mission bio

tapestri
WHITE PAPER

Performance of the Tapestri® Platform for Single-Cell Targeted DNA Sequencing

This white paper describes the Tapestri Single-Cell DNA analysis workflow and reviews different performance metrics of the Tapestri Platform including variant sensitivity, limit of detection, mixed cells and genomic coverage.

Introduction

An average read-out from conventional bulk sequencing misses the underlying genetic diversity across cell populations. The Tapestri Platform was developed to enable the accelerated and accessible detection of genomic variability within and across cell populations. The Tapestri Platform delivers targeted solutions for high-impact application areas, including blood cancers, solid tumors, and genome editing validation. This novel approach to single-cell DNA analysis paired with targeted gene panels offers a powerful strategy for detecting rare subclones, resolving mutational co-occurrence patterns, zygosity and reconstructing phylogenetic lineages. For single cell genomics to eventually inform clinical decisions and impact personalized medicine it must be sensitive, accurate and capable of detecting low abundance clones.

Materials & Methods

The Tapestri Platform provides a targeted, automated and scalable approach to profile single nucleotide variants (SNVs) and indel mutations across thousands of cells at the single cell level. First, thousands of cells are microfluidically encapsulated and lysed in droplets and subsequently protease-treated to liberate DNA from histones and other DNA-binding proteins. In the next step, individual cell lysates are uniquely barcoded and target-specific genes/regions are simultaneously PCR-amplified inside each droplet **(Figure 1)**. Amplified products are pooled and prepared with conventional sequencing library chemistry, sequenced on an Illumina Sequencing instrument and the data is processed and analyzed with fully integrated Mission Bio software. The Tapestri Pipeline tool facilitates read alignment and variant calling whereas Tapestri Insights tool enables data analysis, variant filtering and data visualization.

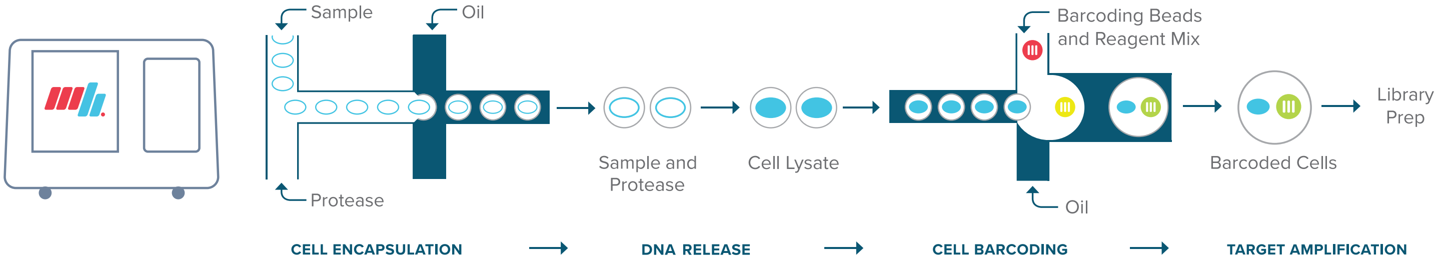


Figure 1. Schematic overview of the Tapestri workflow

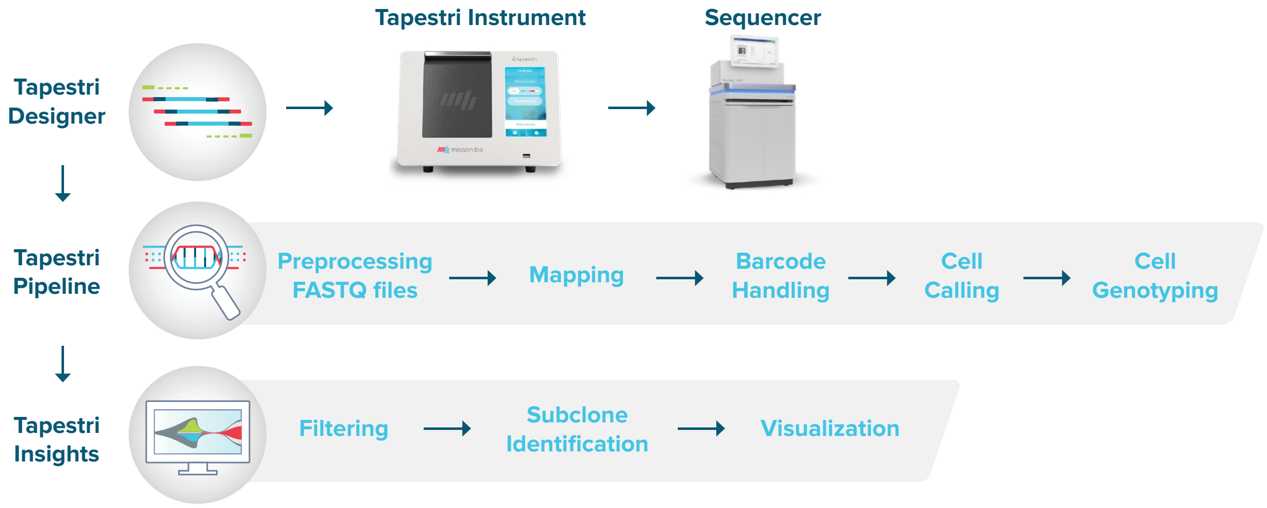


Figure 2. Overview of the data workflow

Preprocessing & Mapping

Fastq files generated by the sequencer are processed using Tapestri Pipeline **(Figure 2)**. First, adapter sequences are trimmed from the sequenced reads using Cutadapt (Martin 2011, Bolger, Lohse et al. 2014), following which the Tapestri barcode structures are extracted from the reads. The reads are then mapped to the genome using the BWA-MEM algorithm (Langmead, Trapnell et al. 2009) (Kim, Pertea et al. 2013). The extracted barcodes on the mapped reads are error-corrected against a whitelist of known barcodes using a hamming distance approach. Reads lacking an insert sequence between gene-specific primers or mapped to off-target loci are discarded.

Barcoding Handling

Tapestri barcodes are two 1536 9-bp barcodes which are attached to beads **(Figure 3)**. These barcodes are extracted from the mapped reads. Barcodes that are likely to be unchanged are first identified and the remaining barcodes are then error corrected. A whitelist based approach is used to first look for the barcodes that are exact matches. Then from the discarded barcodes a levenshtein distance approach is used to correct them. A maximum Levenshtein distance of 3 is used since the barcodes are more than 3 Levenshtein distance apart.

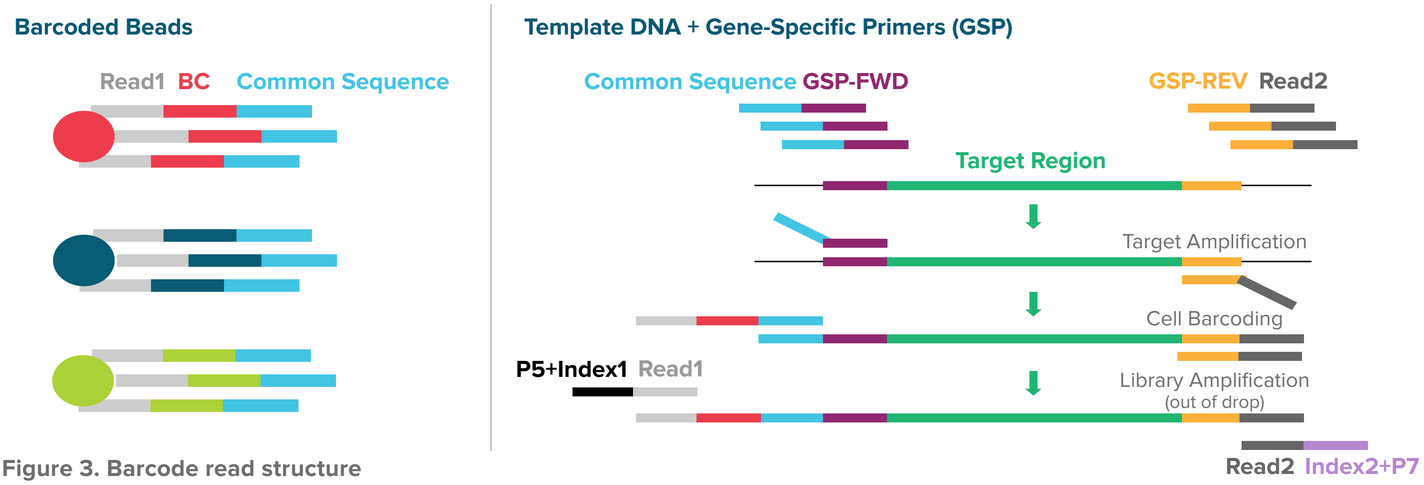


Figure 3. Barcode read structure

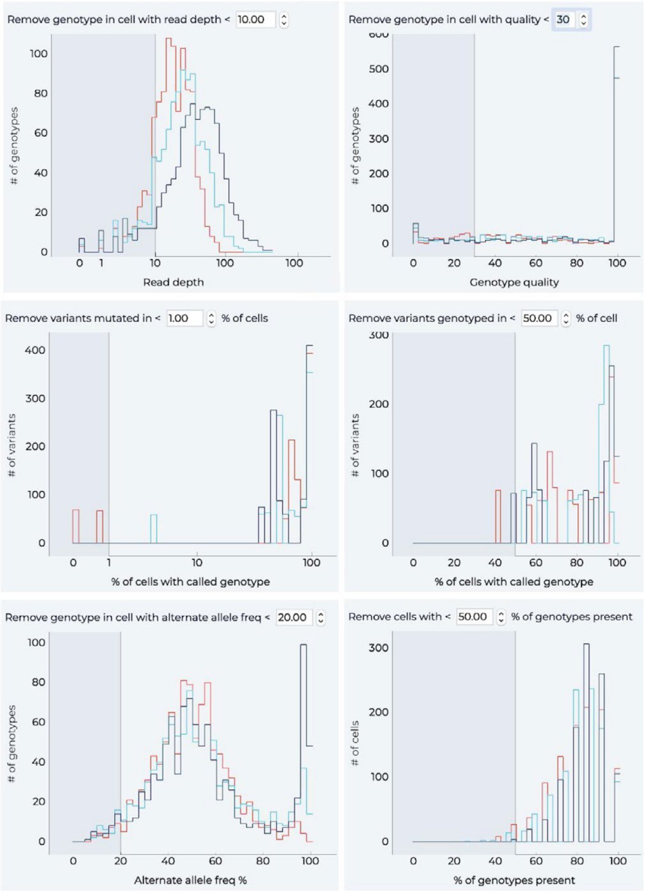
Cell Calling

Based on the Tapestri Single-Cell DNA AML Panel, a threshold of 10*number of amplicons is calculated. From the barcode/amplicon matrix, barcodes which have total reads exceeding this threshold are identified. From this subset matrix the 0.2X of the mean of all amplicons reads for all cells is measured. If this 0.2X value is less than 10 then the threshold is 0.2X mean value if not, its kept at 10. The number of amplicons that are less than this threshold are identified, these amplicons are ignored and the barcodes which have at least 80% amplicons having reads more than the threshold value are chosen. That is the final number of cells.

Variant Calling & Genotyping

The cells are genotyped with the Genome Analysis Toolkit (McKenna, Hanna et al. 2010) using a joint calling approach that follows GATK Best Practices recommendations (DePristo, Banks et al. 2011, Van der Auwera, Carneiro et al. 2013). Each cell is haplotyped in reference confidence mode to enable per-base pair (bp) confidence estimates for a site's being strictly homozygous (reference). The per-bp resolution is maintained while merging the genomic-VCFs (gVCFs) for all cells using GATK's CombineGVCFs tool. Finally, joint genotyping is performed for all cells using GATK's GenotypeGVCFs tool. Loci found to be nonvariant are maintained in the final output. Genotyping parameters are optimized for high sensitivity: a maximum of 2 alternate alleles are reported for each site, the minimum base quality for variant calling is set at 10, and the heterozygosity value is set at 0.001.

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.” Multiallelic variants are decomposed into biallelic variants and then normalized to ensure that each VCF entry is left-aligned and parsimonious (Tan, Abecasis et al. 2015). Blacklisted loci (error prone regions of the genome) are filtered out, and all loci less than 1000 QUAL threshold are tagged for downstream processing. The positions that pass the filtering criteria are called variants. The genotypes and the cell matrix are converted into an open-source loom format for each sample (Zeisel, Hochgerner et al. 2018), which allows efficient storage, data retrieval, and sharing of large omics data sets.



Remove genotype in cell with read depth < 10.00

Remove genotype in cell with quality < 30

Remove variants mutated in < 1.00 % of cells

Remove variants genotyped in < 50.00 % of cell

Remove genotype in cell with alternate allele freq < 20.00

Remove cells with < 50.00 % of genotypes present

Figure 4: The six filters used in variant filtering with default values used

Filtering & Subclone Identification

A series of six quality filtering steps **(Figure 4)** are applied to the cell/genotype matrix to select for high quality data. These filters affect different parameters such as variant quality score, read depth per variant per cell, and limit of detection. Based on selected variants, cell counts corresponding to the selected variants are identified and the genotypes are mapped to identify groups of cells that can be put together as coming from a single clone. The number of clones can vary depending on parameter selections during filtering.

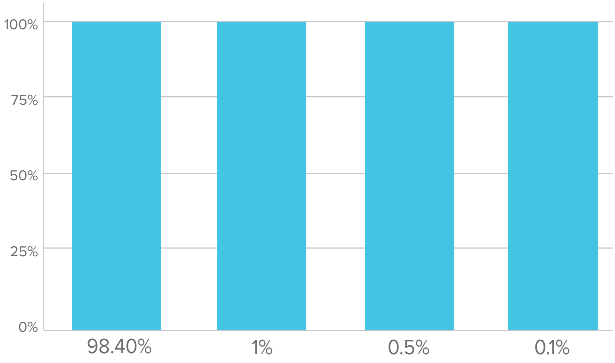


Figure 5: Platform sensitivity across various spike-in percentages

Sensitivity

Sensitivity of the Tapestri Platform was assessed using four different sample types: HCT-15 cells, human prostate cancer cell lines PC-3 and DU-145 and a human melanoma cell line SKMEL-28. Sequencing libraries were processed following standard protocol and sequenced on Illumina NextSeq Sequencing platform. Variants were called using GATK (v3.7) and only cells of high quality were included in the analysis. In order to test the sensitivity, a latin square dilution set was generated in which input cells (PC-3, DU-145, SKMEL28 and HCT15) were combined with each other at ratios of 98.4%, 1%, 0.5% and 0.1% in one sample, 1%, 0.5%, 0.1% and 98.4% in a second, 0.5%, 0.1%, 98.4% and 1% in a third, and 0.1%, 98.4%, 1% and 0.5% in a fourth sample, respectively. Using a cell assignment algorithm, each called cell was assigned to a cell line using a mutation signature. For each cell cluster, a pre-qualified locus was scored as positive if it was present in at least 90% of the cells in the cluster. The sensitivity rate for that cluster was calculated as the total number of pre-qualified loci divided by the sum of the total number of pre-qualified loci and number of negative loci. Sensitivity across the spike-in experiments at 98.4%, 1%, 0.5%, and 0.1% was determined to be 100%. **(Figure 5).**

Limit of Dectection

Limit of detection (LOD) was assessed by spiking three different percentages of K562 cells in a RAJI cell background: 1.0%, 0.5% and 0.1% (n=3 for each percentage). All nine cell suspensions were run on the Tapestri Platform following standard protocol and sequenced on Illumina instruments. A select number

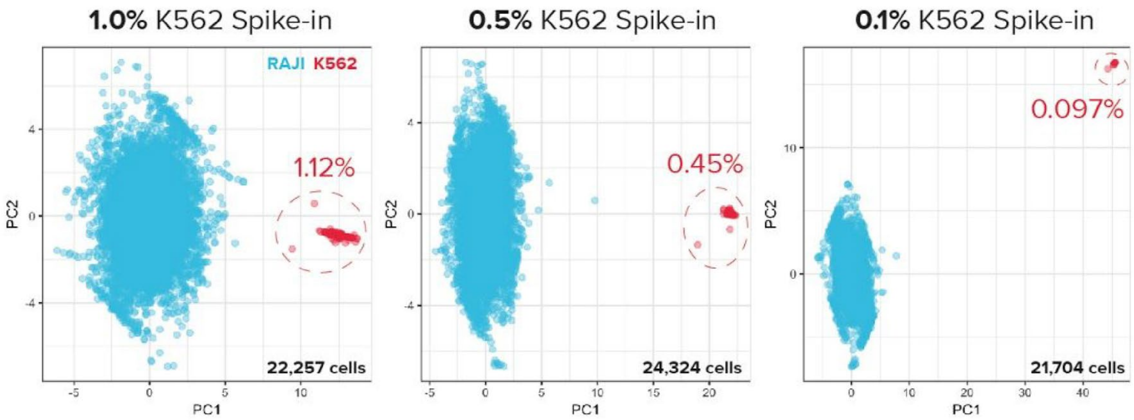


Figure 6: Principal component analysis with cells projected onto the first two PCs. Cells were clustered using k-means and color-coded accordingly.

of single nucleotide variants (SNVs) that are different between both cell lines were used to distinguish RAJI from K562 cells and to calculate the recoverable spike-in percentages **(Figure 6).**

Mixed Cells Detection

Mixed cells are primarily the result of (1) two or more cells encapsulated in one droplet, (2) two or more single cell-containing droplets merging during sample processing or a combination of (1) and (2). The percentage of mixed cells was measured with mixed-genotype experiments in which two different cell lines were combined at a 1:1 ratio. Cells with a heterozygous alternate genotype call were classified as mixed cells as this genotype does not naturally exist at that loci in either cell line **(Figure 7).**

The mixed-genotype experiments suggested a mixed-cell rate of 4.3 % at 10,000 cells with an inferred overall mixed-cell rate of 8.6 % (two times the measured mixing rate to account for K562/K562 and RAJI/RAJI mixing).

ADO Rate

The allele drop-out (ADO) rate for a given sample is calculated using germline heterozygous SNVs, avoiding genomic regions potentially targeted by copy number variations (CNVs), and meetinging the following criteria: (a) Only cells with read depth per variant > 10 are considered. (b) Variants must be genotyped in at least 75% of cells. (c) ADO calculation is carried out only if more than 3 SNPs pass all the thresholds. With the above considerations met, ADO rate is then reported as [100 - (% of homozygous cells - % of wildtype cells)]%. The global ADO rate for a given sample is calculated by averaging all the SNP-specific ADO rates.

Sequencing Coverage Requirements

The recommended minimum coverage for sequencing on the Tapestri Platform is 60-80X. For example, the AML panel that has 125 amplicons would need 7,500 cells * 125 amplicons * 80x coverage or ~75M reads

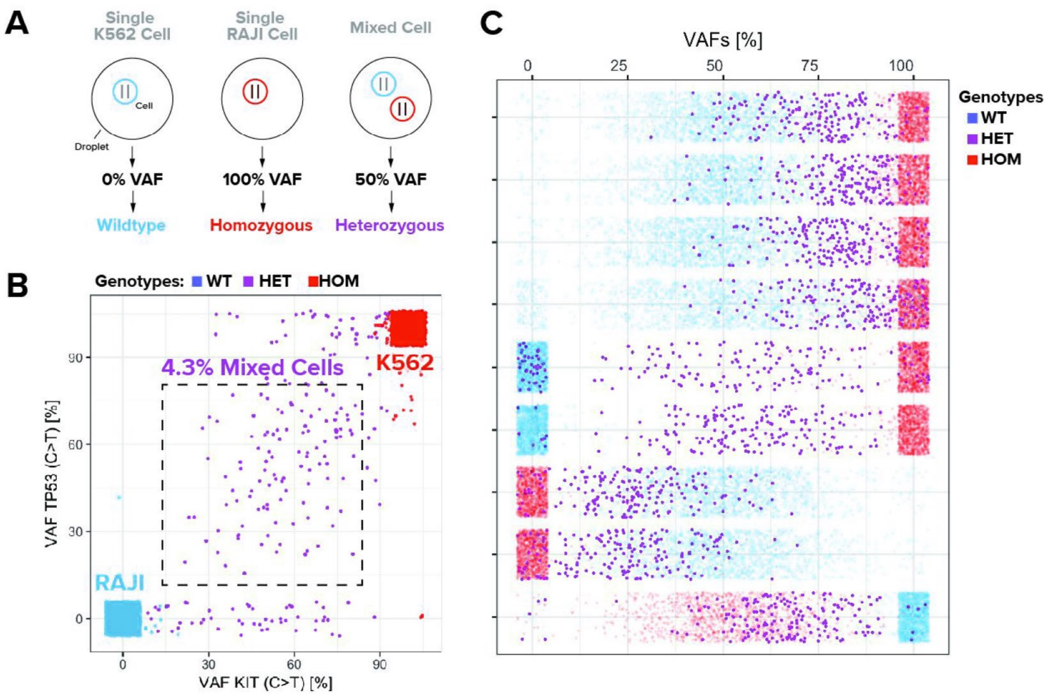


Figure 7: Assessment of mixed cells occurrence on the Tapestri Platform (a) Schematic overview of three scenarios describing the concept of mixed-genotype experiments (b) Scatter plots of all cells across variant allele frequencies (VAFs) of two variants that are naturally occurring homozygous reference (RAJI) or homozygous alternate (K562). Cells in the boxed area putatively represent mixed cells. Cells are color-coded according to the called genotype (blue=WT, purple=HET, red=HOM). (c) VAF distributions at a select number of loci. For loci that are WT/HOM between RAJI and K562 cells the VAFs of mixed cells are distributed at ~50%. For WT/HET loci mixed cell VAFs are distributed at ~25% and for HET/HOM loci mixed cell VAFs are distributed at ~75%.

per sample. **Figure 8** provides an overview of the mapping between percentage of sequencing reads that map to the genome, to targets and to cells on Illumina sequencers for the AML panel. On average the platform maps to ~75% of reads from sequencing to cells. Number of read pairs needed can be estimated by the product of (number of cells * number of amplicons * 80x).

Data Completeness

Three Tapestri Single-Cell panels were run on the Tapestri Platform namely, the AML Panel, the CLL Panel and the Myeloid Panel. The amplicons achieving a minimum depth of 10 reads were taken and compared with the average number of reads per cell per amplicon to examine data completeness. The Tapestri Platform on average achieves > 90% complete data **(Figure 9)**

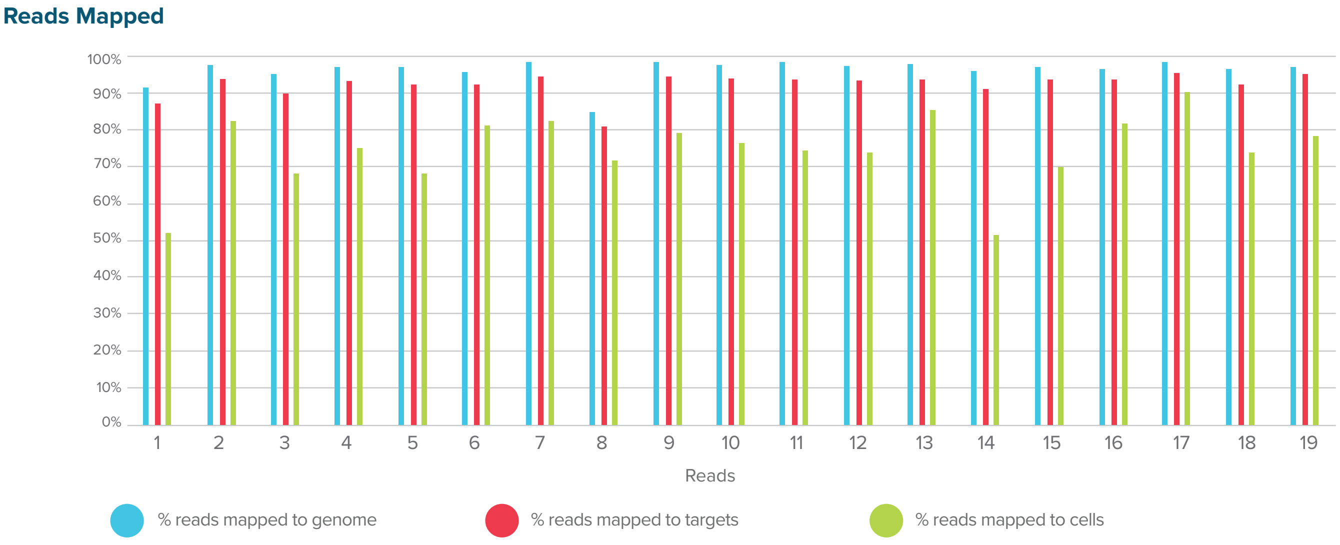


Figure 8: Tapestri platform maps on average 75% of reads from sequencing to cells. Number of read pairs needed can be estimated by the product of (number of cells * number of amplicons * 80x).

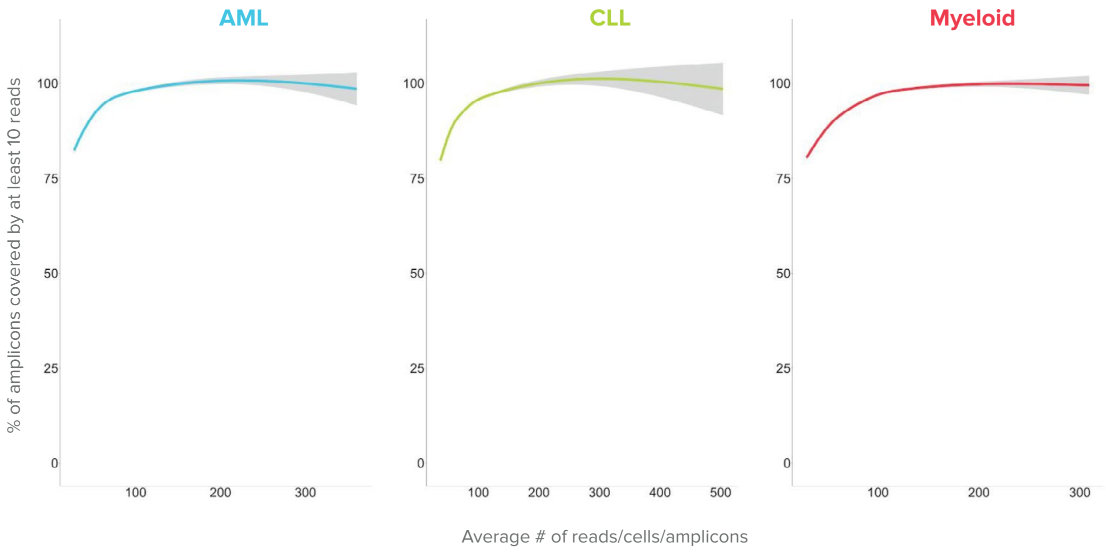


Figure 9: Read coverage for three catalog panels namely, AML, CLL and Myeloid Panel.

Cell Throughput

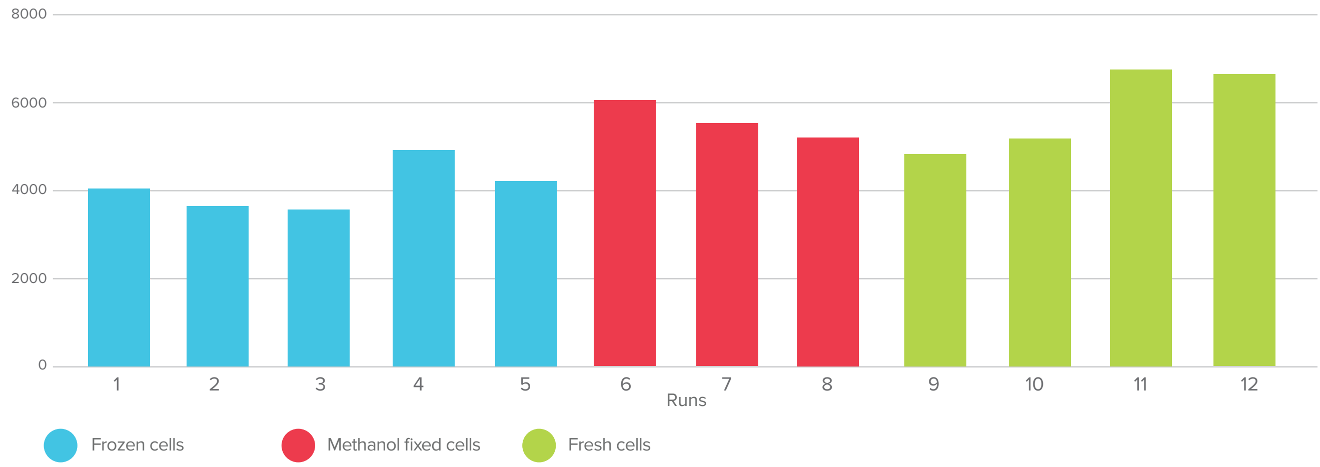


Figure 10: Tapestri Platform cell throughput measured over 12 runs.

Cell Throughout

Platform performance was measured in terms of cell throughput per run. 12 AML panel runs were performed on Illumina instruments with RAJI, BM and PBMC cell lines. Three different cell types were used namely, Fresh (4), Frozen (5) and Methanol fixed (3). Across the 12 runs the Tapestri Platform achieves on average 5056 cells in throughput **(Figure 10)**. *(Please note that sample output varies by sample quality.)*

Panel Uniformity

Panel uniformity was measured across 10 runs with the AML panel (3), the CLL panel (1), the Myeloid panel (2) and Custom Ordered Panels (4) with data coming from different cell types - 1 from fresh cell line, 6 from frozen cell lines and 3 from methanol fixed cell lines. Across the panel and sample types, the Tapestri platform maintain panel uniformity of 96% on average **(Figure 11).**

Panel Uniformity

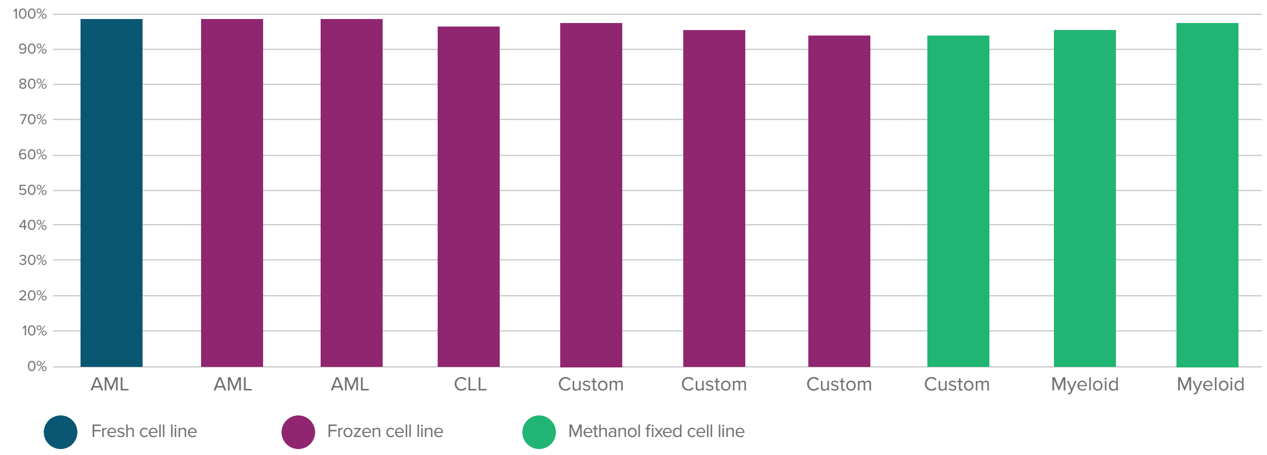


Figure 11: Panel uniformity is greater than 96% on average across 10 runs on the Tapestri Platform.

Conclusion

Tapestri is the industry's first single-cell DNA sequencing platform, enabling precise detection of heterogeneity in disease progression and treatment response. The platform provides unmatched single-cell DNA sequencing sensitivity that provides reproducible and reliable cell throughput. This sensitivity can provide deep resolution of zygosity and co-occurrence of mutations in the same clones. The versatility of the platform allows for usage in applications ranging from examining cell state and lineage tracing in cancers and metabolic disorders to examining genome editing success rates.

QUESTIONS?

missionbio.com
info@missionbio.com

6000 Shoreline Court, Suite 104, South San Francisco, CA 94080 USA
+1 (415) 854-0058

For research use only. Not for use in diagnostic procedures. Mission Bio, Inc. makes no representation or warranty as to the usefulness, completeness, or accuracy of the information contained in this technical note. The products, services, and specifications set forth in this brochure are subject to change without notice and Mission Bio, Inc. disclaims any and all liability for such changes. The information contained herein is provided without warranties of any kind, either express or implied, and Mission Bio, Inc. disclaims any and all liability for typographical, printing, or production errors or changes affecting the products, services, and/or the specifications contained herein. It is the responsibility of the customer to thoroughly analyze all aspects of the customer's proposed application for the products and services. MISSION BIO, INC. DISCLAIMS ALL WARRANTIES, STATUTORY, EXPRESS, IMPLIED OR OTHERWISE, INCLUDING, BUT NOT LIMITED TO, QUALITY, CONDITION, AND SAFETY, AND THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

IN NO EVENT WILL MISSION BIO, INC. BE LIABLE FOR ANY SPECIAL, INCIDENTAL, DIRECT, INDIRECT, PUNITIVE, OR CONSEQUENTIAL DAMAGES OF ANY KIND ARISING OUT OF THE RELIANCE ON THE CONTENTS OF THIS TECHNICAL NOTE OR FROM THE SALE OR USE OF ANY PRODUCT OR SERVICE DESCRIBED HEREIN.