# Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples

Paper by: Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, & Gad Getz

In *Nature Biotechnology*; vol. 31, no. 3 (March 2013)

Review Presentation by: Frank Austin Nothaft

October 31st, 2013

# Problem Statement

- Somatic SNP mutations are common mechanisms that alter gene function in cancer
- However, hard to call because they occur at a low frequency (0.1–100 mutations/Mbp)
- May not be present in all DNA from the locus
  - Cross contamination from normal cells
  - Copy number variation
  - Tumor may be subclonal

# Subclonality

- Tumor has a single set of mutations that cause tumor genesis
- But, tumor evolves to contain multiple different clonal colonies with divergent mutations
- Can analyze through several approaches:
  - Analyze mutations from metastasized tumors
  - Perform ultra deep sequencing
  - Sequence a small number of cells

# Mutation Detection Setup

- Based on methodology from TGCA
- Sample tumor exomes at 100-150x depth, genome at 30-60x (somatic)
- Reference normal sampled at 30-60x (germ-line)

# Measuring Success

- Successful mutation caller will call very rare alleles but will not call false positives
- Tradeoff between sensitivity and specificity:
  - Need to amplify signal (sensitivity) while not amplifying noise (specificity)
- Can lose specificity by calling too many variants in tumor, or too few in germline

# Downsampling - Methodology

- Take subset of reads with known mutations
- Randomly remove reads until desired coverage is reached
- Sensitivity at allele frequency is determined by percentage that can still be called

# Downsampling - Discussion

- Allows us to measure sensitivity at arbitrary coverage

- Cons:
  - Small number of validated events
  - Allele fractions are preserved, so only previously validated fractions can be explored
  - Excludes mutations that were not previously detected
  - Cannot measure specificity

# Virtual Tumor - Methodology

- First, generate a virtual tumor which contains only false positives
  - Generated from two runs of sequencing from same normal sample
- Then, inject high confidence heterozygous event reads from another sample

# Virtual Tumor - Discussion

- Specificity: “True” mutations are known (heterozygous events injected), so all other called mutations are false positives
- Sensitivity: Measured by percentage of heterozygous events detected and weight of reads injected
- Cons:
  - Heterozygous event signature does not match signature of a mutation event

# Detecting Mutations

- Four steps:
  1. Remove low quality reads
  2. Detect variants with Bayesian classifier
  3. Filter to remove false positives
  4. Designate detected variant as germline/somatic

- Mutations are variants that are conclusively not detected in gemline

- Use steps 1 and 3 to improve specificity

# Variant Detection

- Two models:
  - Reference model: Assume non-reference bases are due to sequencing errors
  - Variant model: Assume site contains a true allele
- Variant model:
  - Frequency is unknown but is modeled as the fraction of sample reads that support the mutation
- Detect variant if log(likelihood reference/likelihood variant) exceeds threshold
  - Use fixed 6.3 threshold → $10^{6.3}$:1 in favor of reference

# Filter Variants (1/2)

- Proximal Gap:
  - Remove false positives caused by misaligned indels
  - Reject if ≥3 indels in 11pb window
- Poor mapping:
  - Reject if ≥50% of reads have mapping quality of 0, or no observation of SNP variant with mapQ ≥ 20
- Triallelic site:
  - Reject if normal sample is heterozygous and mutation considered is a third allele

# Filter Variants (2/2)

- Strand bias:
  - Separate reads by strand direction and apply test
  - Reject if LOD is < 2.0
- Clustered position:
  - Reject false positives caused by misalignments
  - Consistent distance to start/end of alignment
- Observed in control:
  - Check matched normal data
  - Reject if observed in ≥2/3% of reads with quality score summing to greater than 20

# Variant Classification

- Three classifications:
  - Somatic if not in matched normal
  - Germ-line if present in matched normal
  - Indeterminate if insufficient data in matched normal

- Indeterminate if germ line likelihood is <95%

# Results - Sensitivity

- Validated against:
  - 3753 validated colorectal cancer mutations (100x coverage)
  - Exome capture from dbGAP (Genotypes and Phenotypes)
  - Virtual tumor from deep coverage genome

- MuTect is highly sensitive
  - @30x, detect F=0.2 with 95.6% sensitivity
    - F=0.1 with 58.9 sensitivity
  - F=0.2 sensitivity hits 99.9% @ 50x coverage
  - 150x coverage yields 66.4% sensitivity for F=0.03

# Results - Specificity

- Check against 1 Gbp of NA12878 data
  - Varying depth in virtual tumor, 30x in virtual normal
- No filters:
  - 5x coverage → 6.7 false positives per Mbp
  - 30x coverage → 20.1 false positives Mpb
- Filters:
  - HC reduces to 1/Mbp
  - HC+PON filters to 0.5/Mbp