Validation of the Firehose somatic variant calling pipeline for clinical whole exome sequencing.

Niall J Lennon, Alyssa Macbeth, Kristian Cibulskis, Adam Kiezun, Cassandra Crawford, Chris Friedrich, Sonic Team, Sheila Dodge, Long Phi Le, and Stacey Gabriel
Broad Institute of MIT & Harvard. Cambridge, MA, USA.

CLINICAL RESEARCH SEQUENCING PLATFORM
BROAD INSTITUTE OF HARVARD & MIT

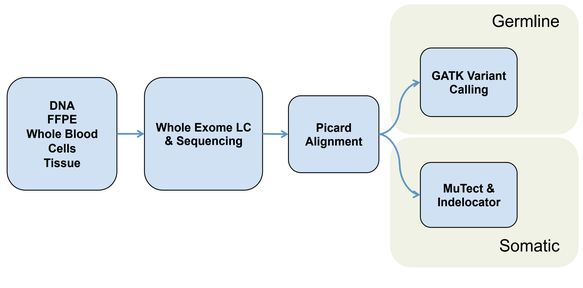
BROAD INSTITUTE GENOMICS PLATFORM

The Clinical Research Sequencing Platform (CRSP) is a CLIA certified, CAP accredited clinical laboratory housed within the Genomics Platform of the Broad Institute. CRSP was founded to further the Broad Institute mission of creating, and making available, tools for genomic medicine and applying them to human diseases.

We present here the validation of the Firehose tools, MuTect and Indelocator, for somatic variant detection, when applied to our clinical whole exome sequencing data from matched tumor and normal patient samples. Validation of high complexity tests such as whole exome sequencing, particularly in the context of somatic alteration, requires careful consideration so as to accurately represent the true sensitivity and specificity of the tools being validated across the range of expected allelic fractions.

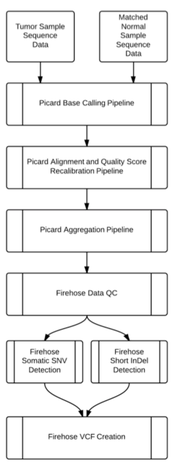
The clinical whole exome sequencing process

Below is a high level process map of the whole exome process in the clinical lab. Our clinical lab takes advantage of industrial process design, automated fluid handling, and advanced LIMS capabilities that have been tried and tested on hundreds of thousands of research samples in the last decade. Similarly, our clinical analytical pipelines leverage best-in-class, Broad-developed tools that are widely used in the research community. These tools include the Genome Analysis Toolkit (GATK) for germline variant calling and several modules of the Firehose framework (MuTect, Indelocator) for somatic variant calling.



Clinical whole exome pipeline for somatic variant calling

Exome data from tumor and matched nomal samples are used for identification of somatic alterations in cancer cases, according to the processing schema below.



Validation strategy and elements

Our validation strategy was informed by the ACMG[1] and CDC[2] guidelines for validations of NGS assays. We partitioned the validation into the following major categories:

1. Analytical Sensitivity
2. Clinical Sensitivity
3. Specificity
4. Precision

Other validation elements that were included in this study and in our germline exome validation but are not presented here include:
1. Sample identity QC - Concordance of exome data with Fingerprinting SNV assay from same stock sample. Specification: LOD>10
2. Contamination estimation - Genotype-free estimation of contamination in each sample dataset. Spec: <1%
3. Index crosstalk - Check for incorrect assignment of indexed reads in a pooled library.
4. Intra-run precision - Concordance of variants found from the same sample on two lanes of the same flowcell
5. Inter-run precision - Concordance of variants from the same sample run on different instruments and by different operators

Validation results

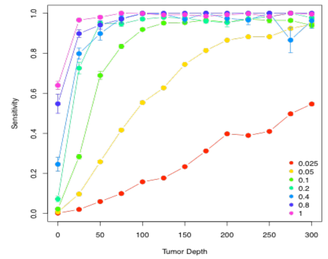
Analytical Sensitivity

To determine the sensitivity of the MuTect algorithm to detect SNVs, and the Indelocator algorithm to detect small insertions and deletions, in tumor samples at varying allelic frequencies, we leveraged a recently published approach[3] that utilized several pools of reference samples, from the 1000 Genomes Project, to create test samples containing several hundred thousand germline SNVs, and several thousand Indels, at a range of allelic frequencies. In addition PCR-free Whole Genome data was created for the same cell lines in order to more accurately establish germline variants for comparison.

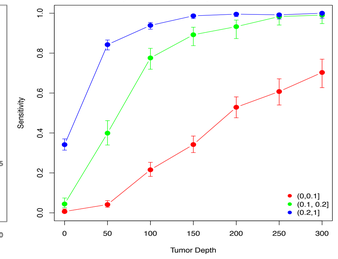
Samples Pooled

Sample ID	Gender
HG00096	Male
HG01500	Male
HG01565	Male
NA19625	Female
HG01583	Male
HG00759	Female
HG01112	Male
HG01051	Male
HG03742	Male
NA18525	Female
HG00419	Female
HG02922	Female
NA19017	Female
NA20845	Male
HG01879	Male
HG01595	Female
NA18939	Female
HG00268	Female
NA19648	Female
NA20502	Female

SNV Sensitivity at varying read depths and allele fractions



Indel Sensitivity at varying read depths and allele fractions



Read Depth	#SNVs at this depth	Sensitivity @ Allele Frequency (AF) AF = 0.025 ± 0.005	AF = 0.05 ± 0.005	AF = 0.1 ± 0.005	AF = 0.2 ± 0.005
50	64807	6%	26%	69%	94%
100	129454	16%	55%	93%	98%
150	188030	24%	75%	95%	98%
200	227646	40%	87%	94%	97%
250	251747	41%	88%	96%	98%
300	266172	55%	94%	95%	94%

Read Depth	#InDels at this depth	Sensitivity (S.E.M) @ Allele Frequency (AF) AF (0, 0.1]	AF (0.1,0.2]	AF (0.2,1]
50	3,476	4%	40%	84%
100	5,019	22%	78%	94%
150	6,228	34%	89%	99%
200	7,117	53%	93%	100%
250	7,691	61%	98%	99%
300	8,101	70%	99%	100%

Based on these results, our analytical sensitivity to call SNVs and InDels at allele fractions of ~10% or higher is >95% at sites with read depths of 100X or greater.

Clinical Sensitivity

To determine the sensitivity of the MuTect algorithm to detect SNVs, and the Indelocator algorithm to detect small insertions and deletions, in a disease-associated sample set, six pairs of matched tumor-normal lung cancer patient samples were sequenced (Table 4). A number of SNVs and Indels in these patient samples had previously been validated as part of The Cancer Genome Atlas project (www.cancergenome.nih.gov).

Clinical Samples

Patient ID	Patient Gender	Primary Disease	Tumor Type
TCGA-78-7155	Male	Non Small Cell Lung Cancer, Adenocarcinoma	Primary
TCGA-95-7039	Female	Non Small Cell Lung Cancer, Adenocarcinoma	Primary
TCGA-55-7907	Male	Non Small Cell Lung Cancer, Adenocarcinoma	Primary
TCGA-44-7670	Female	Non Small Cell Lung Cancer, Adenocarcinoma	Primary
TCGA-05-4410	Male	Non Small Cell Lung Cancer, Adenocarcinoma	Primary
TCGA-05-4390	Female	Non Small Cell Lung Cancer, Adenocarcinoma	Primary

SNV Clinical Sensitivity

Sample	# Validated SNVs	# Called SNVs	MuTect Sensitivity
TCGA-78-7155	29	29	100%
TCGA-95-7039	33	33	100%
TCGA-55-7907	21	21	100%
TCGA-44-7670	17	17	100%
TCGA-05-4410	26	25	96%
TCGA-05-4390	13	12	92%

Indel Clinical Sensitivity

Sample	# Validated InDels	# Called InDels	Indelocator Sensitivity
TCGA-78-7155	40	34	85%
TCGA-95-7039	39	38	97%
TCGA-55-7907	21	20	95%
TCGA-44-7670	28	26	93%
TCGA-05-4410	21	20	95%
TCGA-05-4390	21	20	95%

In the samples studied, our clinical sensitivity to call SNVs and InDels (<10bp) is 99% and 93%, respectively.

Validating the next version of the clinical exome

This poster presents the validation of our original clinical exome process and pipeline. We are currently validating a new exome process and analytical pipeline. The v2 clinical exome at Broad will be available in March 2015 and will feature:
- Updated exome content - 15% more content, better coverage of medically relevant genes.
- Updated Germline Calling - Using GATK Haplotype Caller
- Updated LIMS and accessioning/reporting
- Somatic reports will include variant scoring/ranking using PHIAL[4]

Specificity

False positive rates are an appropriate measure of specificity in assays such as exome sequencing where the number of true negatives is large.

To determine the specificity of the MuTect and Indelocator algorithms to detect variants in a disease-associated sample set, two pairs of characterized tumor-normal cell lines were purchased from the American Tissue Culture Collection (ATCC).

Cell line name	Cell Type	Disease	ATCC #
HCC1143	Epithelial	Ductal breast carcinoma	ATCC CRL-2321
HCC1143 BL	B-cell Lymphoblast	Normal	ATCC CRL-2362
HCC1954	Epithelial	Ductal breast carcinoma	ATCC CRL-2338
HCC1954 BL	B-cell Lymphoblast	Normal	ATCC CRL-2339

In this study, false positive rates for SNVs and InDels were determined through examination of Normal-Normal calls. In the sequencing pools of HCC cancer cell lines, the normal, non-tumor sample was included three times through the entire process from sample preparation to sequence generation. In the analytical pipeline one of the normals was designated as a tumor sample and called against the other normal. This gives 6 pairwise comparisons per cell line, for a total of 12 replicates. Since both samples are the same any variants called are False Positives.

False Positive Rates per Mb of targeted territory

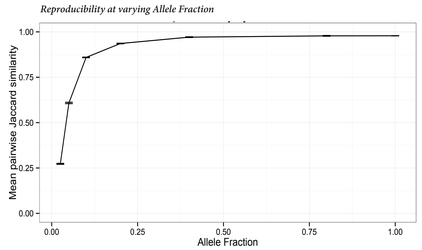
	InDels FPR/Mb	SNV FPR/Mb
Replicate 1	0.06	0.84
Replicate 2	0.15	0.78
Replicate 3	0.06	0.34
Replicate 4	0.15	0.56
Replicate 5	0.06	0.4
Replicate 6	0.12	0.43
Replicate 7	0.06	0.5
Replicate 8	0.03	0.81
Replicate 9	0.12	0.4
Replicate 10	0.21	0.65
Replicate 11	0.12	0.34
Replicate 12	0.18	0.34
Mean	**0.11**	**0.53**
S.E.M	0.01	0.05

Our false positive rate is less than 1/Mb.

Precision

The Jaccard index (or Jaccard similarity coefficient) was used as a metric to evaluate the reproducibility of the Whole Exome Sequencing and Somatic Analysis Test to call SNVs in replicate samples.
The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.
This value was computed for each pair of samples: Nine (9) 20-plex replicates for SNVs => 36 pairs

Reproducibility at varying Allele Fraction



Reproducibility at varying Allele Fraction

Low AF	High AF	Mean Reproducibility	SEM
0.02	0.03	27.30%	0.003
0.045	0.055	60.80%	0.004
0.095	0.105	85.90%	0.002
0.195	0.205	93.60%	0.002
0.395	0.405	97.10%	0.002
0.795	0.805	97.80%	0.002
0.995	1.005	97.80%	0.001

Based on these results, the SNVs called across the exome are >95% reproducible in samples with allele fractions of ~20% or greater.

Conclusions and lessons learned

Overall, this validation study demonstrates great somatic SNV calling and good Indel calling across whole exomes from matched tumor-normal pairs using Firehose tools.

Lessons learned: i) documenting research tools for clinical validation is difficult and the time required should not be underestimated; ii) For somatic calling both read depth and allele fraction must be considered when calculating sensitivity.

References
1. Rehm *et al.* ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 15, 733-47 (2013).
2. Gargis *et al.* Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 30, 1033–36 (2012).
3. Frampton *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 31, 1023–31 (2013).
4. Van Allen *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 20, 682-8 (2014).