Poster ID: I021

Development of a machine learning model for Aspyre Lung Blood: a new assay for rapid detection of actionable variants from plasma in NSCLC patients

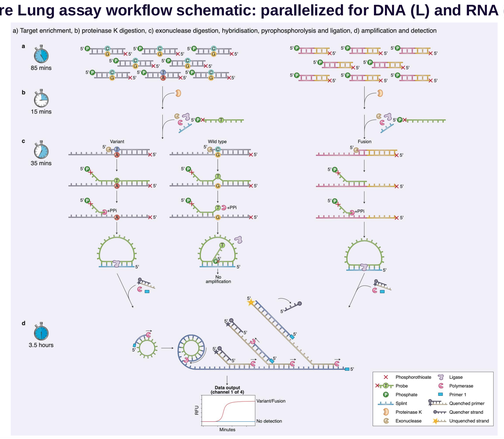
Rebecca Palmer[1], Sam Abujudeh[1], Magdalena Stolarek-Januszkiewicz[1], Ana-Luisa Silva[1], Justyna Mordaka[1], Kristine von Bargen[1], Alejandra Collazos[1], Simonetta Andreazza[1], Nicola Potts[1], Chau Ha Ho[1], Iyelola Turner[1] , Jinsy Jose[1], Dilyara Nugent[1], Prarthna Barot[1], **Christina Xyrafaki[2]**, Ryan Evans[2], Katherine Knudsen[2], Ellie Gillon-Zhang[2], Julia Brown[2], Candace King[2], Cory Kiser[2], Mary Beth Rossi[2], Eleanor Gray[1], Robert Osborne[1], Barnaby Balmforth[1]

1. Biofidelity Ltd., Cambridge, UK 2. Biofidelity Inc., Morrisville, North Carolina, USA

Introduction

Aspyre Lung is a targeted panel of 114 genomic variants across 11 guideline-recommended genes with simultaneous DNA[1] and RNA[2] workflows that makes molecular testing more accessible for NSCLC. In this study we developed a machine learning algorithm to interpret fluorescence data outputs from Aspyre Lung to adapt the assay from tissue to liquid biopsy samples. Data for model training and testing were generated from >13,500 DNA and RNA contrived samples, with variants spiked in at 0.1% to 82% VAF for DNA and 6 to 5000 copies for RNA. The training and testing datasets used 67 reagent batches and 23 operators using 9 qPCR machines at two sites. Variant calling machine-learning models were assessed in terms of median assay wide LoD95, observed sensitivity, false positive rate per sample, per variant LoD95, and per variant observed sensitivity. The model was optimized by varying the training data subsets, features used, and model hyperparameters. Models were assessed against target specifications as well as resistance to operator error, and robustness to variations in global inputs. Across models that passed target specifications, the median assay wide LoD95 range was predicted to be 0.23%-0.79% for DNA SNVs and indels, 1-9 copies for RNA fusions, 41-173 copies for *MET* exon 14 skipping, and per sample false positive rate was <0.4%. Verification with reference samples established the experimentally determined performance characteristics: SNV/indel sensitivity 0.19% VAF, 1 amplifiable copy of gene fusions, and 69 copies *MET* exon 14 skipping events with 100% specificity for all targets. Implementation of these models enables the analysis of both tissue and liquid biopsy samples with high sensitivity, specificity and accuracy within a single workflow.

Aspyre Lung assay workflow schematic: parallelized for DNA (L) and RNA (R)



Materials & Methods

Development and optimisation of an algorithm for variant calling

A large-scale set of experiments generated a data set to train and test the SVM algorithm to perform variant calling. Single variant and multi-variant samples were prepared by spiking in the variant-containing oligonucleotide into the wild type background of DNA or RNA and used as contrived controls. Contrived oligonucleotides made from DNA (SNVs, MNVs, indels) or RNA (fusions, *MET* exon 14 skipping) were externally manufactured (Eurofins or IDT), and quantified by digital PCR (Qiagen) at the Biofidelity R&D facility. cfDNA derived from healthy donor blood, DNA derived from FFPE tonsil tissue and gDNA were used as background for DNA reference samples and quantified by Qubit and dPCR. cfRNA extracted from healthy donor blood and RNA extracted from FFPE variant-free tonsil tissue was used as background for RNA samples and quantified by Qubit. To achieve the required VAF or copies, samples were serially diluted to the appropriate concentrations, and immediately frozen at -20°C (DNA) or -80°C (RNA).

To create a version of the support vector machine (SVM) model for variant calling for plasma, the following parameters were varied: training set, probes used, C, and scale compared to a previously established FFPE model[3]

- Training set – the data used to train models were varied
- Probes used – each variant detected by Aspyre Lung at the nucleic acid level has at least one directly associated oligonucleotide probe. The presence of a variant in a sample affects its directly associated probe (Figure 1) and can affect other probes in more subtle ways. As such, performance benefits may result from using signals from multiple probes to make individual variant calls.
- C - regularization parameter, which controls the trade-off between the hyperplane margin and misclassifications. Smaller values of C prioritize a wider margin and allows for more misclassifications on the training set, while larger values of C places more weight on correct classification at the expense of a narrower margin
- Scale – divides normalized parameters of Cycle of Sigmoid Midpoint (CSm) scale and S-curve height.

SVM Machine Learning Model Optimization

Data output from the laboratory workflow

- The final laboratory stage of the Aspyre Lung assay is amplification, monitored by real-time fluorescent read-out
- Each sample has 24 associated wells: 21 assigned to DNA and 3 to RNA (including positive and negative controls)
- Each well has four dyes detected by measurements at different wavelengths
- An algorithm is required to interpret and convert these fluorescence data into a 'detected', 'not detected' or 'undetermined' readout for each variant in the panel

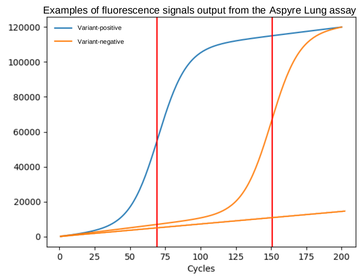


Figure 1 – Model examples of fluorescent curves that may be output by the assay, with two wild-type curves and a sample containing a variant. The variant-positive and one wild-type show typical sigmoidal curves, and the second wild-type sample has no inflexion point (flat-line). Red lines indicate cycle of sigmoid midpoints, the inflection point of the S

Varying inclusion of probe interactions

- Each DNA variant (SNVs, MNVs, indels) can receive input from between one probe (the probe designed to detect that specific variant) to all 80 DNA probes used across the assay
- RNA models tested either just probes from each RNA well or all RNA probes
- Inclusion of greater numbers of probes in models allows for previously unknown probe cross-talk effects and other subtle interactions to be taken into account when calling variants
- It was found (Fig. 2) that inclusion of probes from the same well as the intended target, in addition to the target's own probe and known cross-talk, led to a more sensitive model. The inclusion of all 80 DNA probes across all DNA wells did not provide a boost in sensitivity and was not chosen for the final model

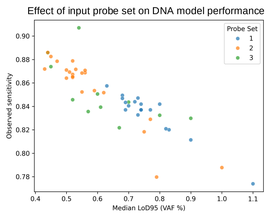


Figure 2 – The effect of modifying inclusion of probe interactions on assay sensitivity. Data shown are from the DNA targets of the assay.

Probe set 1: DNA probes include only their own probes and known cross-talk
Probe set 2: DNA probes include their own probes, those in the same well (usually 4 variants), and known cross-talk.
Probe set 3: DNA probes include all DNA probes of the assay (excluding driver-drug resistance combinations)

Varying data included in training sets

- The data used to train models included multiple sources of variation; reagent batches, raw materials, qPCR machines, operators, sites
- It was anticipated that the more training data used in fitting the models would lead to better performance
- This was broadly true, though excluding data generated from pre-lockdown versions of the assay (particularly for RNA variants) was better than including all available data
- Shown are the results for RNA targets (Fig. 3)

Figure 3 – The effect of varying the training data on assay performance. RNA targets

Training set 1: Original FFPE training set[3]
Training set 2: Set 1 + additional dataset containing plasma-like samples
Training set 3: Set 2 excluding FFPE DNA samples
Training set 4: Set 3 excluding *some* data generated before lock of reagent manufacturing procedures
Training set 5: Set 3 excluding *all* data before lock of reagent manufacturing procedures

Results

Choice of final model

- Final DNA and RNA models were chosen based on their estimated median LoD95, per-variant LoD95 estimates, observed sensitivity, observed and estimated false positive rates per sample (FPR/sample)

Training Set	Probe Set	C	scale	Median LoD95	Observed sensitivity	Observed FPR/sample	Estimated FPR/sample
3	2	0.01	0.6x	0.44% +/- 0.16%	89% +/- 6%	0.09% +/- 0.09%	0.08% +/- 0.07%

Table 1 – Chosen parameters for DNA models. Performance estimated using cross-validation

Training Set	Probe Set	C	scale	Median Fusion LoD95	*MET* LoD95	Observed sensitivity	Observed FPR/sample	Estimated FPR/sample
4	3	0.01	0.6x	0.5C +/- 0.6C	69C +/- 25C	95.8% +/- 0.8%	0.04% +/- 0.04%	0.017% +/- 0.018%

Table 2 – Chosen parameters for RNA models. Performance estimated using cross-validation

Performance of final models on verification data

- A set of highly prevalent and/or representative DNA and RNA variants were selected to verify the performance of the final DNA and RNA models (Table 1):

Aspyre target nucleic acid	Variant type	Gene	Exon	Protein variant	COSM ID
DNA	SNV	*KRAS*	2	G12C	COSM516
			3	Q61H	COSM554
		EGFR	21	L858R	COSM6224
				L861Q	COSM6213
		EGFR	20	T790M	COSM6240
		BRAF	15	V600E	COSM476
	MNV	*ERBB2*	17	V659E	COSM6503262
		KRAS	2	G12V	COSM515
	Deletion	*EGFR*	19	E746_A750del	COSM6223
					COSM6225
	Insertion	*ERBB2*	20	Y772_A775dup	COSM20959
				G778_P780dup	COSM12555
		EGFR	20	A767_V769dup	COSM12376
				A763_Y764insFQEA	COSM26720
RNA	Fusion	*EML4-ALK*	E13_A20	NA	COSF408
			E20_A20ins18		COSF730
		KIF5B-ALK	K24_A20	NA	COSF1058
		KIF5B-RET	K15_R12	NA	COSF1232
		TRIM33-RET	T14_R12	NA	NA
		NCOA4-RET	N6-R12	NA	COSF1341
		CCDC6-RET	C1-R12	NA	COSF1271
		CD74-ROS1	C6_R34	NA	COSF1200
		SDC4-ROS1	S4_R34	NA	COSF1280
		CD74-ROS1	C6_R32	NA	COSF1202
		TPM3-NTRK1	T8_N10	NA	COSF1329
		QKI-NTRK2	Q6_N16	NA	COSF1446
		ETV6-NTRK3	E5_N15	NA	COSF571
	Exon skipping	*MET*	14	L982_D1028del	COSM29312

Table 3 – DNA and RNA variants tested to determine the performance of different SVM models

- For each variant, 4 levels of VAF/copy number were selected to be close to the associated estimated LoD95. For RNA fusions, 6 copies was the lowest level selected to avoid drop-outs associated with random sampling (stochasticity)
- Assay runs included 6 independent batches of reagents, 10 operators, and 6 qPCR machines
- **Median LoD95 for DNA variants was found to be 0.19%**
- **Median LoD95 for RNA fusions was found to be 1 amplifiable copy.**
- Figures 4 and 5 show verification data for a SNV and gene fusion.

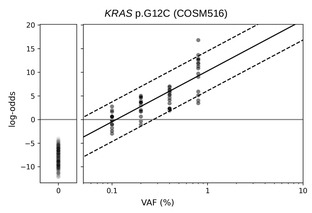


Figure 4 – SVM probability estimate relative to the SVM probability threshold (log-odds ratio) as VAF input changes for COSM516 (*KRAS* p.G12C). Graph shows distribution of log-odds for wild-type (i.e. where VAF is 0%). Log-odds of greater than 0 is called positive. LoD95 is estimated assuming a linear relationship between log VAF and log-odds, with Gaussian noise. Shown are the 5/50/95 percentiles of the linear fit; the point at which these lines cross the calling threshold (log-odds of 0) determines the estimated LoD5/50/95%.

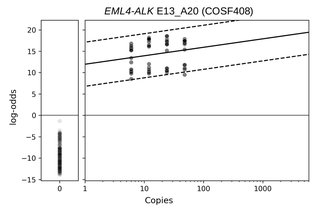


Figure 5 – SVM probability estimate relative to the SVM probability threshold (log-odds ratio) as copy number changes for *EML4-ALK* (COSF408). Left panel shows distribution of log-odds for wild-type (i.e. where copy number is 0). Log-odds of greater than 0 is called positive. LoD95 is estimated assuming a linear relationship between log VAF and log-odds as per Figure 4 (and that LoD95 <1 copy is impossible).

Summary

Aspyre Lung Blood is a pioneering biomarker panel assay detecting 114 variants of NSCLC from ctDNA and ctRNA in blood plasma

- Parallelized workflow for DNA and RNA, short hands-on time (1 hr 40m), total assay time of 14 hrs, easy implementation, no complex bioinformatics or data interpretation required

DNA / RNA	Target Amplification	Enzymatic Cleanup Reaction	Aspyre Reaction	Detection Reaction	Data Analysis
Instrument Requirements	Thermal cycler	Thermal cycler	Thermal cycler	Real-time PCR instrument	Desktop computer
Hands-on Time	30 mins	10 mins	35 mins	15 min	5 mins
Incubation Time	1 hour 25 mins	15 mins	35 mins	3 hours 30 mins	-

- Uses standard laboratory equipment (PCR machine and a real-time PCR machine)
- Cost-effective testing – the assay reports only genomic biomarkers associated with NSCLC, with no additional bioinformatics or expert interpretation required
- Assay sensitivity is 0.19% for SNVs & indels, 1 amplifiable copy for gene fusions, and 69 copies for *MET* exon 14 skipping

Aspyre Lung Blood performance			
	DNA	RNA	
	(SNVs & indels)	Fusions	*MET* exon 14 skipping
Sensitivity (Median panel-wide LoD95)	0.19% VAF	1 amplifiable copy	69 amplifiable copies
Specificity	100%	100%	100%

Aspyre Lung enables accessible, decentralized simplified genomic profiling for NSCLC, supporting both tissue and blood plasma samples in a single instrument run for 1 to 16 samples per batch.

The targeted panel covers 114 genomic variants across 11 genes, combines high sensitivity, specificity, and fast turnaround times through sophisticated machine learning algorithms.

Aspyre Lung Reagents (Research Use Only)

- Simultaneous analysis of DNA and RNA
- Comprehensive lung panel covering biomarkers across 11 key genes for NSCLC
- Runs on existing real-time PCR instruments
- Straightforward implementation
- Reduced sample requirements
- Fast time to result

Figure 5: Aspyre Lung

References

1. Silva *et al*. 2021. Single-copy detection of somatic variants from solid and liquid biopsy. *Sci Rep*. 11(1):6068.
2. Gray *et al*. 2022. Ultra-sensitive molecular detection of gene fusions from RNA using ASPYRE. *BMC Med Genomics*. 15(1):215.
3. Evans *et al*. 2024 Validation of a simple, fast, robust method for multi-variant genomic analysis of actionable NSCLC variants in FFPE tissue. *Frontiers in Oncology*. 14:1420162

*All authors are employees of Biofidelity Inc and may have a financial interest including salary, equity, options, and intellectual property.