitution matrix



a). Substitution rates are calculated by counting number of bases for a given reference base for valid loci.

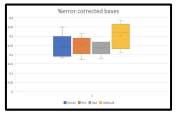


b) Num of bases vs avg deviation in error

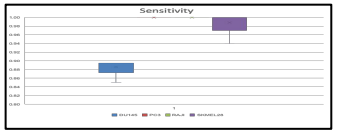
b). We observed significant variation in the substitution rates between runs and hence fixed matrix would not work. Substitution rates starts plateauing after sub sampling 4M reads

Sensitivity and specificity on titration experiments

a)



b)

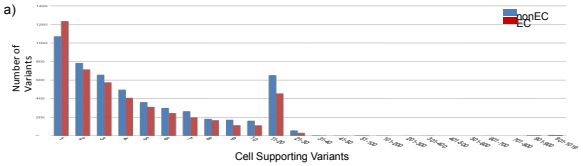


a). 4 different cell lines PC3, RAJI, DU145 and SKMEL28 were mixed at different titrations. Cells were subsampled randomly from each cell line. Error correction was applied to the reads and data from before and after were analyzed for sensitivity and specificity. ~25% reduction in the total number of variants.

b). High sensitivity in most of the variants. There were 3 different FN's that can be improved by further optimization.

Results on PBMC sample with known truth

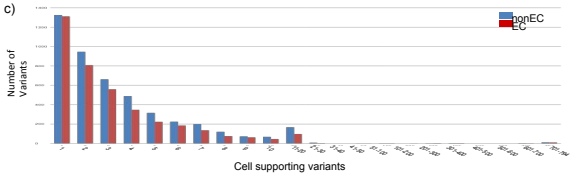
a)



b)

	before	after
variant 1	893	893
variant 2	875	877
variant 3	867	234
variant 4	2	2
variant 5	1	1

c)



d)

	before	after
variant 1	235	234
variant 2	649	651
variant 3	656	186
variant 4	226	226
variant5	19	18

a and c) Two clinical samples were processed through analytical pipeline. Frequency of the variants were counted and compared to before and after error correction. Overall error correction resulted in a decrease in the number of observed variants

b and d) The true variants were known from bulk sequencing. 4/5 of the variants showed same sensitivity before and after. There is one variant with low sensitivity.

Results and Conclusions

To validate this method, we used two different targeted panels on a Latin square model system with PBMC samples with known truth mutations. We also performed titration experiments of 4 cell line mixtures with 98.4%, 1%, 0.5% and 0.1% dilutions. We processed the samples through Tapestri Platform and sequenced over multiple Illumina sequencers (Hiseq 2500, Miseq). We ran the Tapestri analytical workflow with and without error correction. With the error correction pipeline, we reduced our false positive rates by ~25% while maintaining high sensitivity. Further optimization to improve the sensitivity is currently in progress.

Tapestri Solution

See our other posters: #2109, #865, #6581, #5910, #2506, LB-316

Contact: mani@missionbio.com