Abstract #923

Methods to Detect Large Indels and Tandem Duplication in Acute Myeloid Leukemia Using Single-Cell DNA sequencing

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

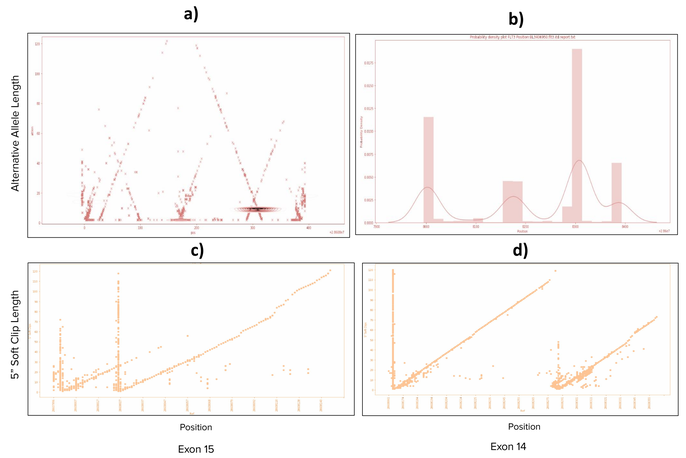
Abstract

Background: FMS-like tyrosine kinase 3 receptor-internal tandem duplication (FLT3-ITD) commonly occurs in one-quarter of patients with acute myeloid leukemia (AML). AML has a poor prognosis, mainly due to relapse. Single-cell DNA sequencing technologies such as Mission Bio Tapestri Platform enables us to understand the clonal heterogeneity of AML patient samples. Large indel calling is prone to errors from library preparation, sequencing biases, and algorithm artifacts. These errors contribute to false positives often in the form of multiple representations of the same variant. Here we present an improved algorithm to identify these large indels and reduce false positives to accurately measure the clonal heterogeneity and enable precision diagnostics.

Methods: The Tapestri Pipeline analytical workflow involves obtaining raw reads from the sequencer, removing adapters, aligning and mapping the reads, calling individual cells and identifying genetic variants within each cell.

We use a soft-clip based approach to detect the internal tandem duplications found in the FLT3 gene. The targeted panel has two amplicons targeting exons 14 and 15 in the FLT3 gene. The soft-clipped reads from these 2 amplicons are scanned for possible insertion events. We then estimate a soft-clip confidence score by mapping the soft-clips again using a threshold. We error correct all the soft-clips which are of low score. The observed insertion event is qualified as an ITD variant if the total number of reads at the loci is greater than 10 and at least 20% of the reads support the insertion. The ITD variant is called homozygous if the allele frequency is greater than 0.9 and heterozygous otherwise. On the 2D space of insert length and position we then apply a generalized median string in Levenshtein space to collapse the different indel variants. The generalized median string is defined as a string that has the smallest sum of distances to the elements of a given set of strings. To do this, we first identify the candidate ITD size bins from the frequency peaks of all the called ITD variants and group the individual variants that are within 20bp boundaries of the frequency peaks into their respective bins. We project the ITD sequence strings within a bin on to Levenshtein vector space domain and calculate the median distance between all strings. We then use the string with the median distance to collapse the ITDs to the consensus sequence and report it in the vcf file.

Sensitivity of Current Algorithm

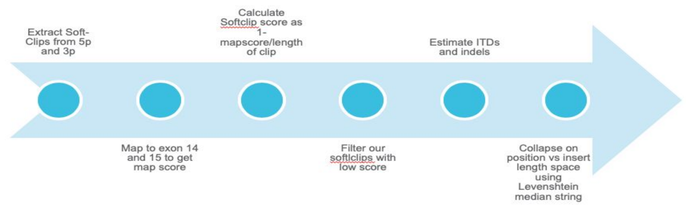


a) For a sample we show here 2 exons, exon 14 and exon 15. For each exon we show the length of alternative allele vs position on the genome. Density is overlaid.

b) We only show the probability density of each cell
c) We calculate the length of soft clip reads for exon 15
d) Soft clip length for exon 14

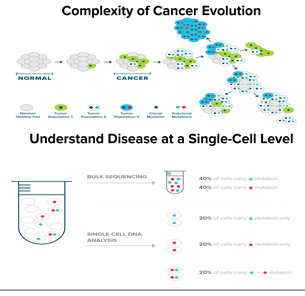
This shows that the current algorithm needs improvement as we would see many false positive FLT3-ITD clones

Improved Algorithm



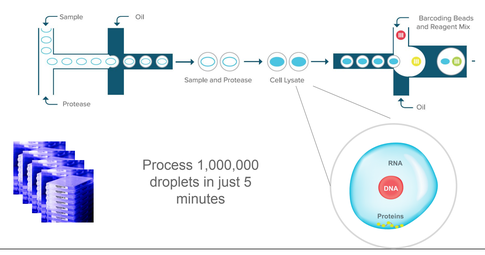
Improved algorithm uses error correction based on a soft-clip score. The soft-clip score is estimated as an independent score from BWA which will penalize a mismatch.

Why Single-Cell ?

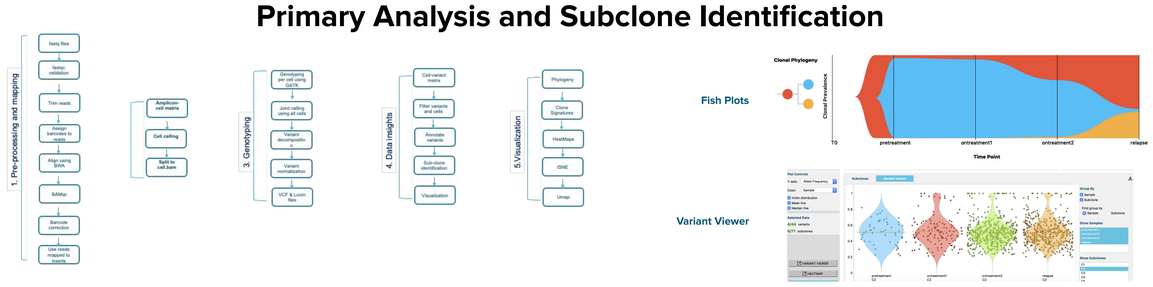


Mission Bio Tapestri Workflow

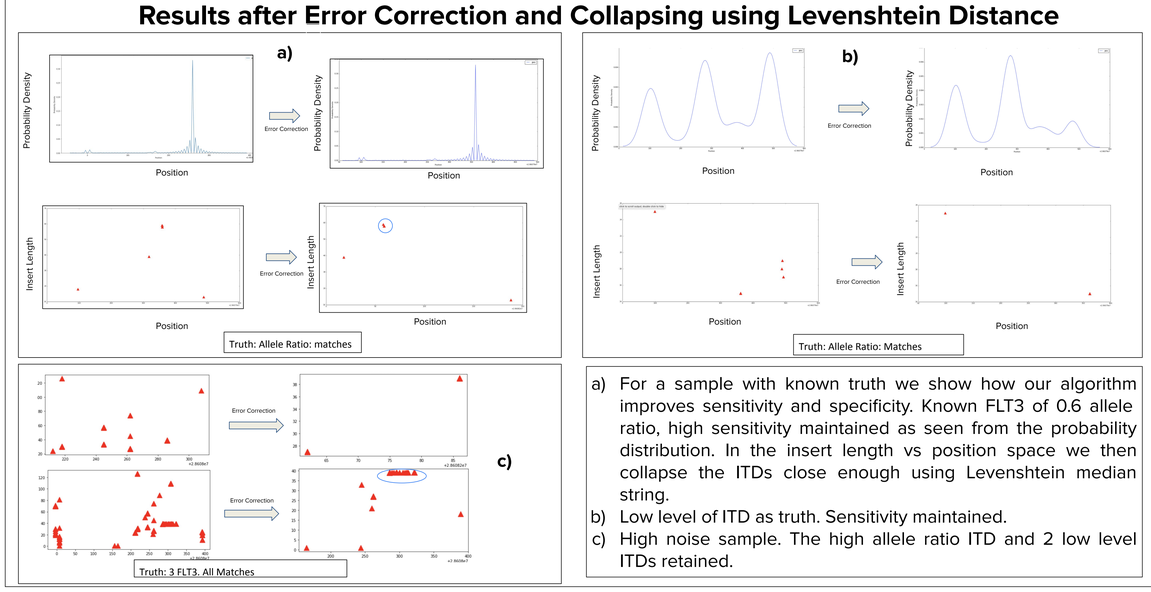
Access to DNA, RNA, and protein from the same cell



Primary Analysis and Subclone Identification

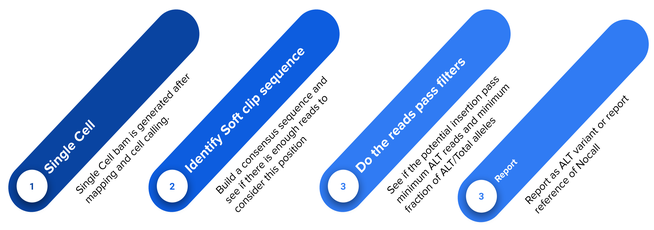


Results after Error Correction and Collapsing using Levenshtein Distance



a) For a sample with known truth we show how our algorithm improves sensitivity and specificity. Known FLT3 of 0.6 allele ratio, high sensitivity maintained as seen from the probability distribution. In the insert length vs position space we then collapse the ITDs close enough using Levenshtein median string.
b) Low level of ITD as truth. Sensitivity maintained.
c) High noise sample. The high allele ratio ITD and 2 low level ITDs retained.

FLT3 ITD Detection from sc-DNA data



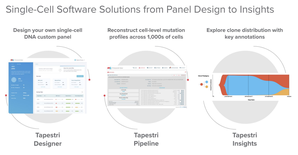
Each cell is scanned for soft-clips and insertions; all insertions and clippings are considered as possible insertions. If the total number of reads is greater than a cutoff (10), and the number and the ratio of non-REF reads are greater than a cutoff (4 and 0.1 respectively), the cell is considered to have a non-REF allele. If the ratio of non-REF reads to REF reads is greater than a cutoff (0.9), a homozygous event is called; otherwise it is considered a heterozygous event. If the cell has enough total reads but not enough ALT reads, it is considered a homozygous reference. Otherwise, it is reported as “no call.”

Results and Conclusions

- We processed AML samples with known FLT3 ITDs through Tapestri platform. We analyzed the raw data via Tapestri analytical workflow including our large indel and ITD detection algorithm with error correction. Using this method, we were able to accurately identify the ITDs and reproduce the true positive clones
- High Sensitivity of previous method maintained.
- Error correction removed the false positive clones and significantly improves specificity.
- Low level LOD truth also maintained

Tapestri Solution and Future Work

Single-Cell Software Solutions from Panel Design to Insights



Catalog Panels for Hematology and Solid Tumor

TAPESTRI SINGLE-CELL DNA PANELS

	# genes	# amplicons	Target regions coverage (Kbp)	Panel Uniformity
AML Panel	19	50	~8.5	>90%
CLL Panel	34	286	~54.9	>90%
Myeloid Panel	47	330	~66.0	>80%
Tumor Hotspot Panel	59	244	~40.0	>80%

Learn more about Mission Bio at our other posters at ISMB

Poster	Session	Title
#923	Session A	Methods to detect large indels and tandem duplication in acute myeloid leukemia using single cell DNA sequencing
#940	Session B	Detecting doublets in Single Cell DNA-Sequencing using Deep Learning
#956	Session B	Error Correction in single-cell DNA sequencing: Finding that rare allele for MRD clone
#982	Session B	Using machine learning to optimize assays for single cell targeted DNA sequencing
#1167	Session B	Analytical Methods to Identify Tumor Heterogeneity and Rare Subclones in Single Cell DNA Sequencing Data from Targeted Panels