# A computational method for somatic vs. germ-line variant status determination from targeted next-generation sequencing of clinical cancer specimens without a matched normal control

James Sun[1], Garrett Frampton[1], Kai Wang[1], Jeffrey S. Ross[1], Vincent A. Miller[1], Philip J. Stephens[1], Doron Lipson[1], Roman Yelensky[1]

[1]Foundation Medicine Inc., Cambridge, MA

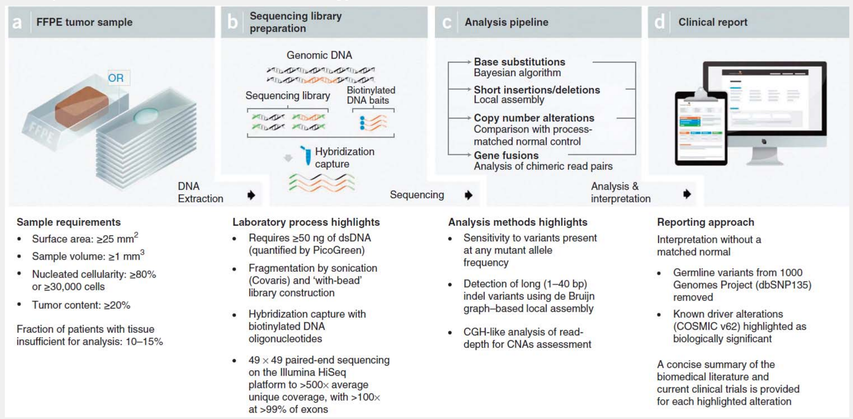
## Background

A key practical constraint in genomic testing in oncology is that matched normal specimens are not commonly obtained. Thus, while most clinically relevant genomic alterations have been previously characterized and do not require normal tissue for interpretation, the use of novel variants whose somatic status is unknown is limited. Here, we describe our approach to assessing somatic status of genomic alterations from tumor tissue alone using a CLIA-certified, NGS-based test that interrogates all exons of 236 cancer-related genes.

## Methods

**Process specimens using Foundation Medicine's NGS-based cancer assay**

Frampton, et. al *Nature Biotechnology* Nov 2013




**Fit a statistical model to determine copy number genome-wide, obtaining copy number and LOH estimates for each chromosomal segment**

Let:
- $S_i$ be a genomic segment
- $l_i$ be the length of $S_i$
- $r_{ij}$ be the LR of exon $j$ within $S_i$
- $f_{ik}$ be the minor allele frequency of SNP $k$ within $S_i$

We seek to estimate $\boldsymbol{p}$ – tumor purity, and $\boldsymbol{C_i}$ – the copy numbers of $\boldsymbol{S_i}$.

Jointly model $r_{ij}$ and $f_{ik}$, given $p$ and $C_i$:

$r_{ij} \sim N(\log_2 \frac{p*C_i+(1-p)*2}{p*(\sum_i l_i C_i)/\sum_i l_i+(1-p)*2}, \sigma_{ri})$

$f_{ik} \sim N(\frac{p*M_i+(1-p)*1}{p*C_i+(1-p)*2}, \sigma_{fi})$

- $M_i \leq C_i$ is number of alt. alleles at $S_i$
- $\sigma_{ri}$ and $\sigma_{fi}$ are noise parameters

Fit model using standards methods - e.g. MCMC, assigning copy # to all segments

For each genomic segment *i*:
- If $C_i = M_i = 0$, the segment has homozygous deletion in tumor
- If $C_i = M_i \neq 0$, the segment has LOH in tumor
- If $C_i \neq M_i \neq 0$, the segment is heterozygous in tumor

**For each mutation identified, use model fit to assess differences in expected allele frequencies (AF) between germline, somatic, and subclonal somatic mutations. Statistical confidence assessed based on read depth and local variability in allele frequency estimates.**

A germline variant at segment *i* has expected AF:

$$AF_{germline} = \frac{pM_i + (1-p)}{pC_i + 2(1-p)}$$

A somatic mutation at segment *i* has expected AF:

$$AF_{somatic} = \frac{pM_i}{pC_i + 2(1-p)}$$

A subclonal somatic mutation at segment *i* has expected AF:

$$AF_{subclonal} \ll \frac{pM_i}{pC_i + 2(1-p)}$$

| | | | sample purity (p) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| copy number | LOH status | status of variant | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| C=1 | LOH | M=0 germline | 50 | 47 | 44 | 41 | 38 | 33 | 29 | 23 | 17 | 9 | 0 |
| | | M=1 somatic | 0 | 5 | 11 | 18 | 25 | 33 | 43 | 54 | 67 | 82 | 100 |
| | | M=1 germline | 50 | 53 | 56 | 59 | 63 | 67 | 71 | 77 | 83 | 91 | 100 |
| C=2 | LOH | M=0 germline | 50 | 45 | 40 | 35 | 30 | 25 | 20 | 15 | 10 | 5 | 0 |
| | | M=2 somatic | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| | | M=2 germline | 50 | 55 | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 | 100 |
| | het | M=1 somatic | 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| | | M=1 germline | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |

**Table 1.** Expected allele frequency table for example copy numbers, given purity (p), copy number (C), and alternative allele count (M). Low purity (<20%) samples are relatively easier for assessing somatic status, but more difficult in assessing tumor LOH. High purity (>90%) samples are easier for assessing tumor LOH, but more difficult in assessing somatic status. Tumor samples that are well-admixed with surrounding normal tissue are optimal – most clinical cancer specimens are in this category.

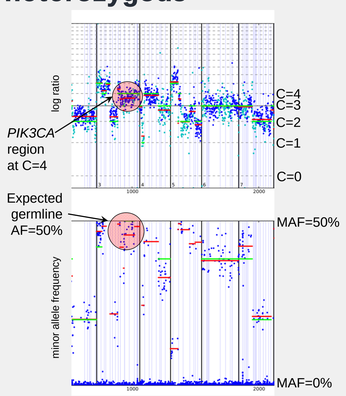
## Results

### Example 1: Known oncogenic driver mutation predicted as somatic and heterozygous

**Candidate mutation:**

*PIK3CA* H1047R at frequency 36%

1. Model shows the tumor has 4 copies of *PIK3CA*, with 2 variant alleles.
2. The genomic segment containing *PIK3CA* is not under LOH in the tumor.
3. *PIK3CA* H1047R at 36% is significantly below threshold of a germline variant (expected AF=50%) but matches a full clonal somatic mutation (expected AF=38%)
4. *PIK3CA* H1047R is somatic and heterozygous in tumor

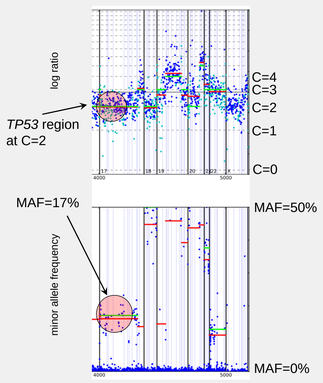



### Example 2: Known tumor suppressor mutation predicted as germline with LOH

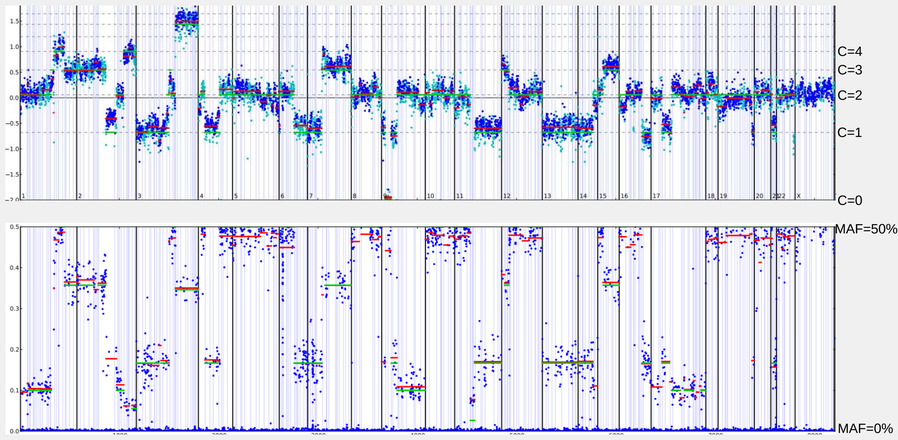
**Candidate mutation:**

*TP53* G356R at frequency 85%

1. Model shows the tumor has 2 copies of *TP53*, with 2 variant alleles.
2. The genomic segment containing *TP53* is under LOH in the tumor.
3. *TP53* G356R with frequency at 85% is significantly above threshold of a somatic variant (expected AF=65%) but matches a germline mutation (expected AF=83%/MAF=17%)
4. *TP53* G356R is germline and homozygous in tumor.




### Example 3: Genome-wide copy number model integrated with variant somatic vs germline status




| chr | start (Mb) | end (Mb) | CN | LOH | chromosome arm level gains and LOH | Status of short variants |
|---|---|---|---|---|---|---|
| chr1 | 1 | 120 | 2 | LOHx | 1p_LOHx | |
| chr3 | 1 | 90 | 1 | LOH1 | 3p_LOH1 | |
| chr3 | 130 | 198 | 6 | none | 3q_gain | |
| chr5 | 1 | 180 | 2 | none | | |
| chr8 | 1 | 146 | 2 | none | | |
| chr13 | 1 | 115 | 1 | LOH1 | chr13_LOH1 | *BRCA2* D651N somatic and homozygous |
| chr17 | 1 | 8 | 1 | LOH1 | 17p_LOH1 | *TP53* R282W somatic and homozygous |
| chr21 | 31 | 48 | 1 | LOH1 | | |

**Table 2.** Genome-wide copy number model with select chromosomes annotated. p-arm of chromosome 1 is under copy-neutral LOH (LOHx), while the entire chromosome 13 is under copy-loss LOH (LOH1). Somatic status of functional mutations is reported.

### Validation: Accuracy >95% with up to 85% call rate in 3 studies

| Study | Call rate | Somatic variants predicted correctly | Germline variants predicted correctly |
|---|---|---|---|
| I. Validating predictions on all called variants in 30 lung & colon cancer FFPE specimens, in which the tumor and the matched-normal tissue were sequenced | 84% (479/567) | 95% (311/326) | 99% (151/153) |
| II. Validating predictions in >2,500 FFPE specimens at sites with known status including 17 somatic hotspot mutations (e.g. *KRAS* G12, *PIK3CA* H1047, *BRAF* V600) and 20 common germline SNPs | 85% (4771/5583) | 96% (2556/2665) | 98% (2062/2106) |
| III. Assessing the impact of stromal admixture on predictions for all called variants using 3 cell lines (HCC-1937, HCC-1954, NCI-H1395) titrated with their matched-normal to 6 levels of tumor content, from 10% to 75% tumor | 83% (184/222) | 97% (60/62) | 97% (118/122) |

**Table 3. Validation studies.** Call rate is the proportion of variants that a germline vs somatic assessment could be made.

## Conclusions

- We leveraged deep, uniform NGS sequencing of clinical cancer specimens to design a computational method that can predict somatic status of mutations without the need for a matched normal control.
- Accuracy of the method is >95%, demonstrated using 3 independent validation approaches.
- The analytical framework also assesses tumor LOH status of identified variants, and the sub-clonality of somatic mutations.
- The method can advance functional prioritization of identified clinical variants and further inform clinical decision making.