Accurate Detection of Low-representation Alleles in Tumor DNA Through Augmented Exome Sequencing

Elena Helman, Michael Clark, Sean Boyle, Anil Patwardhan, Deanna Church, Mark Pratt, Shujun Luo, Nan Leng, Scott Kirk, Mirian Karbelashvili, Parin Sripakdeevong, Christian Haudenschild, Richard Chen, John West

Contact: **elena.helman@personalis.com**

Personalis, Inc. | 1350 Willow Road, Suite 202 | Menlo Park, CA 94025 • [1]National Center for Biotechnology Information, Bethesda, MD

Abstract

Somatic mutations present at a low allelic fraction have been implicated in tumor progression, recurrence, metastasis and drug resistance. The ability to detect these mutations via next-generation sequencing is crucial but often impaired by small sample size, low cellularity, and tumor heterogeneity. In order for sequencing to aid in directing personalized cancer therapies, these low-representation alleles must be accurately identified and interpreted. We sought to determine the sequencing depth and analytical parameters required to detect small variants in cancer by completing exome and transcriptome sequencing of four cancer cell lines. These lines contain a number of known mutations in ***EGFR, KRAS, BRAF,*** and more than 20 other cancer genes at well defined levels of allelic representation, with some mutations at fractions as low as 1%. We completed high-depth exome sequencing with our ACE exome assay, which augments and improves coverage of over 1200 cancer genes, increasing our sensitivity for cancer variants. We utilized modified bioinformatics approaches and integrated whole-transcriptome sequencing to validate our findings. We find that our method is highly sensitive at moderate depths, detecting mutations present at 1%; allelic fraction, and show how the limit of detection changes as depth of coverage changes. We also find high concordance between our estimated allelic fractions and known values. Notably, we recapitulate canonical ***EGFR*** mutations, such as T790M, which has been shown to confer acquired resistance to treatment with ***EGFR*** tyrosine-kinase inhibitor therapies.

Methods

To gauge the limits of detection of our augmented cancer exome on low-representation alleles, we obtained a set of four cell-line mixtures from a commercial vendor (1); these were acquired in triplicate for each mixture type, resulting in 12 independent aliquots. These mixtures contained a number of endogenous mutations present in each of the constituent cancer cell lines, as well as artificially engineered known cancer mutations (**TABLE 1**).

Chromosome	Gene	Variant	Mutation Type	Quantitative Multiplex (Ref)	EGFR Medium (5%) Multiplex (EGF)	EGFR Low (1%) Multiplex (RKO)	K-Ras Medium (5%) Multiplex (KRAS)
7q34	BRAF	V600E	Engineered	10.50%	66.70%	66.70%	
4q11-q12	cKIT	D816V	Engineered	10.00%			
7p12	EGFR	L861Q	Engineered		5.00%	1.00%	
7p12	EGFR	ΔE746 - A750	Engineered	2.00%	5.00%	1.00%	
7p12	EGFR	L858R	Engineered	3.00%	5.00%	1.00%	
7p12	EGFR	T790M	Engineered	1.00%	5.00%	1.00%	
7p12	EGFR	G719S	Engineered	24.50%	5.00%	1.00%	33.30%
12p12.1	KRAS	G13D	Engineered	15.00%			5.00%
12p12.1	KRAS	G12D	Engineered	6.00%			5.00%
12p12.1	KRAS	Q61H	Engineered				5.00%
12p12.1	KRAS	A146T	Engineered				5.00%
1p13.2	NRAS	G12V	Engineered				5.00%
1p13.2	NRAS	Q61K	Engineered	12.50%			5.00%
3q26.3	PI3KCA	H1047R	Engineered	17.50%	50.00%	50.00%	
3q26.3	PI3KCA	E545K	Engineered	9.00%			
7q31	MET	V237fs	Engineered	6.50%			
3p21.3	MLH1	L323M	Engineered	8.50%	50.00%	50.00%	
9q34.3	NOTCH1	P668S	Engineered	31.50%			50.00%
1q21-q22	NTRK1	5'UTR	Engineered	8.50%			
4q12	PDGFRA	G426D	Engineered	33.50%			50.00%
2p23	ALK	P1543S	Endogenous	33%			50.00%
1q25.2	ABL2	P986fs	Endogenous	8%			
5q21-q22	APC	R2714C	Endogenous	33%			50.00%
1p35.3	ARID1A	P1562fs	Endogenous	33.50%			
13q12.3	BRCA2	A1689fs	Endogenous	33%			
13q12.3	CDX2	V306fs	Endogenous	41.50%			
22q13.2	EP300	K291fs	Endogenous	8%			
4q31.3	FBXW7	G667fs	Endogenous	33.50%			50.00%
8p12	FGFR1	P150L	Endogenous	8.50%	50.00%	50.00%	
13q12	FLT3	S985fs	Endogenous	10.50%			
13q12	FLT3	V197A	Endogenous	11.50%			
2q33.3	IDH1	S261L	Endogenous	10%			
16p13	CDH1	3'UTR	Endogenous		50.00%	50.00%	
3p21	CTNNB1	S33Y	Endogenous				50.00%

TABLE 1: Known mutations in four cell-line mixtures, referred to as Ref, EGF, RKO, KRAS.

DNA was extracted from each of the 12 samples and sequenced using our augmented cancer exome, which corrects for GC and coverage bias above a standard exome. Samples were then analyzed using our bioinformatic pipeline that has been optimized for low-representation alleles in highly heterogenous tumor samples. Finally, we assessed resulting somatic SNV and indel calls as well copy-number profiles against known metrics to determine the limits of detection of our exome sequencing process.

In order to test the sensitivity of our process against diverse coverage levels, we computationally combined the triplicates for each sample and then down-sampled these combined aligned sequence files to varying depths of coverage (8 separate downsampled subsets) and ran our pipeline on each of these individually.

Finally, we conducted an orthogonal experiment by mixing, *in-silico*, various ratios of exomes from two normal individuals, one of which is the NIST-GIAB gold standard (3), NA12878. Sensitivity was determined by comparing the SNVs called as unique to NA12878 to the set of known gold variants in high confidence exonic regions. In this artificial system, we can interrogated 3171 variants at 13 different ratios/allele frequencies and observed coverage at each SNV was used to determine robustness against varying coverage.

Results

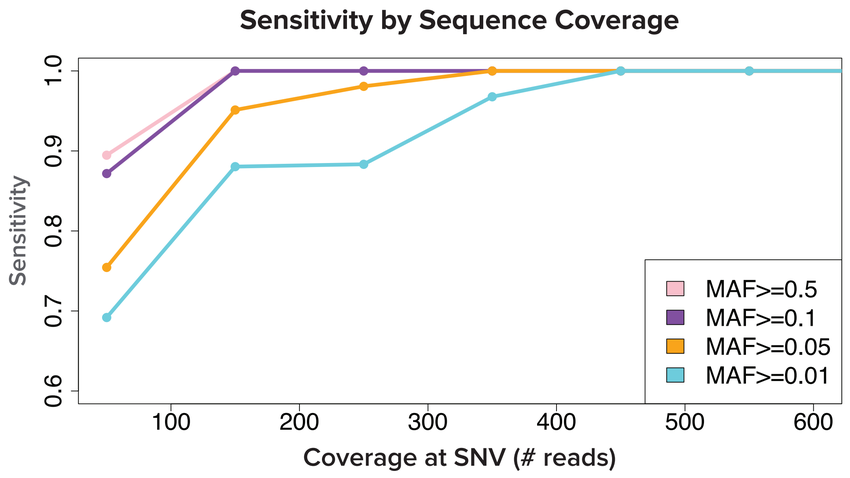


FIGURE 1: Sensitivity of detection at varying sequencing depths across known engineered mutations at four minor allele frequency (MAF) thresholds.

With data from four high coverage cell-line mixtures, downsampled eight times, we were able to compute actual read depth at each known mutation. For mutations at varying levels of allelic frequency, we calculated the sensitivity of our somatic SNV calls at each level of minimum sequence coverage. We reach 99% sensitivity for 10% alleles at only 100-200X, while we capture 1% alleles to the same accuracy at 400-500X.

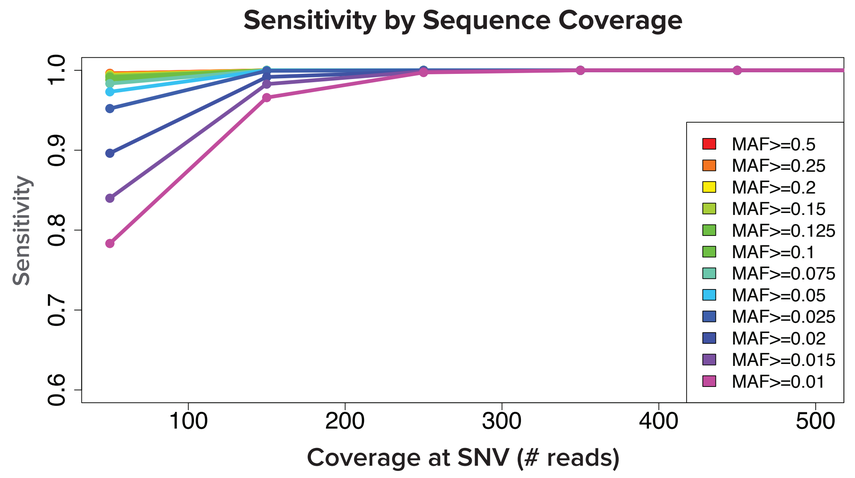


FIGURE 2: Sensitivity of detection at varying sequencing depths across known gold standard variants at several minor allele frequency (MAF) thresholds.

By computationally mixing alignments from two individuals, including the well-characterized NA12878, we were able to assess 38,000 observations of variant calls across a range of allele frequencies and coverages. In this artificial system, our pipeline achieves an even greater sensitivity, with 95% detection of 1% alleles at 100-200X.

Reproducibility

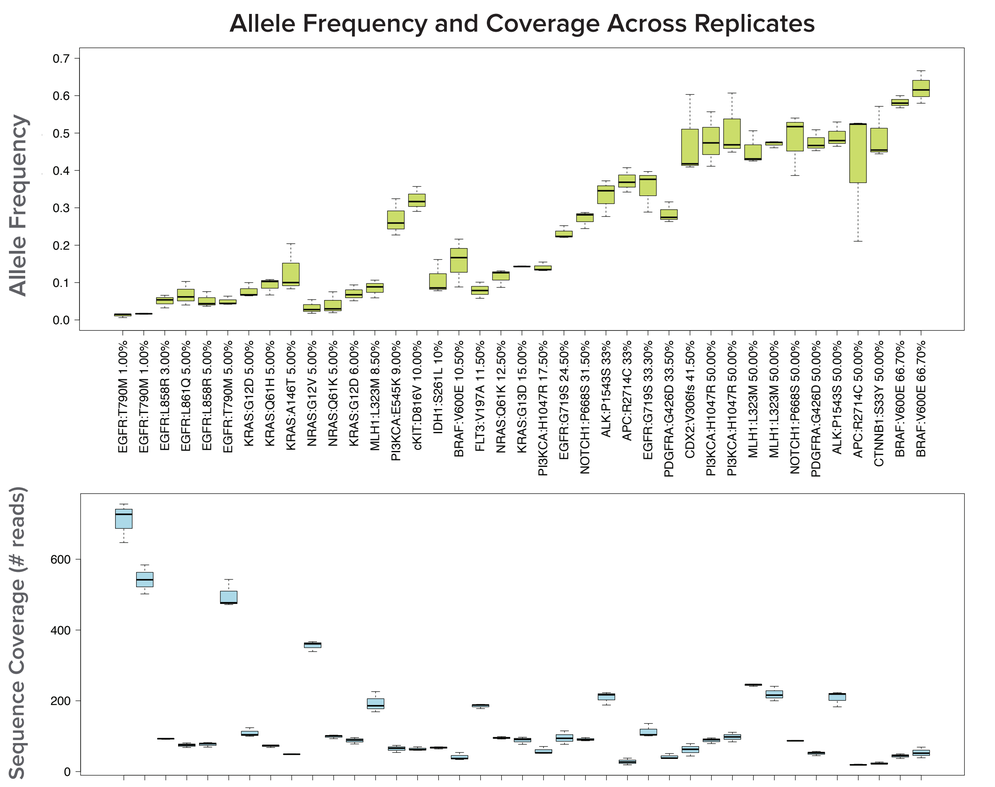


FIGURE 3: Allele frequency (top panel) and sequencing coverage (bottom panel) of each variant across three independent replicates.

Predicted allele frequency and sequencing coverage of each known mutation in the cell line mixtures were highly consistent across triplicate samples. In two cases, ***PIK3CA*** E545K and ***cKIT*** D816V, all three independent replicates predicted an allelic fraction that varied by 20% from what was given as the ground truth allelic fraction. Coverage of certain mutations is consistently elevated across each independent sequencing experiment two-to-seven-fold higher than mean sequencing depth.

Accuracy

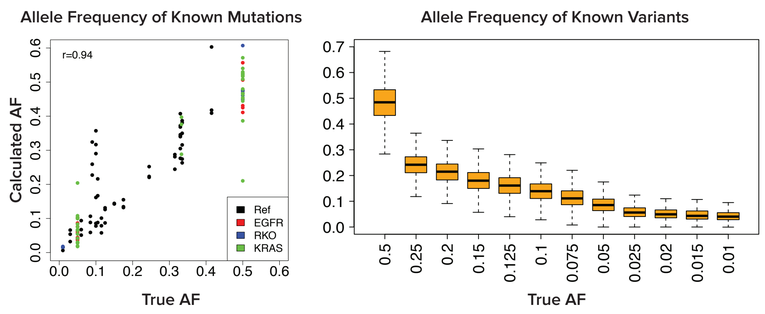
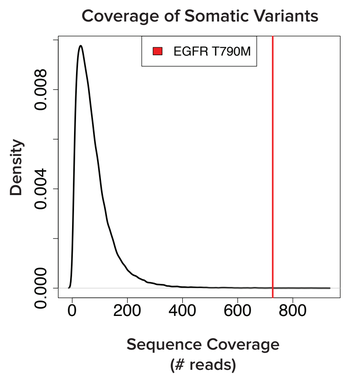


FIGURE 4: Calculated versus known allele frequency (AF) across all samples (left panel) and across all mixture ratios (right panel).

Allelic frequencies predicted by our pipeline are highly concordant (r=0.94) with known fraction of engineered or endogenous mutation. Similarly, with computationally mixed fractions designed to provide a variance around the designated frequency, we are still able to recapitulate these allele frequency ranges.

Cancer Variants



EGFR T790M is an important acquired secondary mutation in cancer because it has been shown to confer resistance to treatment with common tyrosine-kinase inhibitors, such as gefitinib and erlotinib (2). Coverage of this particular variant is enhanced 7-fold in our ACE cancer exome.

FIGURE 5: Sequencing coverage of all somatic variants identified in an engineered cell-line mixture, with coverage of ***EGFR*** T790M highlighted in red.

Copy Number Profiles

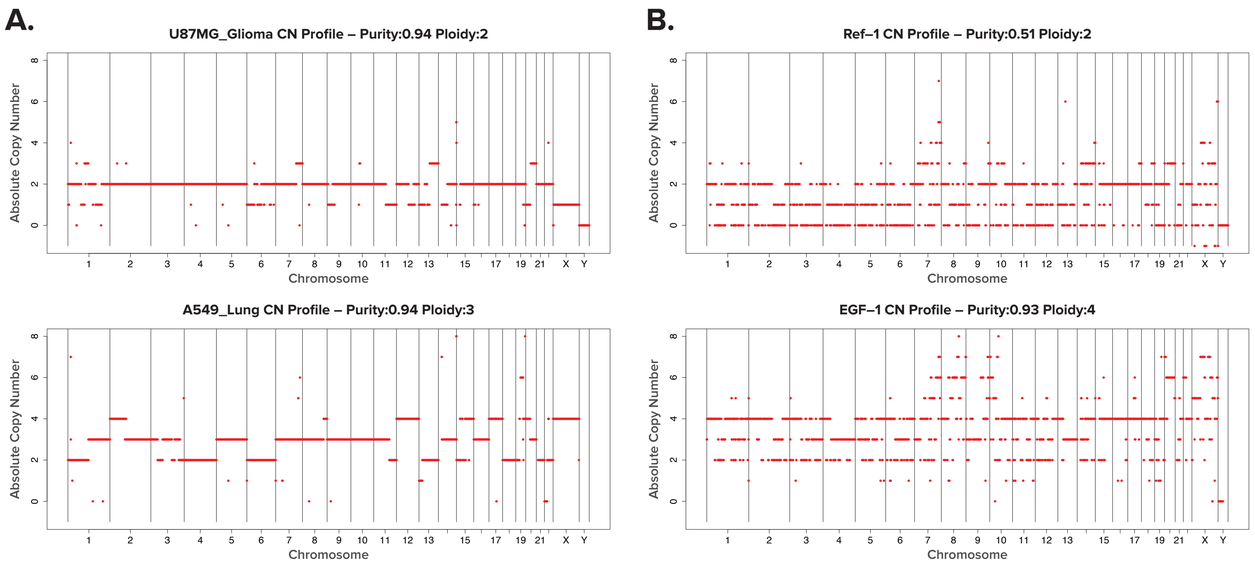


FIGURE 6: Copy number profiles of **A)** U87MG_Glioma, A549_Lung cell lines and **B)** Ref, EGF cell-line mixtures (left to right, top to bottom).

Absolute copy number profiles of cell lines generally demonstrate large structural aberrations and aneuploidy in these samples (**FIGURE 6A**). The cell-line mixtures in this study, however, show extreme copy-number heterogeneity (**FIGURE 6B**), providing further evidence of the mixed origin of these samples as well as our ability to detect low-representation small variants despite highly heterogeneous input (typical of cancer genomes).

Somatic Indel Detection

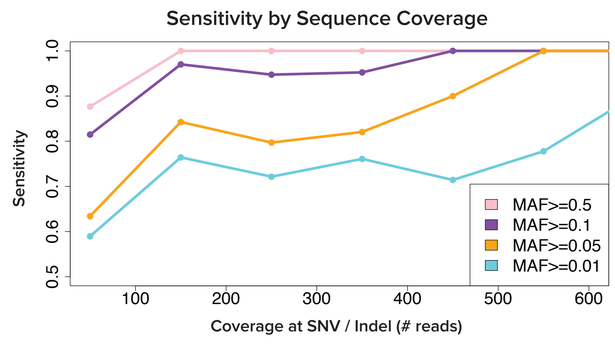


FIGURE 7: Sensitivity of detection at varying sequencing depths across known engineered variants (both SNVs and indels) at four minor allele frequency (MAF) thresholds.

Somatic indel detection remains an area of active research. We assessed 13 additional small insertion/deletion events engineered in the cell line mixtures. We find that even at deep sequencing depths, some true indel events escape detection as a result of filters designed to alleviate false positives due to proximity to homopolymer regions in the reference genome. The ability to distinguish between real mutations near homopolymers, which are prevalent in cancer genomes, and sequencing error due to these stretches is an open question we are currently investigating.

References

1. Horizon Discovery Group plc. 7100 Cambridge Research Park, Waterbeach, Cambridge CB24 9TL. United Kingdom
2. Kobayashi et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. NEJM, 2005 Feb 24; 352(8):786-92.
3. Zook et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology 32, 246-251 (2014)

Conclusions

Important cancer mutations are often present at small allelic fractions because of tumor heterogeneity and low purity samples. For effective cancer exome analysis, accurate detection of these variants is of paramount importance. Existing assays report limits of detection in the 5-10% range. Here, we use mixture experiments to quantitatively determine the detection limits of our augmented exome protocol and analysis pipeline and report its high sensitivity to low-representation cancer mutations.