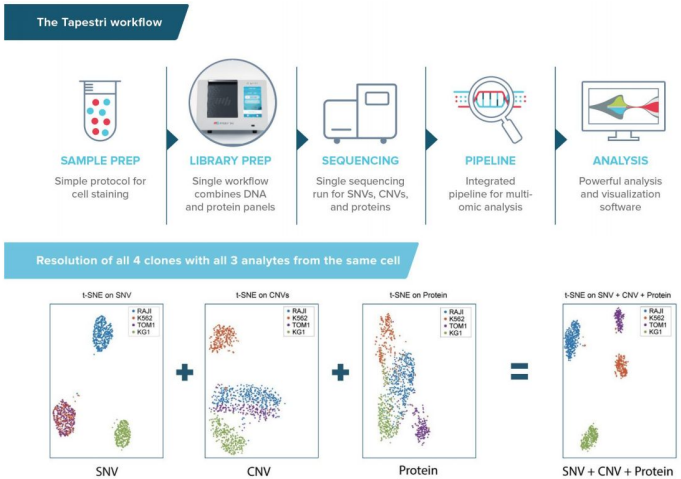
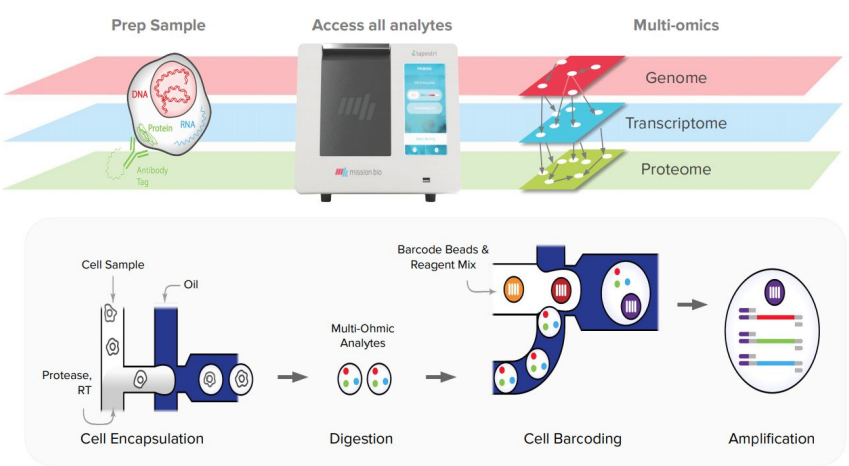
Amplicon design for single cell targeted sequencing assays using machine learning

Shu Wang[1], Saurabh Gulati[1], Saurabh Parikh[1], and Manimozhi Manivannan[1]

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

Resolution Revolution: Access DNA, RNA, and Protein from cells on high throughput single cell platform





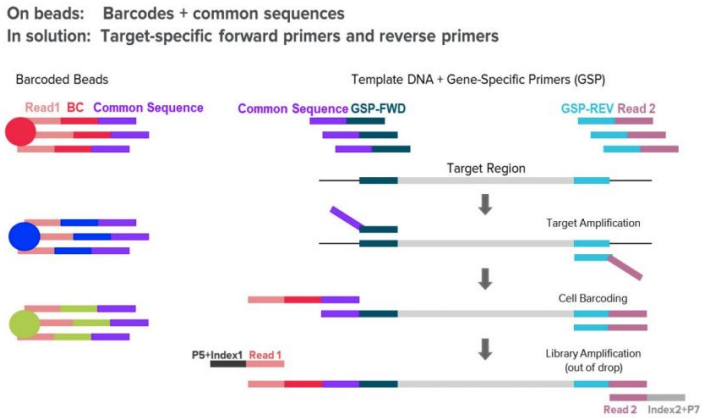
mission bio

Conflicts of interest: S.W., S.G., S.P., M.M. are employees and shareholders of Mission Bio, Inc.

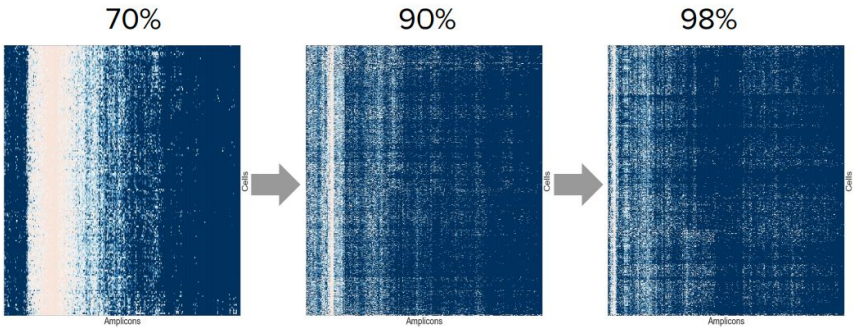
Optimize targeted sequencing panel design using ML approach

High throughput single cell DNA targeted sequencing enables detection of rare mutations in cells and identification of subclones defined by co-occurrence of mutations. The big challenge with multiplex sequencing at single cell level is the non-uniform amplification of the targeted regions during PCR. This results in inadequate coverage of mutations of interest in the panel and hence makes genotyping challenging. To address this challenge, we developed a machine learning engine to optimize amplicon design for uniform amplification by making reliable performance prediction.

Barcode read structure



ML approach to improve quality of panel



ML approaches predicts amplicon performance and improve the panel uniformity for tapestri targeted sequencing assay

Significant improvement of performance in panels for multiple genomes and also of varying panel sizes.

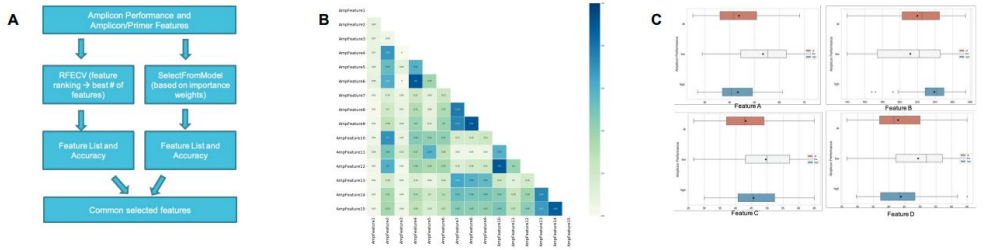
Methods:

Multiple panels with various sizes were designed with amplicons spanning a wide range of design properties such as amplicon GC, length, secondary structure prediction, primer specificity. These panels were synthesized and processed through Tapestri single cell DNA platform. The tested amplicons are classified into low-performer, OK-performer and high flyer based on their normalized reads-per-cell value. Design properties and property distribution of the amplicons and the panel are the features. We used random forest classifier to calculate feature importance and analyzed the range of the top features for each class and their significance of variance between classes. These ranges were then used as parameters in the assay design pipeline. Next, we train machine learning models with performance data to develop a performance prediction engine.

Results:

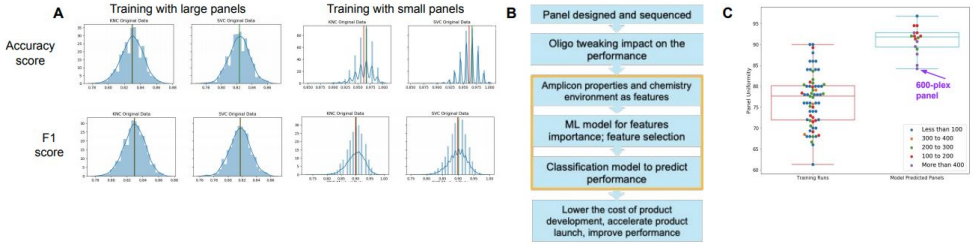
To test the performance of the design pipeline with new parameters, we designed a small (31), medium (128) and large (287) amplicon panel. Multiple runs were conducted for each panel with different cell types. We were able to achieve high panel performance of 97%, 92% and 88% across the three panels. The new parameters resulted in ~10-20% improvement in panel uniformity. We are working on further optimizing the performance prediction engine by using different ML classification models with K-fold cross validation, training using larger group of amplicons and optimizing features using combinations of properties.

Identify important features impacting amplicon performance



A. To identify important amplicon and primer properties that impact amplicon performance in Tapestri targeted sequencing assays, we used random forest classifier to calculate feature importance of properties. Common top features identified using two feature selection methods were selected for carry-on analysis. **B.** Correlation of numeric features identified highly correlated features. Only independent features were kept for feature distribution analysis and building prediction model. **C**. Box plots showing some of the identified predictive features.

KNC and SVC model fit for different data size

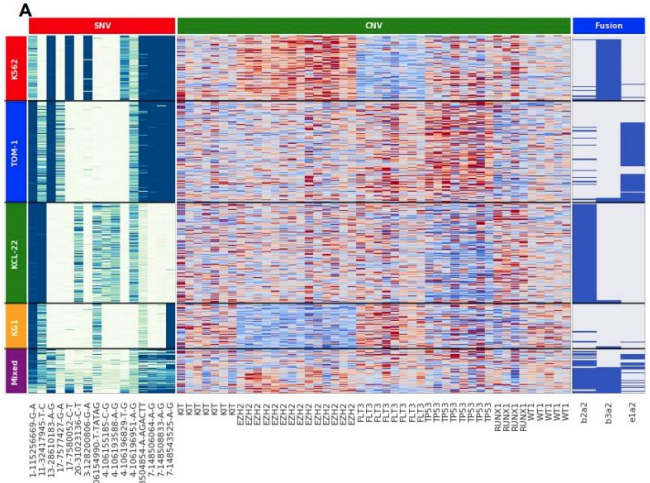


A. Selected amplicon features and performance data for 10 targeted panels were used to train and test performance prediction model. Two ML classification model (KNC and SVC) with K-fold cross validation were trained with 10000 splits of 70/30 for training/testing dataset split, while all splits keep the same ratio of classes in both training and testing datasets. Average accuracy ranges from 0.80-0.88 for large dataset to 0.90-0.98 for small panels. **B.** Schematic showing workflow of using ML model to optimize amplicons performance. **C.** New designer significantly improved amplicon performance and uniformity in targeted assay design across different panel size and genomic contents (human and mouse genomes). 6 newly designed panels were sequenced. Multiple runs were conducted for each panel.

mission bio

Tapestri Designer provides a robust pipeline for targeted panels across multiple analytes

A



B

Fusion	Sensitivity	Specificity
b3a2	96.8%	95.7%
b2a2	93.6%	96.5%
e1a2	70.2%	99.7%

A. A 4 cell line mixture was run on the Tapestri platform with an acute myeloid leukemia (AML) panel and primers to detect 3 BCR-ABL1 fusion transcripts. The data was resolved into 3 modalities of SNVs, CNVs and Fusions. K562 is positive for b3a2, TOM-1 is positive for e1a2 fusion, KCL-22 is positive for b2a2 fusion and KG1 was negative for all 3 fusions. The cells clustered well with the SNV and CNV data, and the fusion data correlated with the clustering. A mixed cell population was observed which shows average of other cell lines in SNV,CNV and fusions. **B.** Using a threshold of 20 reads per cell per fusion transcript for a positive call, we calculated the sensitivity and specificity per fusion transcript across all cells and observed very high specificity for all of them (> 95.7%) with high sensitivity for b3a2 and b2a2 (> 93.6%) and good sensitivity for e1a2 (70.2%) detection.

Conclusion:

- The design pipeline developed with ML model generates panels that have more uniform amplification across amplicons. It shows significant improvement of performance in panels for multiple genomes and also of varying panel sizes.
- Targeted approach of detection results in higher limits of detecting rare cell subpopulations.
- Tapestri Designer provides a robust pipeline for targeted panels across multiple analytes including DNA (SNVs and CNVs) and fusion.
- Demonstrated the capabilities of Tapestri system and analytical pipeline to accurately characterize cell types.