TCGA Mutation Calling Benchmark 4

The latest version of this document will be available at:
http://hgwdev.soe.ucsc.edu/~ewingad/benchmark4/instructions.pdf

Mailing list subscribers will be notified when a new version has been posted.

1. Background and general instructions

TCGA has completed 3 benchmark exercises involving comparison of mutation calls from different centers. Benchmark 1 was whole genome and Benchmarks 2 and 3 were exome data. All benchmarks were restricted to point mutations. Benchmark 4 will be whole genome data and will be the first to use cell lines and the first to include participants from outside TCGA. In addition to the TCGA and ICGC centers, we encourage participation from all groups with an interest in calling mutations in the paired tumor/normal context of cancer genomics.

The overall purpose of this benchmark exercise is comparative evaluation of somatic mutation calls on single nucleotide variants (SNVs) and structural variants (SVs) under a variety of conditions designed to simulate the effects of tumor purity (i.e. normal contamination) and subclonal expansions in a controlled way. In addition, we can eventually know the ground truth as to the veracity of called mutations because the cell lines are publicly available, making all mutation calls amenable to experimental validation. As illustrated in Figure 2 and discussed further below, tumor purity is simulated by combining various fractions of tumor-derived and normal-derived cell line DNA, and subclonal expansions are simulated by generating artificial spike-in SNVs and SVs on the tumor genome background and combining the spike-in genome with the genomes of the original tumor and patient-matched normal cell line sequence. The overall goal is to provide feedback that will be useful in improving mutation calling pipelines that provide vital data for all areas of cancer genomics.

The data for the exercises are distributed as .bam files and are derived from two pairs of cell lines: HCC1143/HCC1143 BL and HCC1954/HCC1954 BL where 'BL' indicates the sample is the paired normal sample derived from blood. Both SNV and SV calls must be returned in VCF format.

The benchmark is divided into three main exercises described below. Each has a separate set of .bam files associated with it listed in a table for each exercise. We strongly encourage all participants to call SNVs and SVs on every dataset. In addition to the main benchmark exercises there are two others: low coverage and compression. Low coverage gives us the opportunity to evaluate mutation calling on low-depth whole genome sequence data (~7x average depth). The compression exercise will include BAM files that have been compressed using methods that lose some information to substantially reduce BAM file size and restored.

1.1. Downloading the data

All .bam files are distributed publicly through CGHub and should be used only for this benchmark exercise due to their artificial nature. To obtain the files for the benchmark exercise, navigate to: https://cghub.ucsc.edu/benchmark_download.html and follow the instructions for installing GeneTorrent and obtaining the public key. Further information on using GeneTorrent and CGHub can be found at the following URL: https://cghub.ucsc.edu/docs. The specific .bam files required for each part of the exercise are detailed below.

1.2. Timeline

Overview:

- Test submissions by Feb. 28th 2013
- Final submissions by March 31st 2013
- Initial results available May 2013

Sample submissions (e.g. calls on only chr20) received prior to Feb. 28th, 2013 will receive feedback on VCF structure and whether information included in the VCF is sufficient for all aspects of the comparison. Please e-mail ewingad@soe.ucsc.edu and describe any sample submissions you wish to have evaluated. We expect downloading, calling variants, and returning the results in VCF format to be completed no later than March. 31st, 2013. Initial comparative results will be available one month following this date, and based on these results a set of mutations from the cell lines will be selected for experimental validation via PCR with site-specific primers. We expect the validation phase to last another month with results available by the end of May 2013. Deeper sequencing on the cell lines genomes will be carried out during February and March.

1.3. Contacts

- For questions contact the mailing list: tcga-mutation@soe.ucsc.edu
- To join the mailing list contact Singer Ma: singer@soe.ucsc.edu
- For all other inquiries contact Adam Ewing: ewingad@soe.ucsc.edu

2. Mutation calls

In general, participants should follow the TCGA VCF 1.2 specification, which builds on the global VCF 4.1 spec by requiring certain fields and consistent nomenclature.

TCGA VCF 1.2 specification:
https://wiki.nci.nih.gov/display/TCGA/TCGA+Variant+Call+Format+%28VCF%29+1.2+Specification
VCF 4.1 specification:

http://www.1000genomes.org/wiki/Analysis/Variant\%20Call\%20Format/vcf-variant-call-format-version-41

We request that participants return both germline and somatic mutation calls in the same VCF file, using the reference genome (GRCh37) as the basis for comparison (i.e. the REF alleles). **It is very important to include all variants, germline and somatic in addition to all filtered variants (i.e. variants for which there is some evidence but which do not make the cut as reported mutations)**, annotating the filter(s) for each variant in the FILTER field. Somatic status is specified by the SS sub-field in the FORMAT column, and from adding SOMATIC to the INFO field, as described in Table 4a of the TCGA VCF 1.2 specification. The allele fraction for each mutation should be noted in the VCF file. This can be accomplished using the AD, DP, and FA (Fraction of reads supporting ALT) sub-fields in the FORMAT column (see Table 4b in the TCGA VCF 1.2 spec). Copy number of breakends should be specified using the CN sub-field in the FORMAT column.

At a minimum, we request that participants provide SNV calls. A full submission consists of SNVs, short indels (< 100bp), SVs (>=100bp), and CNVs. Guidelines for each class of mutation are as follows:

- SNV minor allele fraction (MAF) should be expressed in VCF via the AD sub-field specified in the FORMAT column e.g. 7,3 would indicate 7 REF reads and 3 ALT reads for a MAF of 0.3. Immediately consecutive SNVs (e.g. AT → GC) should be expressed as two separate records (one A → G and a second for T → C).

- Insertions and deletions (INDELs) smaller than 100bp should be expressed using the VCF indel notation, i.e. if REF is G and ALT is GTAC, there was an insertion of TAC after G. Allele fraction should be specified the same way as for SNVs. INDELs should be left-aligned (Figure 1); this can be accomplished by running GATK LeftAlignVariants: http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantutils_LeftAlignVariants.html

- CNVs should be expressed using <CNV> as the ALT allele. Start and end positions for a CNV segment are expressed via the POS field and END sub-field in INFO. Copy number is expressed using the CN FORMAT sub-field, and should be in terms of absolute copy number values (e.g. 2 = diploid, 3 = triploid).

- SVs should be expressed as individual breakends using the VCF breakend description for the ALT allele and adding SVTYPE=BND to the INFO column. If a precise breakend is not known, the phrase IMPRECISE should be in the INFO column. The somatic status ('SS') field can be used to indicate somatic status as well; this should be done in addition to the 'IMPRECISE' sub-field. As with CNVs, confidence intervals for imprecise variants should be given using the CIPOS sub-field, and copy number expressed using the CN sub-field. Precise breakends, if available, should be left-shifted similar to indels to avoid

missing concordant mutations calls due to ambiguous breakends. This can be accomplished with leftShiftBreakends.py, provided by VCFcomparator: https://github.com/adamewing/VCFcomparator/blob/master/etc/leftShiftBreakends.py

All variant calls should include a score indicating confidence that a given variant is somatic. This should be a monotonic function distinct from the quality score, designated as SSC in the FORMAT field. The following should be included in the header:

```
##FORMAT=<ID=SSC,Number=1,Type=Integer,Description="Somatic Score">
```

Phasing information may be present, as it can be described in VCF format, but will not be considered as part of a scoring function.

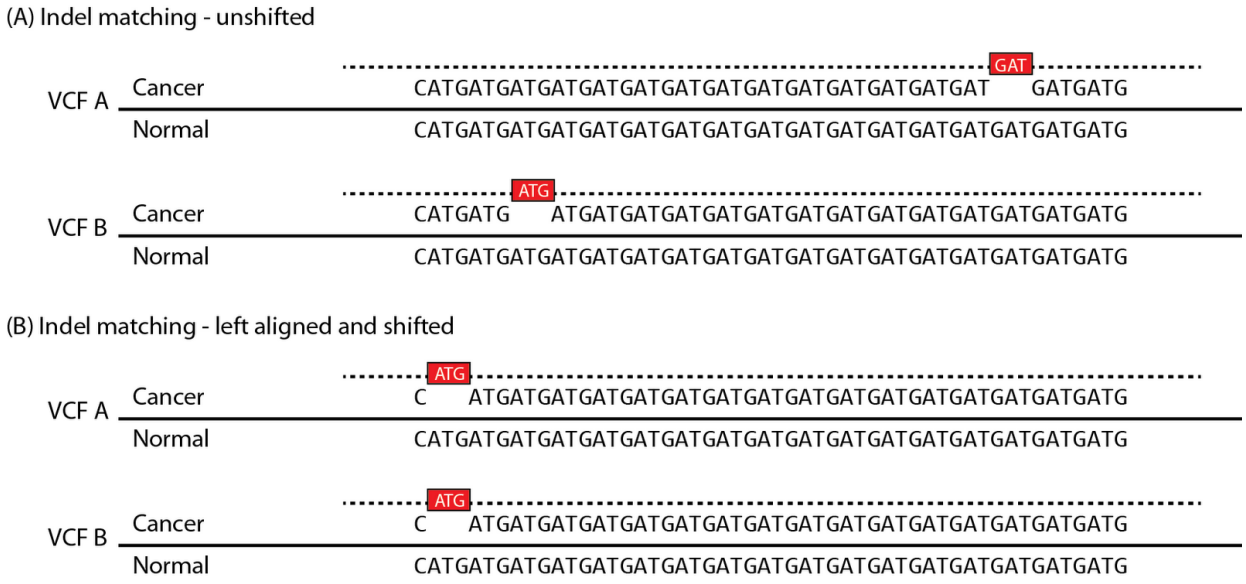

Figure 1: Left alignment of ambiguous INDELs. (A) Two equivalent deletion calls, 'GAT' and 'ATG' in a short tandem repeat that would be considered to be different calls by a direct comparison. (B) Left alignment of indels converts these each to the leftmost equivalent mutation. Now they are identical, and so will be recognized as equivalent mutation calls.

2.1. Submitting Results

- Mutations should be returned as one VCF file per tumor/normal comparison. Thus, a complete submission of parts 0, 1, and 2 will consist of 24 VCF files: 2 VCFs for Part 0, 12 VCFs for Part 1, and 10 VCFs for Part 2.

- Mutation calls in VCF format will be uploaded via anonymous FTP to ftp://kolossus.soe.ucsc.edu along with a data dictionary describing the correspondence between VCF files and .bam file comparisons. Filenames should indicate the submitting participant. After the upload is complete, please email ewingad@soe.ucsc.edu to let us

know that new files have been submitted.

3. Evaluating mutation calls

3.1. Comparing mutation calls

VCFs returned via FTP will be compared to one another using the VCF comparator described in the appendix of this document. This will yield results on the level of individual mutation calls and an overall score for the comparison between each possible pairing of mutation type and VCFs from two centers. This will enable us, for example, to use clustering methods to identify groups of mutation callers with common characteristics on different mutation types.

For parts 0 and 1 of the benchmark, PCR validation will allow one to conclusively determine whether a given mutation is false positive (FP). We will validate a subset of calls made only by a single-participant as well as calls made concordantly by a subset of *n* participants for *n= 2, 3, ...* to produce a figure of FP rate versus concordance of mutation calls. The spike-in subclonal mutations in part 2 allow exploration of the false-negative (FN) rate, which has not been possible for previous TCGA benchmarks. Examples of non-concordant calls, false positive calls, and false negative calls can then be explored in detail to identify actionable explanations.

Results will be available both as a detailed list of all *n*-concordant variants, and as a summary that details the detection characteristics of each algorithm in terms of FP and FN rate for SNVs, Indels, SVs, and CNVs (if applicable). Additional information about concordant variants will include (dis)agreement on filter status, and (dis)agreement on somatic status. Further details on the VCF comparator can be found in the appendix.

4. Benchmark Exercises

4.1. Part 0: High coverage tumor vs. normal cell lines

This exercise consists of two comparisons: (1) HCC1143 vs. HCC1143 BL and (2) HCC1954 vs. HCC1954 BL. HCC1143 and HCC1954 are both derived from grade 3 breast ductal carcinomas and HCC1143 BL / HCC11954 BL are the corresponding patient matched normal samples derived from blood.

Table 1: Files for Part 0

Description	.bam file	Coverage
HCC1143	G15511.HCC1143.1.bam	~50x
HCC1143 BL	G15511.HCC1143_BL.1.bam	~60x

HCC1954	G15512.HCC1954.1.bam	~58x
HCC1954 BL	G15512.HCC1954_BL.1.bam	~71x

4.2. Part 1: Tumor/normal mixtures vs. normal

This exercise simulates varying levels of normal contamination in tumor samples. There are two sets of six comparisons as illustrated in Figure 2. Each tumor/normal mixture will be compared against the 30x normal sample (distinct from the normal sample used in the mixtures) to identify somatic SNVs and SVs for a total of 12 comparisons: 6 HCC1143 Normal vs. HCC1143 Tumor/Normal Mixture comparisons and 6 HCC1954 Normal vs. HCC1954 Tumor/Normal Mixture comparisons.

Table 2: Files for Part 1

Description	**.bam file**	**Coverage**
HCC1143 Normal	HCC1143.NORMAL.30x.compare.bam	~30x
HCC1954 Normal	HCC1954.NORMAL.30x.compare.bam	~30x
HCC1143 5% Normal, 95% Tumor	HCC1143.mix1.n5t95.bam	~30x
HCC1143 20% Normal, 80% Tumor	HCC1143.mix1.n20t80.bam	~30x
HCC1143 40% Normal, 60% Tumor	HCC1143.mix1.n40t60.bam	~30x
HCC1143 60% Normal, 40% Tumor	HCC1143.mix1.n60t40.bam	~30x
HCC1143 80% Normal, 20% Tumor	HCC1143.mix1.n80t20.bam	~30x
HCC1143 95% Normal, 5% Tumor	HCC1143.mix1.n95t5.bam	~30x
HCC1954 5% Normal, 95% Tumor	HCC1954.mix1.n5t95.bam	~30x
HCC1954 20% Normal, 80% Tumor	HCC1954.mix1.n20t80.bam	~30x
HCC1954 40% Normal, 60% Tumor	HCC1954.mix1.n40t60.bam	~30x
HCC1954 60% Normal, 40% Tumor	HCC1954.mix1.n60t40.bam	~30x
HCC1954 80% Normal, 20% Tumor	HCC1954.mix1.n80t20.bam	~30x
HCC1954 95% Normal, 5% Tumor	HCC1954.mix1.n95t5.bam	~30x

4.3. Part 2: Tumor/normal/subclone mixtures vs. normal

This exercise simulates subclonal expansions that contain novel mutations (both SVs and SNVs), ~500 SNVs and ~200 SVs were spiked in to each subclone using a method that allows for the near-seamless addition of novel SNVs and SVs, including indels, to be added to existing .bam files (see https://github.com/adamewing/bamsurgeon for details.). Structural variants include insertions, deletions, inversions, duplications, compound events, and small indels. Subclonal mutations were added to appear heterozygous in the subclone, and the subclone was combined with reads from tumor and normal .bams to create a subclone spike-in mixture. The comparisons will be similar to those in Part 1, with 5 HCC1143 Normal vs. HCC1143 Tumor/Normal/Subclone Mixture comparisons and 5 HCC1954 Normal vs. HCC1954 Tumor/Normal/Subclone Mixture comparisons for a total of 10 comparisons in this exercise.

Table 3: Files for Part 2 (all samples ~30x)

Description	**.bam file**
HCC1143 Normal	HCC1143.NORMAL.30x.compare.bam
HCC1954 Normal	HCC1954.NORMAL.30x.compare.bam
HCC1143 25% Normal 74% Tumor 1% Subclone	HCC1143.spiked1.n25t74s1.bam
HCC1143 25% Normal 70% Tumor 5% Subclone	HCC1143.spiked1.n25t70s5.bam
HCC1143 25% Normal 65% Tumor 10% Subclone	HCC1143.spiked1.n25t65s10.bam
HCC1143 25% Normal 55% Tumor 20% Subclone	HCC1143.spiked1.n25t55s20.bam
HCC1143 25% Normal 35% Tumor 40% Subclone	HCC1143.spiked1.n25t35s40.bam
HCC1954 25% Normal 74% Tumor 1% Subclone	HCC1954.spiked1.n25t74s1.bam
HCC1954 25% Normal 70% Tumor 5% Subclone	HCC1954.spiked1.n25t70s5.bam
HCC1954 25% Normal 65% Tumor 10% Subclone	HCC1954.spiked1.n25t65s10.bam
HCC1954 25% Normal 55% Tumor 20% Subclone	HCC1954.spiked1.n25t55s20.bam
HCC1954 25% Normal 35% Tumor 40% Subclone	HCC1954.spiked1.n25t35s40.bam

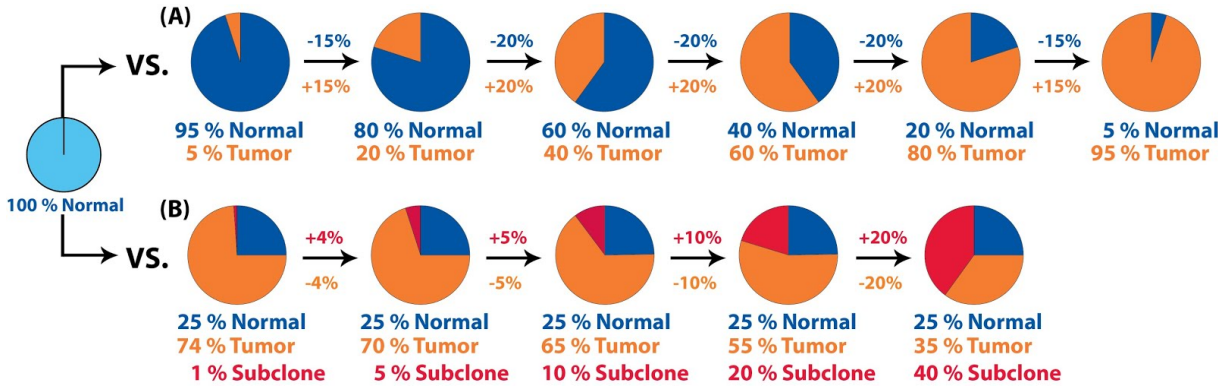

Figure 2: Mutation calling workflow for parts 1 and 2. Normal 30x .bam is compared to all mixed .bams and subclone spike-in .bams. Blue represents the contribution of sequence derived from the normal cell line (HCC1143 BL or HCC1954 BL), orange represents sequence derived from the tumor cell line (HCC1143 or HCC1954) and red represents sequence derived from the tumor cell line with spiked-in mutations to simulate a subclone (different sets of mutations were added to HCC1143 and HCC1954). Note that the 30x normal .bam being compared against (light blue) is a distinct set of reads from the normal fraction of the mixed .bam files (dark blue).

4.4 Other exercises: Low coverage benchmark

Table 4: Files for low coverage benchmark (all samples ~7x)

Description	.bam file
HCC1143 Normal	HCC1143.NORMAL.7x.compare.bam
HCC1954 Normal	HCC1954.NORMAL.7x.compare.bam
HCC1143 25% Normal 65% Tumor 10% Subclone	HCC1143.lowcover.7x.n25t65s10.bam
HCC1143 25% Normal 55% Tumor 20% Subclone	HCC1143.lowcover.7x.n25t55s20.bam
HCC1954 25% Normal 65% Tumor 10% Subclone	HCC1954.lowcover.7x.n25t65s10.bam
HCC1954 25% Normal 55% Tumor 20% Subclone	HCC1954.lowcover.7x.n25t55s20.bam

High-coverage whole genome sequencing is still prohibitively expensive for the time being, therefore we are interested in how well mutations can be called from low coverage data (~7x average). Two normal/tumor/subclone mixtures were created with different spiked-in mutations than the corresponding higher-coverage mixtures from part 2. This exercise consists of four comparisons: HCC1143 Normal vs. both HCC1143 normal/tumor/subclone mixtures and the same for HCC1954.

4.5 Other exercises: Mapping benchmark

The choice of mapping algorithm can have a significant impact on mutation calls. To quantify this effect, participants in the mapping benchmark will dump paired BAM files to .fastq format, re-map the reads using their aligner, and submit the resulting BAM file. Mutation calling output on the original BAMs will be compared with mutations called on the re-mapped BAMs.

FASTQ extraction should be performed using SamtoFastq in the picard suite:
http://picard.sourceforge.net/command-line-overview.shtml#SamToFastq

BAM files must validate with no errors via ValidateSamFile:
http://picard.sourceforge.net/command-line-overview.shtml#ValidateSamFile

At a minimum, we request that the high-coverage samples (Table 1 / Part 0) be used for this exercise.