P034

Using Machine Learning to Optimize Assays for Single-Cell Targeted DNA Sequencing

Shu Wang[1], Manimozhi Manivannan[1] , Saurabh Gulati[1], Dong Kim[1], Sombeet Sahu[1], Nianzhen Li[1], Adam Sciambi[1], Niranjan Vissa[1] and Nigel Beard[1]
[1] Mission Bio, South San Francisco, CA, USA
Conflicts of interest: S.W., M.M., S.G., D.K., S.S., N.L., A.S., N.V. and N.B. are employees and shareholders of Mission Bio, Inc.

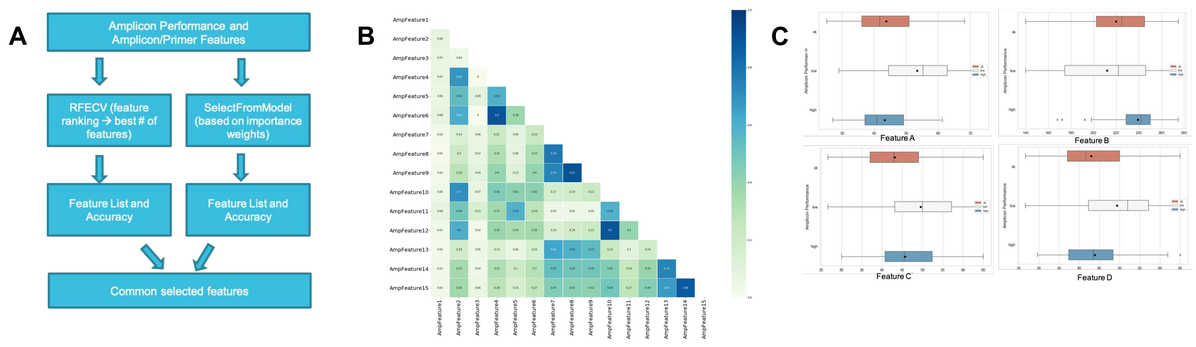
Abstract

Background
High throughput single-cell DNA sequencing allows for detection of rare mutations in cells and identification of sub-clones defined by co-occurrence of mutations. The big challenge with multiplex sequencing at single-cell level is the non-uniform amplification which results in inadequate coverage of mutations of interest. To address this challenge, we developed a machine learning engine to optimize amplicon design for uniform amplification by making reliable performance prediction.

Methods
10 different panels were designed with amplicons spanning a wide range of design properties. The tested amplicons are classified into low, average or high performer amplicons based on their normalized reads-per-cell value. The design properties of the amplicons are the features. Highly correlated features were identified and pruned. We used random forest classifier to calculate feature importance. Top features were identified using two different feature selection methods. We then 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 underlying Tapestri Designer.

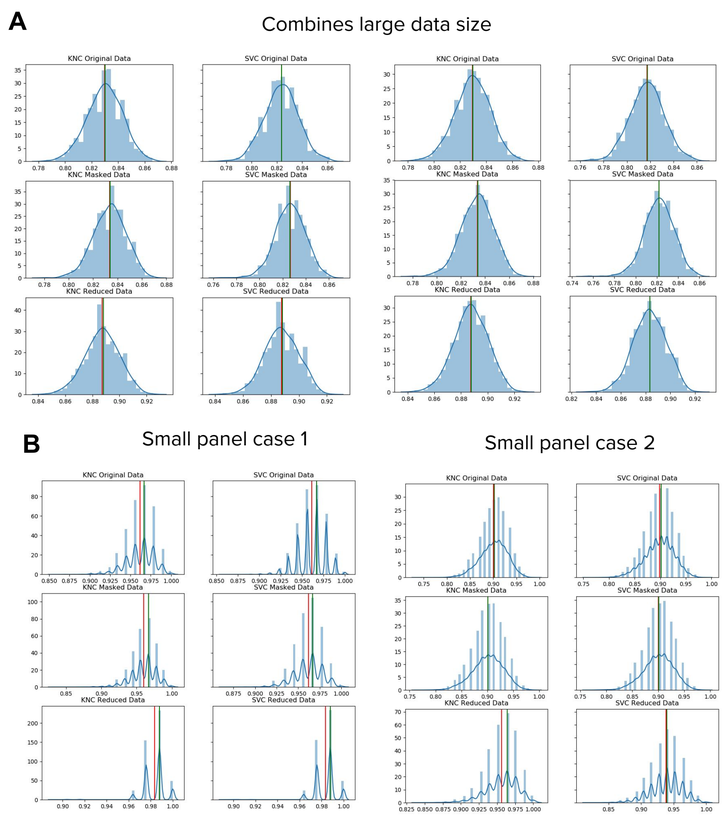
Results
We designed three different panels using the new pipeline. We achieved high panel performance of 97%, 92% and 88% across the three panels. The new parameters resulted in ~10-20% improvement in panel uniformity. We are working on further optimizing the performance prediction engine by using different ML classification models with K-fold cross validation, training using larger group of amplicons and optimizing features using combination of properties.

Identify important features impacting amplicon performance



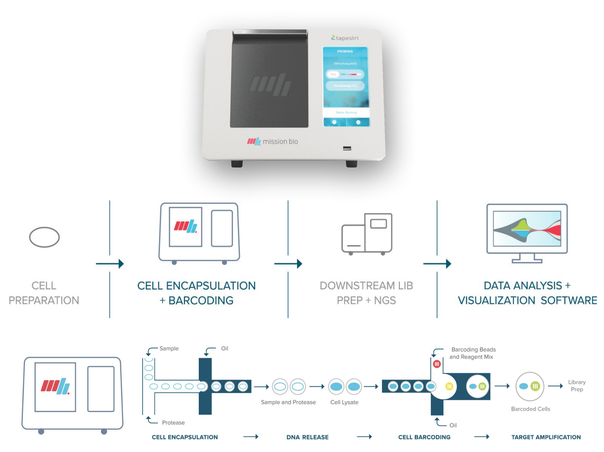
A. To identify important amplicon and primer properties that impact amplicon performance in Tapestri targeted sequencing assays, we used random forest classifier to calculate feature importance of properties. Amplicons are classified into low-performer, OK-performer and high-flyer based on their normalized reads-per-cell value. 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 plot of top feature distribution.

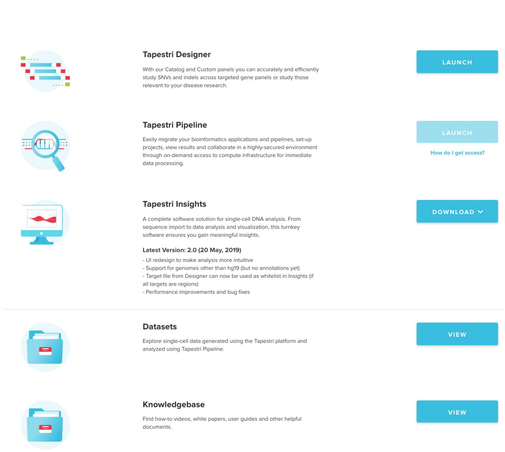
KNC and SVC model fit for different data size



Selected amplicon features and performance data for 10 targeted panels were used to train and test performance prediction model. **A.** 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. We also masked high-flyer or removed high-flyer from data set to understand accuracy of predicting amplion performance passing the minimal requirement at 0.2x mean. Average accuracy is 0.80-0.88 for large dataset. **B.** Small panels testing dataset showed higher accuracy score at 0.90-0.98.

Tapestri Workflow and Products



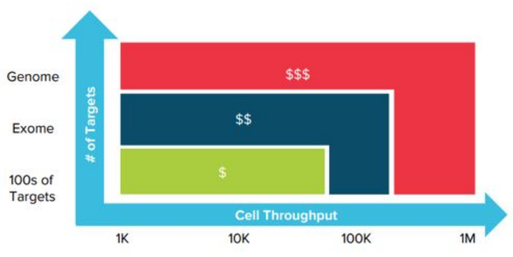


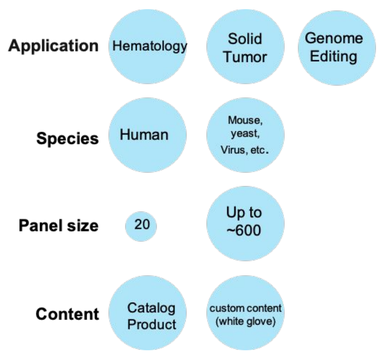
Single Cell Targeted Panel: Quality, Flexibility, Scalability

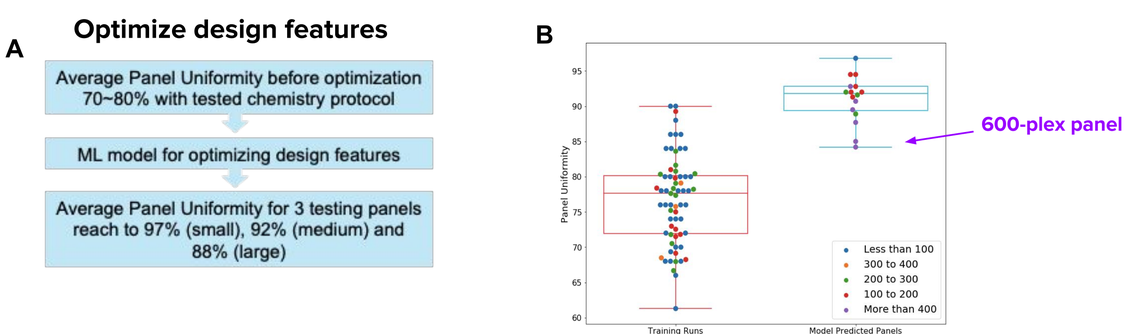
Quality: High molecular capture rate and panel uniformity
Measure every mutation with confidence

Flexibility: Choose how many cells and the size of panel
Fit to the experimental need and budget

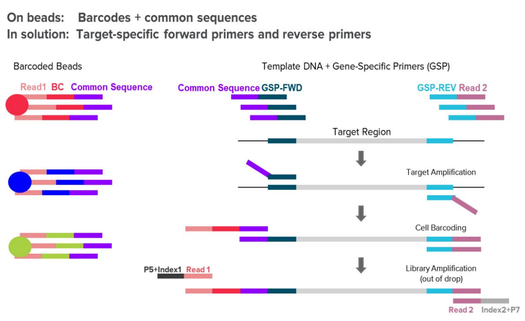
Scalability: Number of cells and targets scale over time
Build to increase panel plexy and cell throughput



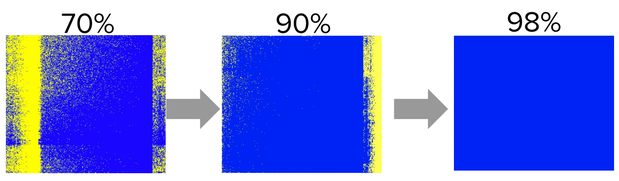




Barcode read structure

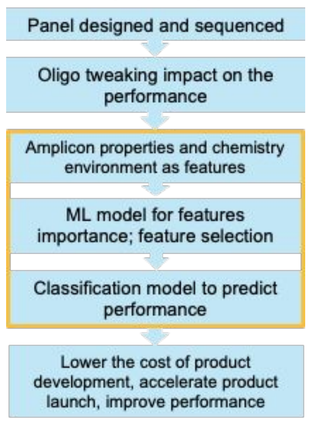


ML approach to improve the quality of panel



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

Predict amplicon performance



A. Schematic workflow of using ML model to optimize amplicon design. After design feature optimization, three testing panels including a small mouse panel (31 amplicons), a medium hg19 panel (128 amplicons) and a large hg19 panel (287 amplicons) were designed using new set of parameters. Multiple runs were conducted for each panel. New designer achieves high panel performance of 97%, 92% and 88% across the three panels. **B.** New designer significantly improved amplicon performance and uniformity in targeted assay design across different panel size and genomic contents. 6 newly designed panels were sequenced. Multiple runs were conducted for each panel. **C.** Schematic workflow of building performance prediction model to improve assay performance and product development process.

Tapestri Solutions

Catalog Panels for Hematology and Solid Tumor

TAPESTRI SINGLE-CELL DNA PANELS

	# genes	# amplicons	Target regions coverage (Kbp)	Panel Uniformity
AML Panel	19	50	~18.5	>90%
CLL Panel	34	286	~54.9	>90%
Myeloid Panel	47	330	~66.0	>80%
Tumor Hotspot Panel	59	244	~40.0	>80%

59 GENES - TUMOR HOTSPOT PANEL

ABL1	CSF1R	FGFR1	IDH2	MLH1	RB1
AKT1	CTNNB1	FGFR2	JAK1	MPL	RET
ALK	DDR2	FGFR3	JAK2	MTOR	SMAD4
APC	EGFR	FLT3	JAK3	NOTCH1	SMARCB1
AR	ERBB2	GNA11	KDR	NRAS	SMO
ATM	ERBB3	GNAQ	KIT	PDGFRA	SRC
BRAF	ERBB4	GNAS	KRAS	PIK3CA	STK11
CDH1	ESR1	HNF1A	MAP2K1	PTEN	TP53
CDK4	EZH2	HRAS	MAP2K2	PTPN11	VHL
CDKN2A	FBXW7	IDH1	MET	RAF1	

20-GENE AML PANEL

ASXL1	GATA2	KIT	PTPN11	TET2
DNMT3A	IDH1	KRAS	RUNX1	TP53
EZH2	IDH2	NPM1	SF3B1	U2AF1
FLT3	JAK2	NRAS	SRSF2	WT1

47-GENE MYELOID PANEL
(Available Q2 2018)

ASXL1	DNMT3A	IDH1	MYC	PTPN11	STAG2
ATM	ERG	IDH2	MYD88	RAD21	STAT3
BCOR	ETV6	JAK2	NF1	RUNX1	TET2
BRAF	EZH2	KDM6A	NPM1	SETBP1	TP53
CALR	FLT3	KIT	NRAS	SF3B1	U2AF1
CBL	GATA2	KMT2A	PHF6	SMC1A	WT1
CHEK2	GNAS	KRAS	PPM1D	SMC3	ZRSR2
CSF3R	HRAS	MPL	PTEN	SRSF2	

34-GENE CLL PANEL

ATM	CD79B	EZH2	MAP2K1	PAX5	SPEN
BCOR	CHD2	FAT1	MED12	PLCG2	TP53
BIRC3	CREBBP	FBXW7	MYD88	POT1	XPO1
BRAF	CXCR4	KLHL6	NFKBIE	RPS15	ZMYM3
BTK	DDX3X	KRAS	NOTCH1	SETD2	-
CARD11	EGR2	LRP1B	NRAS	SF3B1	-

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