Mutascope
Reference Manual

Last updated 4/9/13

Mutascope Version 1.0.0

Authors:

Shawn Yost <yostshawn@gmail.com>

Olivier Harismendy, PhD <oharismendy@ucsd.edu>

Table of Content

Overview

Mutascope is a software suite designed to analyze data from high throughput sequencing of PCR amplicons, with an emphasis on normal-tumor comparison for the accurate and sensitive identification of low prevalence mutations.

PCR amplicons sequencing is a very specific application of high-throughput sequencing. In contrast to common applications such as whole genome or whole exome sequencing, the sequenced fragments are not randomly distributed along the genome, but have fixed start and end coordinates corresponding to the amplicon and sequencing read length. These characteristics have direct consequences on the error profile along the amplicons, which are not accounted for by common genotype callers. **Mutascope** exploits this characteristic to improve variant calling in high throughput sequencing of PCR amplicons[1]. For example, **Mutascope** assigns a read-group to each sequenced read that corresponds to the PCR amplicon and strand it comes from. This information is then used when calculating specific error rates to call significant variants with as low as 1% frequency in the tumor.

1.1. Main Features

- Read alignment and grouping
    - Leverages BWA to align the reads
    - Assigns each read to a read group based on the amplicon of origin and the sequenced read set (read1 or read2) in a paired end sequencing run
    - Selects only reads originating from amplicons boundaries
- Accurate Variant Calling
    - Measures the error rate at each read position, substitution type and read type, using normal DNA reads
    - Calculates binomial probability for the presence of a variant in the tumor
    - Classifies germline and somatic mutation using Fisher exact test
    - Calls germline genotypes using a Bayesian probabilistic approach
    - Excludes potential false positive mutations from non-specific PCR products
- Modular Structure allowing to run some or all analysis steps
- Quality Control
    - Generates a set of quality metrics (alignment and coverage statistics)
    - Calculates the coverage ratio between tumor and normal to identify potential copy number aberrations.
    - Plots the allele frequency of both germline and somatic variants in the tumor relative to normal sample to identify potential sample swap/mixing.
- Comprehensive VCF
    - Somatic and germline classification
    - Somatic quality score
    - Germline genotype quality score
    - Quantitative and qualitative filtering including
        - Somatic Fisher-Exact P-value
        - Low prevalence Binomial P-value
        - Presence of homopolymer

1.2. Availability

Mutascope relies on tools developed by others such as SAMTools[2], BWA[3,4], Picard Tools UCSCTools [5], R [6], and GATK[7] (Section 2.1).

Mutascope is available through the sourceforge repository (https://sourceforge.net/projects/mutascope/files/)

2. Installation

To install **Mutascope** unzip '**Mutascope**.zip' file and follow the INSTALL instructions.

2.1. Prerequisite Tools

The following software is required to be installed on your system before you run **Mutascope**:

- Perl: v5.12.3
- R: v2.15.1
- MCLUST (R-package): v4.0
- SAMTools: v0.1.18
- BWA: v0.5.9-r16 – v0.5.10

2.2. Prerequisite Datasets

- **REF.fa**: Human genome: fasta file and BWA index. BWA index is generated by:

  ```
  bwa index -a bwtsw REF.fa
  ```

- **REF.2bit**: Human genome 2bit: This file should be called `REF.2bit` (if the reference is called `REF.fa`). Generated using the UCSCTools[5] script '`faToTwoBit`' provided in the 'scripts/' folder.
- **dbSNP.vcf:** VCF formatted version of the dbSNP database. A VCF formatted dbSNP can be found in the GATK bundle[7]. Follow these instructions to access the FTP server and obtain the latest version of the bundle.

3. Quick Start

This Quick Start procedure is specific for Illumina paired end sequencing of a tumor-normal pair. It will let you go directly from a set of 4 fastq files to a VCF file containing germline and somatic variants.

```
Mutascope.pl runPipeline -bed amplicons.bed -fasta REF.fa -dbsnp dbsnp.vcf
    -pdir ./projectname/ -normal STRING -tumor STRING -nr1_fastq FASTQ –
    nr2_fastq FASTQ -tr1_fastq FASTQ -tr2_fastq FASTQ
```

Where:

- `STRING` is a character string you choose to name normal and tumor output files, respectively. Only letters and numbers permitted. No special characters.
- `FASTQ` corresponds to one of the 4 fastq files: Normal reads 1 & 2 or tumor reads 1 & 2.

It also assumes that:

- The prerequisite packages in Section 2.1 are in your path (besides the MCLUST package).
- You have created the project directory `projectname/`.
- You have a list of the amplicons in a BED format (Section 4.1).
- You have prepared a BWA indexed genome in the same folder as REF.fa (Section 2.2)
- You have prepared a 2bit genome in the same folder as REF.fa (Section 2.2)

4. General Usage

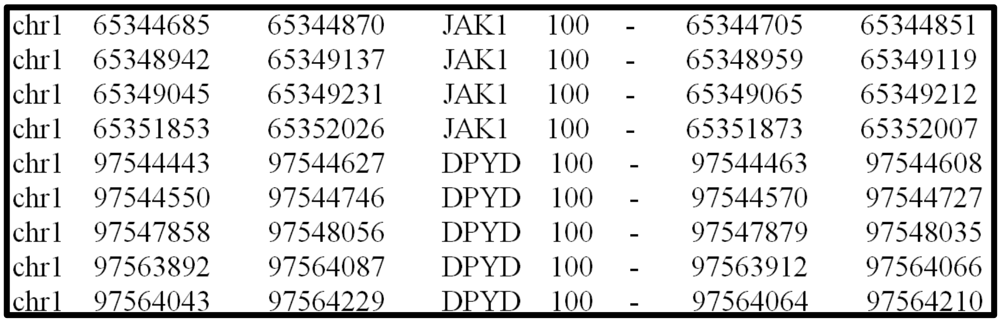
4.1. Required input files

BED

This file contains the list of PCR amplicons, one amplicon per line in a format derived from the UCSC BED standard. For **Mutascope** it has the following 8 values per row.

```
CHR START END AMPID Score Strand WOPSTART WOPEND
```

Which corresponds to Chromosome [CHR], Amplicon start [START], Amplicon end [END], Amplicon Name [AMPID], Arbitrary number [Score], Strand of the genome that is amplified [Strand], Amplicon without primer start [WOPSTART] and Amplicon without primer end [WOPEND] (Figure 1). The *score* field is not used by **Mutascope**, and can be set arbitrarily. The *strand* field is used when specifying that the sequenced reads are strand specific. The *ampid* field is used as the amplicon name.



chr1	65344685	65344870	JAK1	100	-	65344705	65344851
chr1	65348942	65349137	JAK1	100	-	65348959	65349119
chr1	65349045	65349231	JAK1	100	-	65349065	65349212
chr1	65351853	65352026	JAK1	100	-	65351873	65352007
chr1	97544443	97544627	DPYD	100	-	97544463	97544608
chr1	97544550	97544746	DPYD	100	-	97544570	97544727
chr1	97547858	97548056	DPYD	100	-	97547879	97548035
chr1	97563892	97564087	DPYD	100	-	97563912	97564066
chr1	97564043	97564229	DPYD	100	-	97564064	97564210

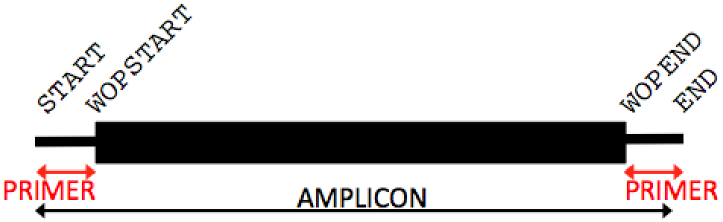


Figure 1: (Top) schematic representation of an amplicon, its primer and the location of START, WOPSTART, END and WOPEND values. (Bottom) Example of a properly formatted BED file for use in Mutascope.

FASTQ

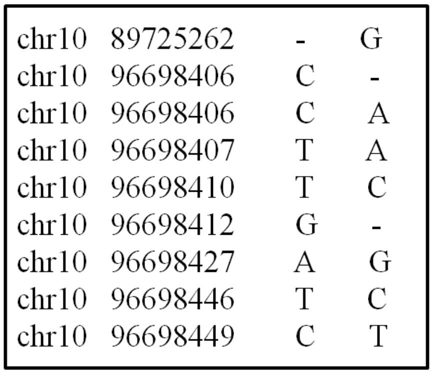
A set of four fastq files, two for the sequence of normal DNA, two for the sequence of tumor DNA are required by **Mutascope**. For each sample, **Mutascope** needs the first set of reads sequenced (read1) and the second set of reads sequenced (read2) in a paired-end sequencing run. Currently **Mutascope** only works with paired-end sequencing data.

4.2. Intermediate Files

Intermediate files are generated by the different **Mutascope** modules and saved in the `intermediate/` folder.

Blacklist (`.blacklist`)

List of positions identified by the module [`makeBlackList`] as more likely to be false positive bases on in silico PCR. A blacklist file is a list of mutations that are excluded from analysis in **Mutascope**. As an example using a set a set of 1676 amplicons that cover 150kb of the genome, ~0.08% of all possible substitutions were removed from analysis. The file format is TAB delimited: Chromosome, Position, Reference base, Alternate Base. A '-' is used in the *alternate base* (resp. *reference base*) field if there is a nucleotide deletion (resp. insertion) at that position.



```
chr10	89725262	-	G
chr10	96698406	C	-
chr10	96698406	C	A
chr10	96698407	T	A
chr10	96698410	T	C
chr10	96698412	G	-
chr10	96698427	A	G
chr10	96698446	T	C
chr10	96698449	C	T
```

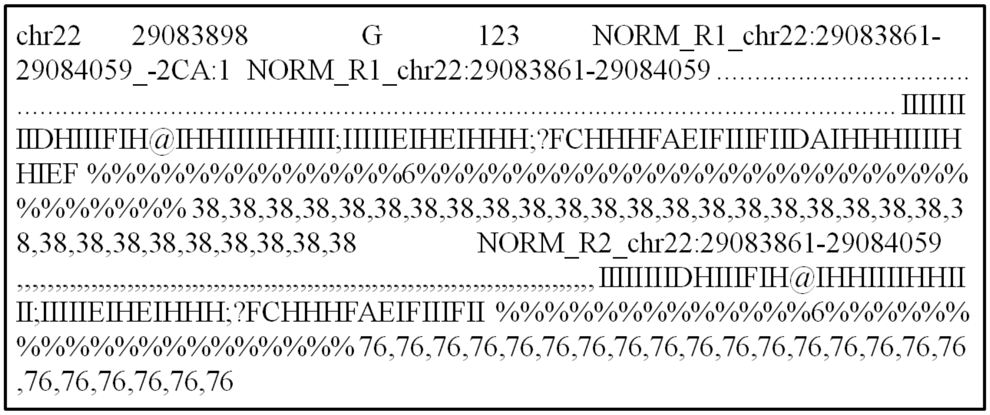
Figure 2: Example of a blacklist file formatted for Mutascope. The file is generated by the makeBlackList module or by the runPIpeline module when run for the first time.

Extended Pileup (`.xpileup`):

Extended pileup file. An extended pileup file that is generated by the `xpileup` module. Each row is tab-delimited with the following fields:

- chromosome
- position
- reference base
- total coverage
- *indel information*: a string containing space separated information about indels in each read group. This field is left blank if there is no indel at the given position. The indel info for each read group is formatted as a underscore delimited string: [sample]_[read type]_[ampliconID as chr:start-stop]_[indel info], where [indel info] is the number of reads containing that indel, repeated for each indel starting at the given position (+ for insertion/- for deletion, indel length, indel sequence, ":", total coverage). For example "+1A:16" indicates an insertion of one "A" in a total of 16 fold read coverage.
- *read information:* a tab delimited string containing the information in each read group. For each read group the information is space delimited as:
    - Read group ID: [sample]_[read type]_[ampliconID as chr:start-stop]
    - Base sequence (SAMtools pileup format)
    - Base Quality (SAMtools pileup format)
    - Mapping Quality (SAMtools pileup format)
    - Position in the read (SAMtools pileup format)

Note: Do not use a regular pileup file when running '**Mutascope**.pl `callSomatic`'



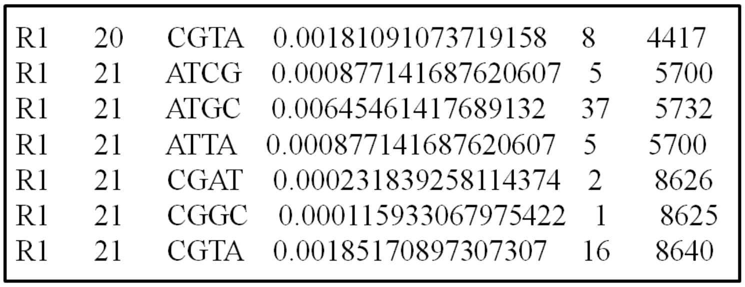
```
chr22	29083898	G	123	NORM_R1_chr22:29083861-
29084059_-2CA:1  NORM_R1_chr22:29083861-29084059 ...................................
........................................................................................................................ IIIIII
IIDHIIIFIH@IHHIIIIHHIII;IIIIIEIHEIHHH;?FCHHHFAEIFIIIFIIDAIHHHIIIIH
HIEF %%%%%%%%%%%%6%%%%%%%%%%%%%%%%%%%%%
%%%%%%% 38,38,38,38,38,38,38,38,38,38,38,38,38,38,38,38,38,38,38,38,38,3
8,38,38,38,38,38,38,38,38,38		NORM_R2_chr22:29083861-29084059
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, IIIIIIIDHIIIFIH@IHHIIIIHHII
II;IIIIIEIHEIHHH;?FCHHHFAEIFIIIFII %%%%%%%%%%%6%%%%%%
%%%%%%%%%%%%% 76,76,76,76,76,76,76,76,76,76,76,76,76,76,76,76,76
,76,76,76,76,76,76
```

Figure 3: Example of an xpileup formatted file

Error Rates `(.errorRate)`

ErrorRates files contain the error rate information from the normal sample. It is a TAB delimited file. Each row consists of the following 6 fields:

- The first three fields define the class of nucleotide substitutions evaluated depending on
  - Read type (either R1 or R2)
  - Position (position in the read)
  - Mutation Type (the type of mismatch: CGTA == C → T OR G → A)
- The next 3 fields are calculated from the sample `.xpileup`, masked for dbSNP and blacklist substitutions:
  - Error Rate: the average fraction of mismatches in the class
  - Number of Mismatches in the class
  - Total Number of Bases in the class



R1	20	CGTA	0.00181091073719158	8	4417
R1	21	ATCG	0.000877141687620607	5	5700
R1	21	ATGC	0.00645461417689132	37	5732
R1	21	ATTA	0.000877141687620607	5	5700
R1	21	CGAT	0.000231839258114374	2	8626
R1	21	CGGC	0.000115933067975422	1	8625
R1	21	CGTA	0.00185170897307307	16	8640

Figure 4: Example of an errorRate file

4.3. Results Files

.VCF

Following the VCF 4.1 format, specific FORMAT and INFO fields have been added. Various FILTERS are also preset by **Mutascope** and are featured in the VCF file.

FIELD	NAME	Description
QUAL		-$\log_{10}$(Binomial P-value)
INFO	DP	Read Depth
INFO	BQ	Average RMS base quality of all samples
INFO	MQ	Average RMS mapping quality of all samples
INFO	VT	Variant type, can be SNP, INS, or DEL
INFO	ER	Error rate used in the `callSomatic` module
INFO	FEP	Somatic P-value calculated in the `callSomatic` module
INFO	GQ	-$\log_{10}$ of the Genotype Quality Score: 2nd Best Score – Best Score
INFO	PL	-$\log_{10}$ of the Bayesian likelihood for REF,HET,ALT genotypes
FORMAT	GT	Genotype
FORMAT	AF	Allele Frequency
FORMAT	DP4	Number of ref-forward, ref-reverse, alt-forward and alt-reverse bases
FORMAT	AD	Ref,Alt coverage
FORMAT	SS	Variant type relative to non-adjacent Normal: 0=wildtype, 1=germline, 2=somatic, 3=LOH, 4=post transcriptional modification, 5=unknown
FORMAT	BQ2	REF RMS base quality and ALT RMS base quality
FORMAT	MQ2	REF RMS mapping quality and ALT RMS mapping quality
FORMAT	NR	The total number of read groups overlapping the position
FORMAT	DS	The distance to the second allele (#ALT1-#ALT2)/SUM(#ALT)
FORMAT	ST	The number of strands supporting the alternate allele (1 or 2)
FORMAT	RGB	Read Group Bias Score for the alternate allele using a chi-squared test
FILTER	MIF	Multiple INDELs called at a single position
FILTER	HRUN	≥5bp homopolymer repeat within 1bp of the start or end of the INDEL
FILTER	MCF	Low coverage (Normal = X and Tumor = Y)
FILTER	SBFILT	Variant shows Strand Bias
FILTER	RGBF	Variant shows Read-Group Bias
FILTER	DSF	First and second alternate allele frequencies are too close (DS < 0.5)
FILTER	MAMQ	Low average Mapping Quality score
FILTER	MFEP	Low Fisher-exact P-value for Somatic variants (FEP < 0.0005)
FILTER	MBPF	Low Binomial P-value to call the position a variant (QUAL < 54)
FILTER	MAABQ	Low average alternate allele Base Quality score
FILTER	CNIF	The SNP is within 10bp of an Indel
FILTER	SNPCF	The SNP is within 10bp of 2 other SNPs
FILTER	ICF	The Indel is within 10bp of 1 other Indel

ampliconLogR.txt

LogR files are tab-delimited with the following columns: CHR (chromosome), START (start position of the amplicon), STOP (stop position of the amplicon), AMPLICON (amplicon name in the BED file), NORMAL_RC (number of reads aligning to this amplicon in the NORMAL sample), TUMOR_RC (number of reads aligning to this amplicon in the TUMOR sample), and LOGR (logR ratio of tumor vs. normal).

```
#CHR     START       STOP     AMPLICON NORMAL_RC TUMOR_RC LOGR
 chr1  43814982   43815163     MPL1        59        1357    4.36
 chr1  115256500 115256680    NRAS1      2177        1229   -0.99
 chr1  115258702 115258884    NRAS8      1188        4589    1.78
chr10  43609048   43609226     RET1      1969        1802   -0.29
chr10  43609902   43610075     RET2      2339        4773    0.86
```

Figure 5: Example of an ampliconLogR file

4.4. Quality Files

alignment_stats.txt

The alignment stats file reports the number of reads passing the sequential analysis steps:

- R1/2_sequenced_reads: total number of reads in the fastq file
- R1/2_mapped_reads: Number of reads mapped by BWA
- R1/2_uniquely_aligned_reads: Number of reads BWA aligned to a unique position
- R1/2_reads_with_high_SW_score: Number of reads with a minimum SW score determined by the *refinement* module).
- R1/2_mapped_at_expected_location: Number of reads mapping within the set distance from the amplicon boundary

Among this last set of reads, the file indicates further:

- R1/2_with_indel: N of reads containing a gap in the alignment
- R1/2_with_soft_clipping: N of reads with bases clipped by BWA. An excess of clipped bases can indicate that the reads contain adaptors or are of low quality.

readsPerAmplicon.txt

The reads per amplicon file contains the number of reads aligning to each amplicon in the BED file. This value is similar to classic coverage depth, except were reads (R1 and R2) overlap. The file is tab-delimited with the following columns; chromosome, start of amplicon, stop of amplicon, amplicon name, number of reads assigned to that amplicon.

readsPerAmplicon_cov2X.txt

This file indicates the fraction of amplicons covered by a number of reads within two fold (>0.5 or <2) of the average N of reads per amplicons in that sample. This measure of uniformity is referred to as cov2x.

readsPerAmplicon_meanSD.txt

This file indicates the mean and standard deviation of the number of reads per amplicon.

readsPerAmplicon_sensitivity.txt

This file indicates the number of amplicons (first row) and fraction of amplicons (second row) with fewer than 50, 100, 500, and 1000 (cumulative) and greater than or equal to 1000 reads per amplicon.

readsPerAmplicon_cumulativeCoverageDist.pdf

This plot displays the cumulative distribution of the normalized number of reads per amplicon. The read count for each amplicon is normalized by dividing by the average number of reads per amplicon for that sample.

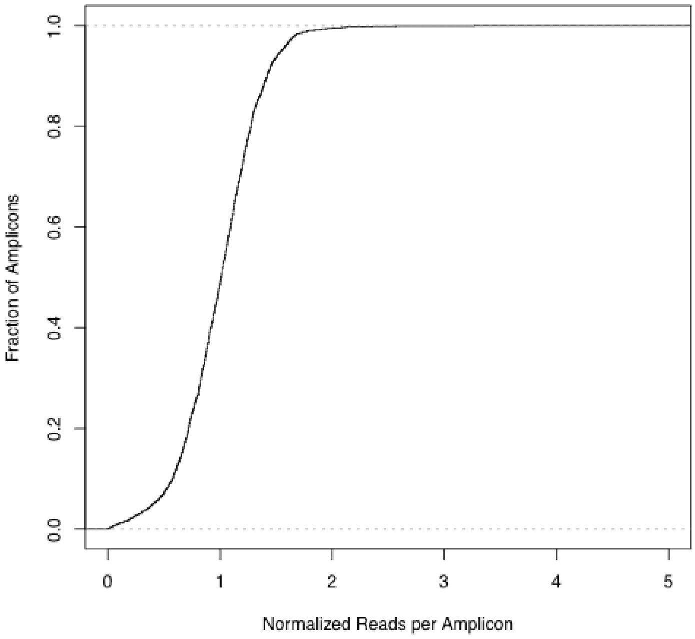


Figure 6: Example of a cumulativeCoverageDist plot

NORMAL_vs_TUMOR_filtered_variantPlot.pdf

This scatter plot displays the allelic fraction of a somatic (red) or germline (blue) variants (circle: SNV, square: indels) in the tumor (Y axis) in comparison to the normal DNA (X axis). The title of the plot indicates the total number of germline and somatic variants called and that passed the filters

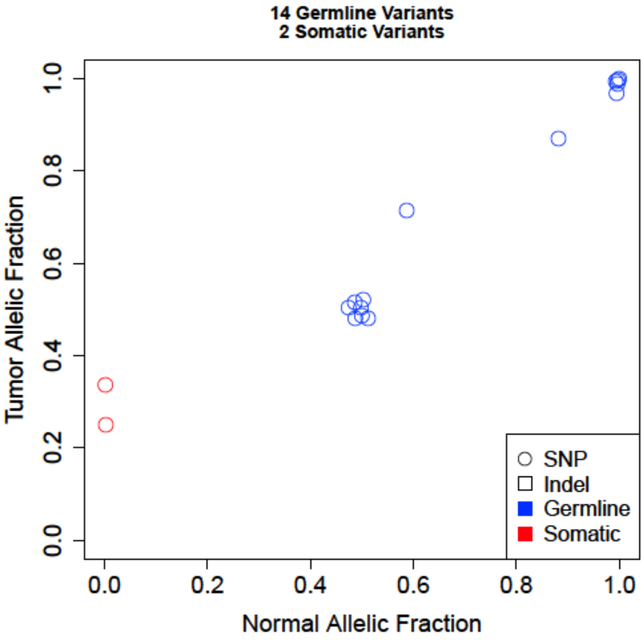


Figure 7: Example of a variant plot where SNPs and Indels are plotted as circles and squares, respectively, and germline and somatic variants are blue and red.

4.5. Directory Structure

The different **Mutascope** modules are relying on the following, expected directory structure. This structure is established in the working directory when **Mutascope** `RunPipeline` is run in the first time.

o `intermediate/`	where intermediate files like bam, errorRate, .xpileup are written
o `results/`	where VCF and ratios are written
o `qualities/`	where quality control metrics are written

4.6. General Flow

Mutascope consists of a set of 8 modules written in Perl. The `RunPipeline` module is a master module which, when called, runs the seven other modules sequentially (Figure 8). Alternatively, each of the seven modules can be called separately, provided the correct input file and directory structure are present. Currently **Mutascope** is designed to write all output files in a PROJECT directory created in advance. **Mutascope** modules expect that certain files generated by the previous steps exist in the PROJECT directory (as indicated by the dashed arrows from one module to another).

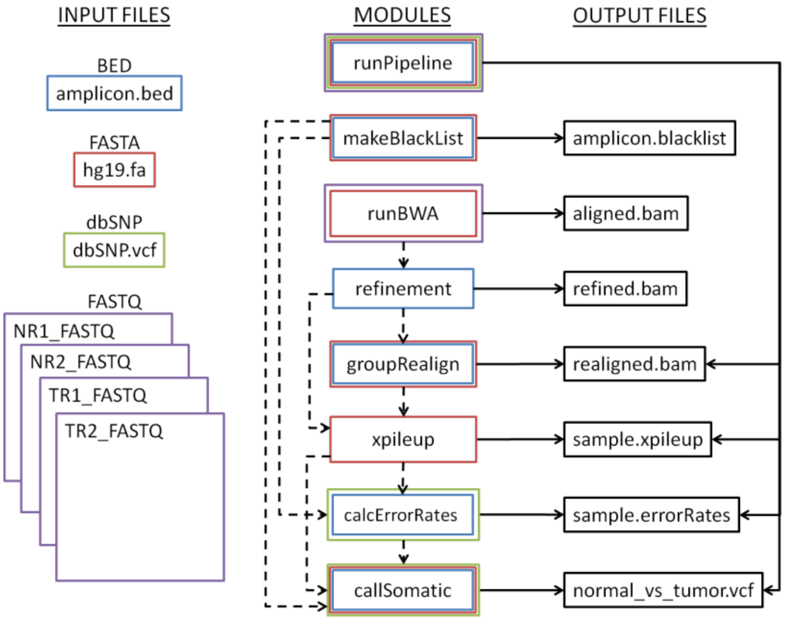


Figure 8: General workflow of **Mutascope**. The three columns consist of the input and output files as well as the **Mutascope** modules. The modules are framed by different colors indicating the required input files. The continuous arrows indicate intermediate or result files generated by the module. The dashed arrows indicate the different sequence of modules that are possible. For example, `callSomatic` requires that you ran and generated the output files from the `makeBlackList`, `xpileup`, and `calcErrorRates` modules.

Full Pipeline

`Mutascope.pl runPipeline` (Section 5.1) runs the full **Mutascope** pipeline from fastq to VCF. This is the preferred mode of operation.

Stepwise Modules

1. `Mutascope.pl makeBlackList` (Section 5.2): This module identifies the set of mutations to ignore during variant calling. These black listed mutations are a set of mutations that can be caused by misalignments of reads to homologous regions in the genome. This module needs to be run once for each set of PCR amplicon (BED file).
2. `Mutascope.pl runBWA` (Section 5.3): This module aligns reads using BWA's Smith-Waterman alignment algorithm.
3. `Mutascope.pl refinement` (Section 5.4): This module refines the alignment/BAM files generated in step 2 to assign a read group to each read, remove low quality aligned reads and soft-clip primer sequences. The read group given to each read assigns it to a specific PCR amplicon in the BED file. This read group is used when calculating error rates and calling variants. The module also generates a set of alignment and reads per amplicon quality metrics.

4. `Mutascope.pl groupRealign` (Section 5.5): This module uses GATK to realign together reads around indels from both the tumor and normal samples. This step is optional but recommended.
5. `Mutascope.pl xpileup` (Section 5.6): Generates the xpileup file for both the tumor and normal sample from the files generated in step 3 or 4.
6. `Mutascope.pl calcErrorRates` (Section 5.7): Calculates the error rates of the normal and tumor samples from the files generated in step 5.
7. `Mutascope.pl callSomatic` (Section 5.8): Calls and filters variants using the xpileup files generated in step 5 and the error rates generated in step 6. This module also generates the logR ratio for each amplicon and a plot of the germline and somatic variants passing all the filters.

5. Specific Usage

5.1. runPipeline

`runPipeline` is the master module used to run the full **Mutascope** pipeline. This module requires the following module line options: -bed, -fasta, -dbsnp, -pdir, -normal, -tumor, -nr1_fastq, -nr2_fastq, -tr1_fastq, and -tr2_fastq.

Command	**Format**	**Option**	**Description**
-bed	BED	**REQUIRED**	A file in BED format containing the PCR amplicon information (Section 4.1)
-fasta	FASTA	**REQUIRED**	The FASTA file of the reference genome used to align the reads. **Mutascope** assumes the same folder also contains the 2bit version of the genome (Section 2.2)
-dbsnp	VCF	**REQUIRED**	A dbSNP file in VCF format (Section 2.2).
-pdir	PATH	**REQUIRED**	The path to the PROJECT directory (Section 4.4).
-normal	STRING	**REQUIRED**	The name of the 'normal' sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
-tumor	STRING	**REQUIRED**	The name of the 'tumor' sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
-nr1_fastq	FASTQ	**REQUIRED**	The Read1 FASTQ file of the 'normal' sample
-nr2_fastq	FASTQ	**REQUIRED**	The Read2 FASTQ file of the 'normal' sample
-tr1_fastq	FASTQ	**REQUIRED**	The Read1 FASTQ file of the 'tumor' sample
-tr2_fastq	FASTQ	**REQUIRED**	The Read2 FASTQ file of the 'tumor' sample
-BQ	INT	*Optional*	The minimum base quality used to calculate error rates and call variants (Default 10).
-MQ	INT	*Optional*	The minimum read mapping quality used to calculate error rates and call variants (Default 1).
-saaf	FLOAT	*Optional*	Minimum Percent of alternate alleles in the TUMOR for calling Somatic variants (Default 0.5).
-gaaf	INT	*Optional*	Minimum Percent of alternate alleles reads in the NORMAL sample for calling Germline variants (Default 15). Decreasing this will increase the variant calling time.
-mcovs	INT, FLOAT	*Optional*	The minimum coverage and lower coverage percentile used to calculate the minimum coverage to call variants; whichever is greater (Default 10,0.5).
-e	FLOAT	*Optional*	The default error rate for SNPs and Indels missing a measured error rate (Default 0.005).

-het	FLOAT	*Optional*	The heterozygous probability used in the Bayesian likelihood genotyping method (Default 0.001).
-length	INT	*Optional*	The length of the sequencing reads. If the sequencing length varies then use the length of the longest sequencing read (Default 151).
-t	INT	*Optional*	Threading option for BWA (Default 1).
-dist	INT	*Optional*	Read vs Amplicon start or stop alignment offset, in nucleotides (default 2)
-strand_specific	TRUE/FALSE	*Optional*	Reads should align to a specific strand. This strand is determined by the BED file's strand information. For amplicons with a + (-) strand, R1 should align to the forward (reverse) strand and R2 should align to the reverse (forward) strand. (Default FALSE).
-bwa_path	FILE	*Optional*	BWA executable (full path) (Default is your $PATH)
-samtools_path	FILE	*Optional*	SAMTools executable (full path) (Default is your $PATH)
-noBlacklist		*Optional*	Specifies to not generate/use a black-list file when calculating error rates and calling variants.
-useTumorErrors		*Optional*	Specifies to use the TUMOR error rates when calling somatic variants
-h			Outputs the help menu for the given command.

5.2. makeBlackList

makeBlackList generates a list of mutations to exclude from analysis (blacklist file). This module identifies alternate alignment possibilities using BWA and reads of length length created from the sequence of PCR amplicons (BED). Specific base-pair mismatches (SNP or indels) resulting from these alternate alignments are added to the file for further exclusion from the analysis. The file generated is BED.blacklist. This step is essential to remove potential false positive calls resulting from unspecific PCR products. makeBlacklist is only run once for a given set of amplicons. The specific substitutions or indels are then removed from the .xpileup file.

Input/Output

The required inputs are the list of PCR amplicons in a BED format and a reference genome in both .fa and .2bit format.

The output is a file called BED.blacklist, which contains a list of potential false positive mutations that are ignored in the further analysis.

Usage

Command	**Format**	**Option**	**Description**
-bed	BED	**REQUIRED**	A file in BED format containing the PCR amplicon information (Section 4.1)

-fasta	FASTA	**REQUIRED**	The FASTA file of the reference genome used to align the reads. **Mutascope** assumes the same folder also contains the 2bit version of the genome (Section 2.2)
-bwa_path	FILE	*Optional*	BWA executable (full path) (Default is your $PATH)
-length	INT	*Optional*	The length of the sequencing reads. If the sequencing length varies then use the length of the longest sequencing read (Default 151).
-t	INT	*Optional*	Threading option for BWA (Default 1).
-h			Outputs the help option for the given command

5.3. runBWA

runBWA is a module to run BWA using the bwasw command (Smith-Waterman alignment).

Input/Output

The required inputs are a reference genome in a FASTA format, sample name, and the read1 and read2 FASTQ files.

This program generates 2 alignment files in the *intermediate* sub-directory called "SAMPLE_read1_bwaSWAln.bam" and "SAMPLE_read2_bwaSWAln.bam".

Usage

Command	Format	Option	Description
-fasta	FASTA	**REQUIRED**	The FASTA file of the reference genome used to align the reads. **Mutascope** assumes the same folder also contains the 2bit version of the genome (Section 2.2)
-sample	STRING	**REQUIRED**	The name of the sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
-pdir	PATH	**REQUIRED**	The path to the PROJECT directory (Section 4.4).
-r1_fastq	FASTQ	**REQUIRED**	The **Read1** FASTQ file
-r2_fastq	FASTQ	**REQUIRED**	The **Read2** FASTQ file
-bwa_path	FILE	*Optional*	BWA executable (full path) (Default is your $PATH)
-samtools_path	FILE	*Optional*	SAMTools executable (full path) (Default is your $PATH)
-t	INT	*Optional*	Threading option for BWA (Default 1).
-h			Outputs the help option for the given command.

5.4. refinement

`refinement` is a module to remove low quality aligned reads along with un-clipping soft-clipped bases and trimming the primer sequence from the aligned reads. It also assigns a read group to each sequencing read that corresponds to the PCR amplicon and read (1 or 2) it was generated from. This is a key step in **Mutascope** because the read groups assigned here are used to generate the xpileup file, calculate error rates, and call variants. The module also outputs a set of alignment and reads per amplicon quality control metrics.

Input/Output

The required inputs are a BED file of the PCR amplicons, the sample name, and an output directory containing a sub-directory called *intermediate* containing the output files from *runBWA* (SAMPLE_read2_bwaSWAln.bam).

This program outputs a BAM in the *intermediate* sub-directory called "SAMPLE_merged_unClipped_primersClipped.bam".

Usage

Command	Format	Option	Description
`-bed`	BED	**REQUIRED**	A file in BED format containing the PCR amplicon information (Section 4.1)
`-pdir`	PATH	**REQUIRED**	The path to the PROJECT directory (Section 4.4).
`-sample`	STRING	**REQUIRED**	The name of the sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
`-dist`	INT	*Optional*	Read vs Amplicon start or stop alignment offset, in nucleotides (default 2).
`-length`	INT	*Optional*	The length of the sequencing reads. If the sequencing length varies then use the length of the longest sequencing read (Default 151).
`-strand_specific`	TRUE/ FALSE	*Optional*	Reads should align to a specific strand. This strand is determined by the BED file's strand information. For amplicons with a + (-) strand, R1 should align to the forward (reverse) strand and R2 should align to the reverse (forward) strand. (Default FALSE).
`-samtools_path`	FILE	*Optional*	SAMTools executable (full path) (Default is your $PATH)
`-h`			Outputs the help menu for the given command.

5.5. groupRealign

`groupRealign` is a module to run GATK's realignment program on the tumor and normal samples together to improve SNP and indel calling. Realigning reads around indels improves both SNP and indel calling. Realigning both the matched normal and tumor sample together will help remove false positive LOH and/or somatic variants that can be caused by different alignments around a homopolymer run.

Input/Output

The required inputs are a BED file of the PCR amplicons, a FASTA file of the reference genome, the normal sample name, tumor sample name, and an output directory containing a sub-directory called *intermediate* containing the output files from *refinement*

- `Normal_merged_unClipped_primersClipped.bam`
- `Tumor_merged_unClipped_primersClipped.bam`

This program outputs two BAM files in the *intermediate* sub-directory called

- `Normal_merged_unClipped_primersClipped_realigned.bam`
- `Tumor_merged_unClipped_primersClipped_realigned.bam`

Usage

Command	Format	Option	Description
-bed	BED	**REQUIRED**	A file in BED format containing the PCR amplicon information (Section 4.1)
-fasta	FASTA	**REQUIRED**	The FASTA file of the reference genome used to align the reads. **Mutascope** assumes the same folder also contains the 2bit version of the genome (Section 2.2)
-pdir	PATH	**REQUIRED**	The path to the PROJECT directory (Section 4.4).
-normal	STRING	**REQUIRED**	The name of the 'normal' sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
-tumor	STRING	**REQUIRED**	The name of the 'tumor' sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
-samtools_path	FILE	*Optional*	SAMTools executable (full path) (Default is your $PATH)
-h			Outputs the help menu for the given command.

5.6. xpileup

`xpileup` is a module that uses SAMTools pileup and custom perl scripts to generate a xpileup file used for variant calling and calculating error rates. The xpileup file is an extended pileup file containing the read group information assigned during the `refinement` module.

Input/Output

The required inputs are a BED file of the PCR amplicons, a FASTA file of the reference genome, sample name, and an output directory containing a sub-directory called *intermediate* containing the output files from `groupRealign` (SAMPLE_ merged_unClipped_primersClipped_realigned.bam).

This program outputs a gzip xpileup file in the *intermediate* sub-directory called "SAMPLE.xpileup.gz".

Usage

Command	Format	Option	Description
-bed	BED	**REQUIRED**	A file in BED format containing the PCR amplicon information (Section 4.1)
-fasta	FASTA	**REQUIRED**	The FASTA file of the reference genome used to align the reads. **Mutascope** assumes the same folder also contains the 2bit version of the genome (Section 2.2)
-pdir	PATH	**REQUIRED**	The path to the PROJECT directory (Section 4.4).
-sample	STRING	**REQUIRED**	The name of the sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
-length	INT	*Optional*	Length of the sequencing reads. If LENGTH varies than input the length of the longest read (Default 151)
-samtools_path	FILE	*Optional*	SAMTools executable (full path) (Default is your $PATH)
-h			Outputs the help menu for the given command.

5.7. calcErrorRates

`calcErrorRates` is a module to calculate error rates of a given SAMPLE using non-dbSNP sites and non-blackList sites. The errorRate file generated contains the calculated error rates given the position within a read, read1/read2, and the mismatch type. These error rates are used when determining if a position contains a variant. The base quality and mapping qualities used in this step should be the same values used in the `callSomatic` module. Bases are filtered out in this step because they are filtered in the `callSomatic` module to improve variant calling and the bases used in the `callSomatic` module should be the same bases used to calculate the error rates.

Input/Output

The required inputs are a BED file of the PCR amplicons, a VCF file of dbSNP, sample name, and an output directory containing a sub-directory called *intermediate* containing the output files from *xpileup* : `SAMPLE.xpileup`.

This program outputs an error rate file in the *intermediate* sub-directory called `SAMPLE.errorRates`.

Usage

Command	Format	Option	Description
`-bed`	BED	**REQUIRED**	A file in BED format containing the PCR amplicon information (Section 4.1)
`-dbsnp`	VCF	**REQUIRED**	A dbSNP file in VCF format (Section 2.2).
`-pdir`	PATH	**REQUIRED**	The path to the PROJECT directory (Section 4.4).
`-sample`	STRING	**REQUIRED**	The name of the sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
`-BQ`	INT	*Optional*	The minimum base quality used to calculate error rates and call variants (Default 10).
`-MQ`	INT	*Optional*	The minimum read mapping quality used to calculate error rates and call variants (Default 1).
`-noBlacklist`			Specifies to not generate/use a black-list file when calculating error rates and calling variants.
`-h`			Outputs the help menu for the given command.

5.8. callSomatic

`callSomatic` is the module that calls variants from a tumor-normal pair. It also applies a set of filters to remove false positive variants. This module uses the tumor-normal pair information as well as the calculated error rates to call variants. A Bayesian-likelihood algorithm is used to determine germline genotypes and a binomial test using calculated error rates is used to call somatic SNVs. Indels are filtered based on coverage, proximity to other indels, and homopolymer runs. **Mutascope** uses a set of standard filters, such as SNP clusters and proximity to an indel, as well as some specific filters, such as the read group bias and minimum average alternate allele base quality score filter.

Input/Output

The required inputs are a BED file of the PCR amplicons, a fasta version of the reference sequence, a VCF file of dbSNP, normal and tumor sample name, and an output directory containing a sub-directory called *intermediate* containing the output files from `xpileup` (NORMAL.xpileup.gz and TUMOR.xpileup.gz).

This program outputs a single VCF file after applying the filters below. The file is located in the *results* sub-directory and called “NORMAL_vs_TUMOR _filtered.vcf”. It also outputs a PDF containing a plot of the germline and somatic variants that passed all the filters as well as a file containing the LogR ratio for each amplicon.

FILTER	Description
MIF	Multiple INDELs called at a single position (≥2)
HRUN	Homopolymer repeat within 1bp of the start or end of the INDEL (≥5bp)
MCF	Low coverage (Normal = X and Tumor = Y)

SBFILT	Variant shows Strand Bias (Reference contains coverage on both strands while the alternate is seen online on 1 strand)
RGBF	Variant shows Read-Group Bias ($< 10^{-15}$)
DSF	The distance between the first and second alternate allele frequencies is too close (DS < 0.5)
MAMQ	Low average Mapping Quality score (< 10)
MFEP	Low Fisher-exact P-value for Somatic variants (> 0.0005)
MBPF	Low Binomial P-value to call the position a variant ($-\log_{10}$(P-value) < 54)
MAABQ	Low average alternate allele Base Quality score (> 0.99 Probability of being in the High Quality cluster)
CNIF	The SNP is within 10bp of an Indel
SNPCF	The SNP is within 10bp of 2 other SNPs
ICF	The Indel is within 10bp of 1 other Indel

Usage

Command	Format	Option	Description
`-bed`	BED	**REQUIRED**	A file in BED format containing the PCR amplicon information (Section 4.1)
`-dbsnp`	VCF	**REQUIRED**	A dbSNP file in VCF format (Section 2.2).
`-fasta`	FASTA	**REQUIRED**	The FASTA file of the reference genome used to align the reads. **Mutascope** assumes the same folder also contains the 2bit version of the genome (Section 2.2)
`-pdir`	PATH	**REQUIRED**	The path to the PROJECT directory (Section 4.4).
`-normal`	NAME	**REQUIRED**	The name of the 'normal' sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
`-tumor`	NAME	**REQUIRED**	The name of the 'tumor' sample containing only numbers and letters (No symbols). The same name needs to be used for the same sample for all **Mutascope** modules.
`-BQ`	INT	*Optional*	The minimum base quality used to calculate error rates and call variants (Default 10).
`-MQ`	INT	*Optional*	The minimum read mapping quality used to calculate error rates and call variants (Default 1).
`-saaf`	FLOAT	*Optional*	Minimum Percent of alternate alleles in the TUMOR for calling Somatic variants (Default 0.5).
`-gaaf`	INT	*Optional*	Minimum Percent of alternate alleles reads in the NORMAL sample for calling Germline variants (Default 10). Decreasing this will increase the variant calling time.

`-mcovs`	INT, FLOAT	*Optional*	The minimum coverage and lower coverage percentile used to calculate the minimum coverage to call variants; whichever is greater (Default 10,0.5).
`-e`	FLOAT	*Optional*	The default error rate for SNPs and Indels missing a measured error rate (Default 0.005).
`-het`	FLOAT	*Optional*	The heterozygous probability used in the Bayesian likelihood genotyping method (Default 0.001).
`-noBlacklist`		*Optional*	Specifies to not generate/use a black-list file when calculating error rates and calling variants.
`-useTumorErrors`		*Optional*	Specifies to use the TUMOR error rates when calling somatic variants
`-h`			Outputs the help option for the given command.

6. Frequently Asked Questions

6.1. Can I analyze a Tumor sample without a matched Normal?

Yes but it is not recommended. **Mutascope** has not been tested without a normal sample. A matched normal sample sequenced in the same run as the tumor sample is essential to measure accurately the experimental error rate and reliably call rare somatic mutations. If you do not have a matched normal sample, we recommend using another sample (may be tumor) run on the same sequencing run. You specify the "-useTumorErrors" option. You can also combine multiple samples from the same run and use that as a normal sample. The main problem with analyzing the tumor sample this way is that the variant classification (germline, LOH, or somatic) may be incorrect. Some somatic variants in the tumor sample may exist in the non-matched 'normal' sample; thus changing the variant classification from somatic to germline. The opposite may also occur; a germline variant in the tumor sample may not exist in the non-matched 'normal' sample and thus changing the variant classification from germline to somatic. Therefore you may not be able to trust the variant's classification called by **Mutascope** because you do not have a matched normal sample.

6.2. What happens if the normal DNA was sequenced independently from the tumor DNA ?

The error rates can be slightly different from sequencing run to sequencing run. This could potentially lead to false positive low frequency somatic variants. First you should look at the error rates of the normal sample and the tumor sample and see how much they vary. You could still use the normal sample's error rates if, on average, the error rates are within 2-fold of the tumor sample's error rates. Otherwise you may want to consider instead using the tumor sample's error rate file in the `callSomatic` script, assuming the mutation rate of your sample is low and the mutations are at low prevalence. You can specify to use the tumor samples error rates by specifying *–useTumorErrors*.

6.3. How long does it take to run Mutascope ?

For a set of 1676 amplicons (~150kb) and a matched tumor-normal pair with ~1 million reads for read1 and read2 for both the tumor and normal sample (4 million total); **Mutascope** takes 2 hours and 40 minutes using 8 CPU and 4G of RAM during the BWA alignment step and 1 CPU and 2G of RAM during the remaining modules of **Mutascope**.

6.4. What version of dbSNP should I use?

Any version of dbSNP greater than 131 should work as it contains most variants present at ≥1% of the population. **Mutascope** does not currently select for common SNPs. Excluding too many sites can affect the accuracy of the error rate and can affect overall **Mutascope** performance. However, for dbSNP137 and lower this should not be a concern. Alternatively, one can create a subset of dbSNP containing only common variants to use with Mutascope.

*6.5. Can I use **MUTASCOPE** to discover copy number aberration*

Yes and no. **Mutascope** outputs a file containing the coverage ratio between tumor and normal for each amplicon. This file can then be used to call copy number aberrations. In contrast to somatic SNV calling, this method will however be sensitive to sample heterogeneity. In addition, the allelic fraction of the germline heterozygotes variants in the tumor can be indicative of loss of heterozygosity, a common consequence of many copy number aberrations.

*6.6. Can I use **MUTASCOPE** to analyze Ion Torrent PGM or Proton data ?*

Maybe. **Mutascope** has not been tested using Ion Torrent PGM or Proton data. The PGM data should be aligned using Ion Torrent's recommended alignment algorithm. We do not currently recommend using **Mutascope** on Ion Torrent or Proton data but encourage users to explore this possibility and report on their experience.

*6.7. How does **MUTASCOPE** compares to other somatic variant callers ?*

Mutascope has a higher sensitivity and positive predictive value than the standard somatic variant callers. Currently, and to our knowledge, **Mutascope** is the only variant caller that was designed specifically to detect somatic variants with a frequency lower that 5% using amplicon-based targeted sequencing.

*6.8. How does **MUTACOPE** call indels ?*

In contrast to the probabilistic approach used for SNVs**, Mutascope** currently uses a frequency approach to call indels. `callSomatic` uses the *gaaf* and *saaf* options to determine the genotype of the indel. **Mutascope** does not calculate error rates for indels and thus uses the given default error rate (-*e*) as the error rate in the Binomial test. The default error rate should not be changed unless the average error rates in the error rate file generated by **Mutascope** is higher than –*e*. The dataset used to test **Mutascope** did not have a set of reference indels to check the performance of indel calling. However, we did compare the indel calling results of **Mutascope** to a set of standard variant callers (i.e. GATK, SAMTools, and VarScan) and found that indels called by **Mutascope** were also called by the standard variant callers. Similar to GATK, identifying homopolymer runs around the indels and filtering out false positive indels was key to obtain this performance.

*6.9. How can I just call variants using **MUTASCOPE** without doing the refinement or groupRealign steps?*

Future versions of **Mutascope** will feature an *addReadGroups* module to do that. You can then generate the xpileup files and error rate files and finally call variants with the callSomatic module. The key to **Mutascope**'s variant calling is the read groups that are assigned to each read. These read groups are used when calculating error rates and calling variants.

6.10. What does a sample swap look like?

If you are sequencing matched tumor-normal samples, then the variant plot should look like Figure 7. However, if there is a sample swap or you specified the wrong FASTQ files or name when running **Mutascope** then your variant plot will look like Figure 9, showing an aberrant distribution of allelic fractions (e.g. 1 in the normal and 0.5 in the tumor).

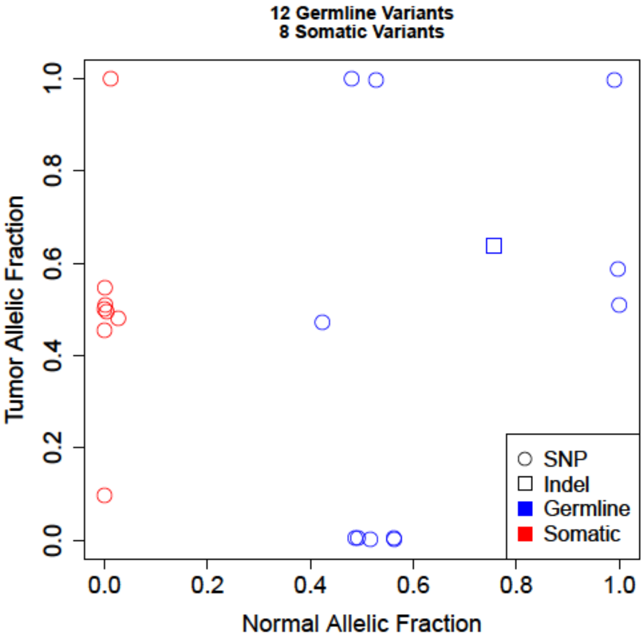


Figure 9: Example of a variant plot where the tumor and normal samples are not from the same patient

7. References

1. Harismendy O, Schwab RB, Bao L, Olson J, Rozenzhak S, et al. (2011) Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome Biol 12: R124.
2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
3. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
4. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
5. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.
6. Team RC (2012) R: A Language and Environment for Statistical Computing.
7. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.