P099

Error Correction in Single-Cell DNA Sequencing: Finding Rare Allele for MRD Clone

Dong Kim[1], Manimozhi Manivannan[1] , Sombeet Sahu[1], Shu Wang[1], Saurabh Gulati[1], Nianzhen Li[1], Adam Sciambi[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., A.S., N.L., A.P., H.V.,N.B. are employees and shareholders of Mission Bio, Inc.

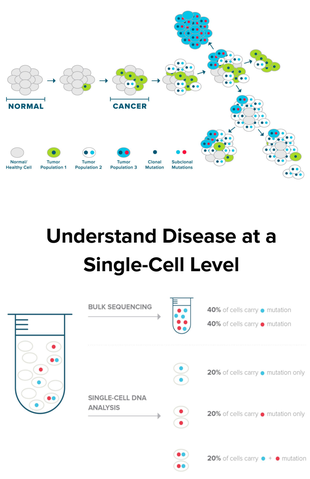
Abstract

Background
The Tapestri single-cell DNA sequencing platform contains a small number of errors from polymerase incorporations, structure induced template switching, PCR mediated recombination or DNA-damage, and errors from sequencing. These errors can range from 0.5% to 2% depending on the sequencer. To improve variant calling and minimal residual disease (MRD) detection we developed a novel consensus sequence-based method for correcting the errors and reducing the false-positive rates.

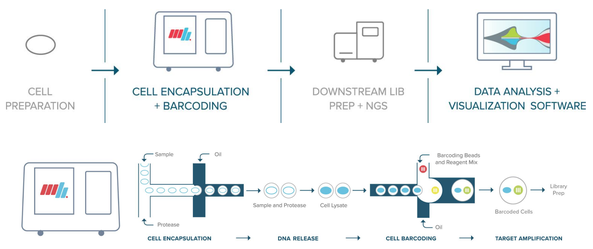
Methods
Here we present a model to correct base position errors in Tapestri single-cell DNA analytical workflow. The error correction method involves 2 steps. First, we train the model with cell BAM files from multiple panels. A pileup is generated around a mismatch position with k-1/2 bases on both sides. During training, a batch of normalized pileups and true reference bases are feed into NN-based model. For each batch, error is calculated between predicted bases and true reference bases. Gradients of this error are calculated with respect to weights of each layer. The gradients are then used to update the weight of the model in a back propagation step. Once a validation accuracy reaches, the trained model and base transition matrix are used to predict a likelihood of a base on given position. If prediction is the same as reference, the model corrects observed non-reference base to reference base. To filter out the noisy reads before passing the data to variant caller, we suppress the quality scores of reads having very low coverage.

To validate this method, we used two different targeted panels on a Latin square model system (4 cell line mixtures with 98.4%, 1%, 0.5% and 0.1% dilutions) with known truth mutations. We ran the Tapestri analytical workflow with and without error correction. With the error correction pipeline, we were able to significantly reduce our false positive rates while maintaining our sensitivity.

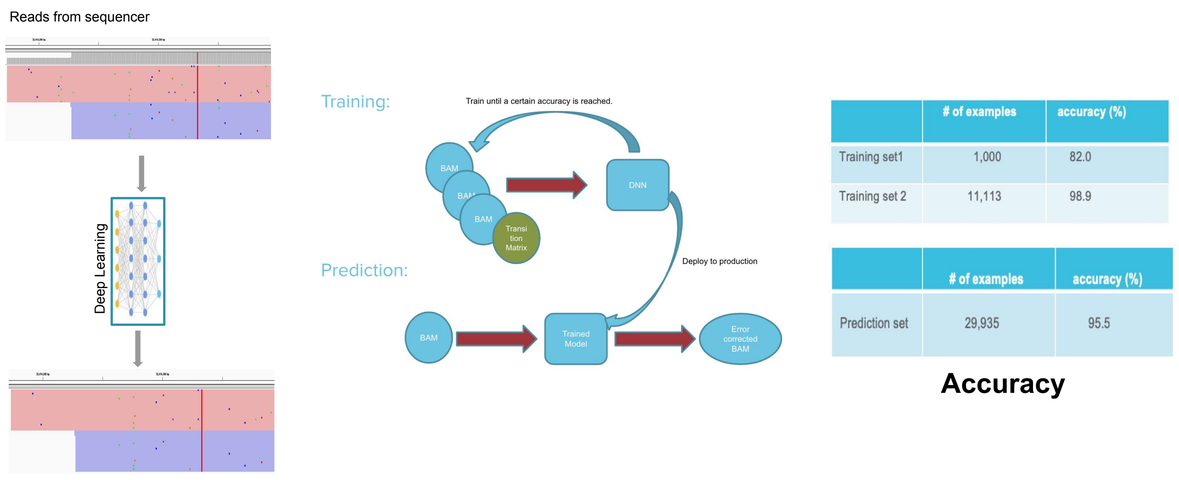
Why Single-Cell ?



Mission Bio Tapestri Workflow Overview



Error correction workflow using DNNs

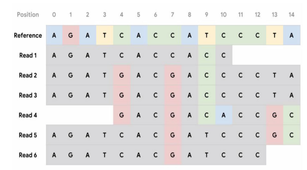


	# of examples	accuracy (%)
Training set1	1,000	82.0
Training set 2	11,113	98.9

	# of examples	accuracy (%)
Prediction set	29,935	95.5

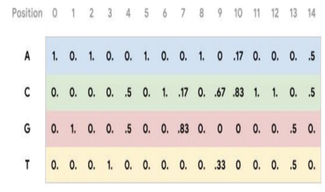
Accuracy

Preparing input data for CNN



a) Pileup of mapped reads

b) Pileup frequency



c) Normalized pileup

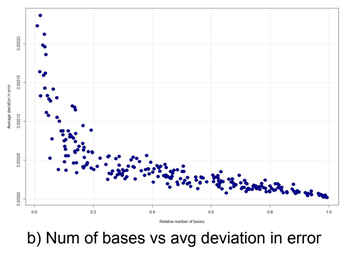
a) For each aligned read in a BAM file, a pileup is generated around a mismatch position with n flanking bases to the left and right. b) Frequency of A,C,G, and T for each location in a window. c) Frequencies of each base are normalized by the total number of bases appearing on a given position.

Generating substitution matrix

REF	ALT Base A	T	G	C
A	0.990037710	0.002613237	0.006092478	0.001256648
T	0.002857063	0.989192132	0.001616415	0.006334639
G	0.003451487	0.001457358	0.994483174	0.000607 6378
C	0.001618104	0.003412833	0.000589 7472	0.994379 3161

a) Substitution matrix

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

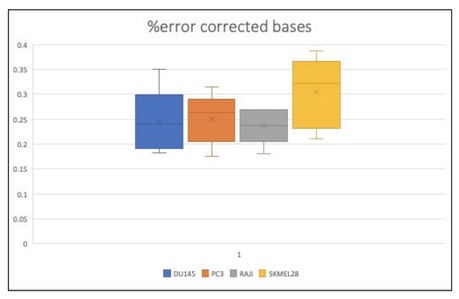


b) Num of bases vs avg deviation in error

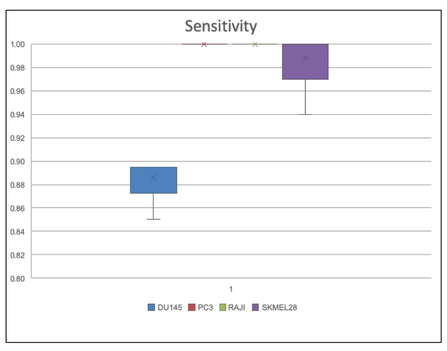
b). We observed significant variation in the substitution rates between runs and hence fixed matrix would not work. Substitution rates starts plateauing after sub sampling 4M reads

Sensitivity and specificity on titration experiment

a)



b)

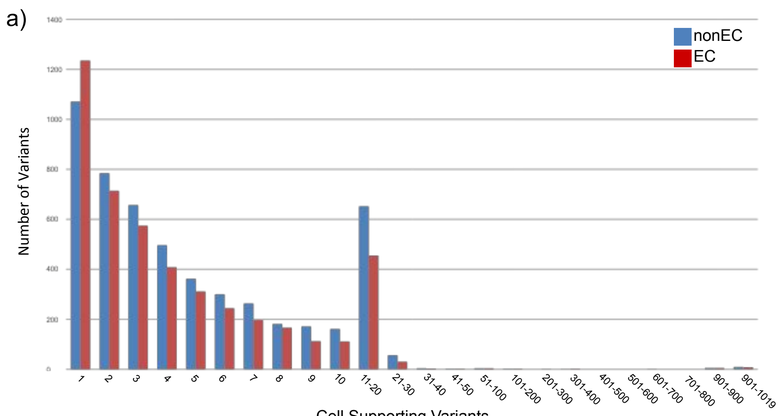


a). 4 different cell lines PC3, RAJI, DU145 and SKMEL28 were mixed at different titrations. Cells were subsampled randomly from each cell line. Error correction was applied to the reads and data from before and after were analyzed for sensitivity and specificity. ~25% reduction in the total number of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

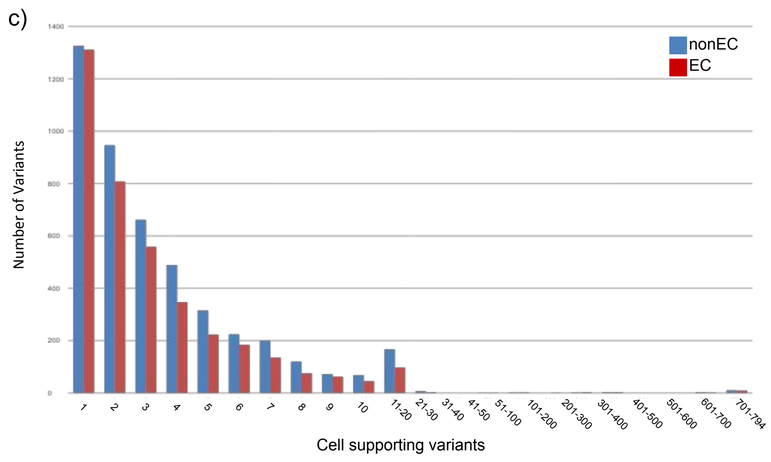
a)



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)

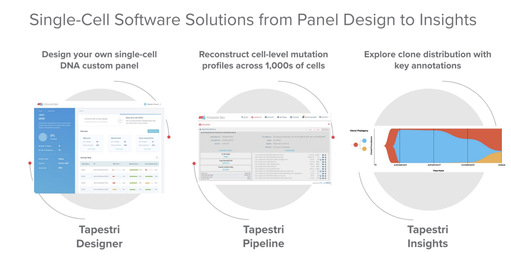
	before	after
variant 1	235	234
variant 2	649	651
variant 3	656	186
variant 4	228	226
variant5	19	16

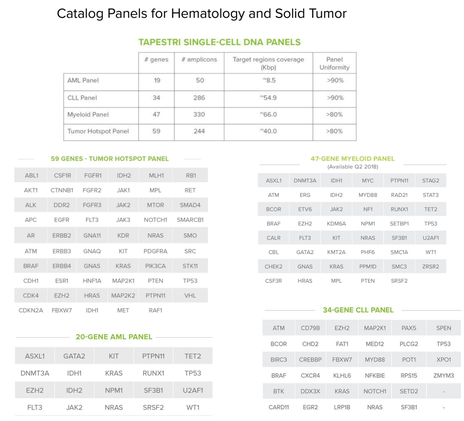
a and c) Two clinical samples were processed through analytical pipeline. Frequency of the variants were counted and compared to before and after error correction. Overall error correction resulted in a decrease in the number of observed variants
b and d) The true variants were known from bulk sequencing. 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





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