

# MosaicHunter User Guide

Version 1.0

May 18th, 2016

<http://mosaichunter.cbi.pku.edu.cn/>

mosaichunter@mail.cbi.pku.edu.cn

August Yue Huang, Adam Yongxin Ye, Zheng Zhang, Yanmei Dou

Center for Bioinformatics, Peking University

## Contents

I	Program.....	3
I.1	Getting Started.....	3
I.1.a	Installation & quick-start .....	3
I.1.b	Preparing your reference.....	4
I.1.c	Preparing your reads .....	4
I.1.d	Prepare sample-specific files .....	5
I.1.e	Running MosaicHunter .....	5
I.2	Framework and Best Practice .....	6
I.2.a	Program framework and work flow .....	6
I.2.b	Best practice pipelines .....	7
I.3	Bayesian Models and Program Modes .....	14
I.3.a	Single mode, paired mode & trio mode .....	14
I.3.b	Whole-genome sequencing vs. whole-exome sequencing.....	16
II	Manual for Parameters and Filters .....	17
II.1	Top Parameters .....	17
II.2	Filter Introduction and Parameters .....	20
II.3	Output Files and Format.....	33
III	Miscellaneous.....	35

III.1 Contact Information .....35

III.2 Acknowledgement .....36

III.3 License .....36

III.4 FAQ.....36

# I Program

**MosaicHunter** is a bioinformatic command-line tool for identifying postzygotic single-nucleotide mosaicism (SNM) sites from whole-genome or whole-exome sequencing data of unpaired, trio, or paired samples. It implements and wraps the Bayesian genotypers and several stringent filters in a java framework.

MosaicHunter is developed and implemented by August Yue Huang, Adam Yongxin Ye, Zheng Zhang, Yanmei Dou in the Center for Bioinformatics at Peking University, with helps from many other colleagues.

## I.1 Getting Started

### I.1.a Installation & quick-start

#### Prerequired tools

java ( $\geq 1.7$ ) (<https://java.com/en/>)

UCSC blat (<http://genome.ucsc.edu/FAQ/FAQblat.html#blat3>)

Note you will need to include the directory containing runnable UCSC blat in your PATH (or set the parameter `blat_path` for `misaligned_reads_filter`).

#### Suggested tools

git (<https://git-scm.com/downloads>)

ant (<https://ant.apache.org/bindownload.cgi>)

samtools (<http://sourceforge.net/projects/samtools/files/>)

Picard (<http://broadinstitute.github.io/picard/>)

GATK (<https://www.broadinstitute.org/gatk/download/auth?package=GATK>)

#### Download MosaicHunter program and required resource files

Clone the MosaicHunter repository

```
cd your_path
```

```
git clone https://github.com/zzhang526/MosaicHunter.git
```

Some large required resource files can be accessed by URL:

[http://soms.nibs.ac.cn:6235/glk/human\\_g1k\\_v37.fasta](http://soms.nibs.ac.cn:6235/glk/human_g1k_v37.fasta)

[http://soms.nibs.ac.cn:6235/dbsnp/dbsnp\\_137.b37.tsv](http://soms.nibs.ac.cn:6235/dbsnp/dbsnp_137.b37.tsv)

You also need to download these files, and put them into `your_path/resources/` directory.

### Quick usage of a pre-compiled binary-code release

In order to make it easy to install MosaicHunter, we provide the pre-compiled release, which is in the unpacked `build/` directory. To use it, simply type

```
cd your_path/MosaicHunter
java -jar build/mosaichunter.jar
```

### Build MosaicHunter from source

We also provide the source code of MosaicHunter. If you want to build it from source, you need to install ant first (<http://ant.apache.org/bindownload.cgi>), then type

```
cd your_path/MosaicHunter
ant
```

The newly compiled `mosaichunter.jar` file will overwrite the pre-compiled one in `build/` directory, and you can run MosaicHunter by

```
java -jar build/mosaichunter.jar
```

#### I.1.b Preparing your reference

MosaicHunter requires a fasta file (.fasta, .fa) for your reference genome. It is better to make sure that your reference file has the same name and order of contigs to your .bam file(s) and .bed file(s).

Reads aligned to any contigs which are not appeared in the reference file will be ignored.

When running MosaicHunter, you need to set the top parameter `reference_file`.

#### I.1.c Preparing your reads

MosaicHunter currently accepts aligned reads from whole-genome or whole-exome sequencing. We recommend bwa for read mapping and GATK/Picard for read preprocessing (Picard-SortSam, Picard-MarkDuplicates, GATK-IndelRealignment, GATK-BaseRecalibration). These preprocessing steps are important to reduce false positives in identifying mosaic sites.

MosaicHunter uses the sorted and indexed bam file(s) to identify mosaic sites, as set by the top parameter `input_file` to the path of your major .bam file.

For the 'trio' mode, you also need to specify `father_bam_file`, `mother_bam_file`; and for the 'paired' mode, `control_bam_file` are required to be specified.

In addition, we suggest to only keep proper-mapped reads (for paired-end reads; with flag 0x2), and drop reads with NM > 4, to make the input cleaner. The command to achieve it can be like

```
samtools view -h -f 0x2 input.bam | perl -ne 'print if (/^@/| ( /NM:i:(\d+)/&&$1<=4))' | samtools view -Sb - >cleaner.bam
```

#### I.1.d Prepare sample-specific files

False mosaic sites can be caused by alignment errors in the genomic regions containing indels and CNVs, thus we suggest you to generate a list of such error-prone regions for MosaicHunter to mask. We recommend to identify and filter indels and CNVs from the processed .bam file following the pipelines of GATK and CNVnator, respectively. The regions of called indels (with +/-5bp flanks) and CNVs should be then converted to the .bed format and specify the corresponding parameter `indel_region_filter.bed_file` when running MosaicHunter.

#### I.1.e Running MosaicHunter

The standard command-line usage of MosaicHunter is

```
java -jar your_path/build/mosaichunter.jar <predefined_configuration>  
-P param_1=value_1 [-P param_2=value_2 [...]]
```

(Note: `predefined_configuration` could be 'genome', 'exome', or 'exome\_parameters')

or

```
java -jar your_path/build/mosaichunter.jar -C <config.properties> -P  
param_1=value_1 [-P param_2=value_2 [...]]
```

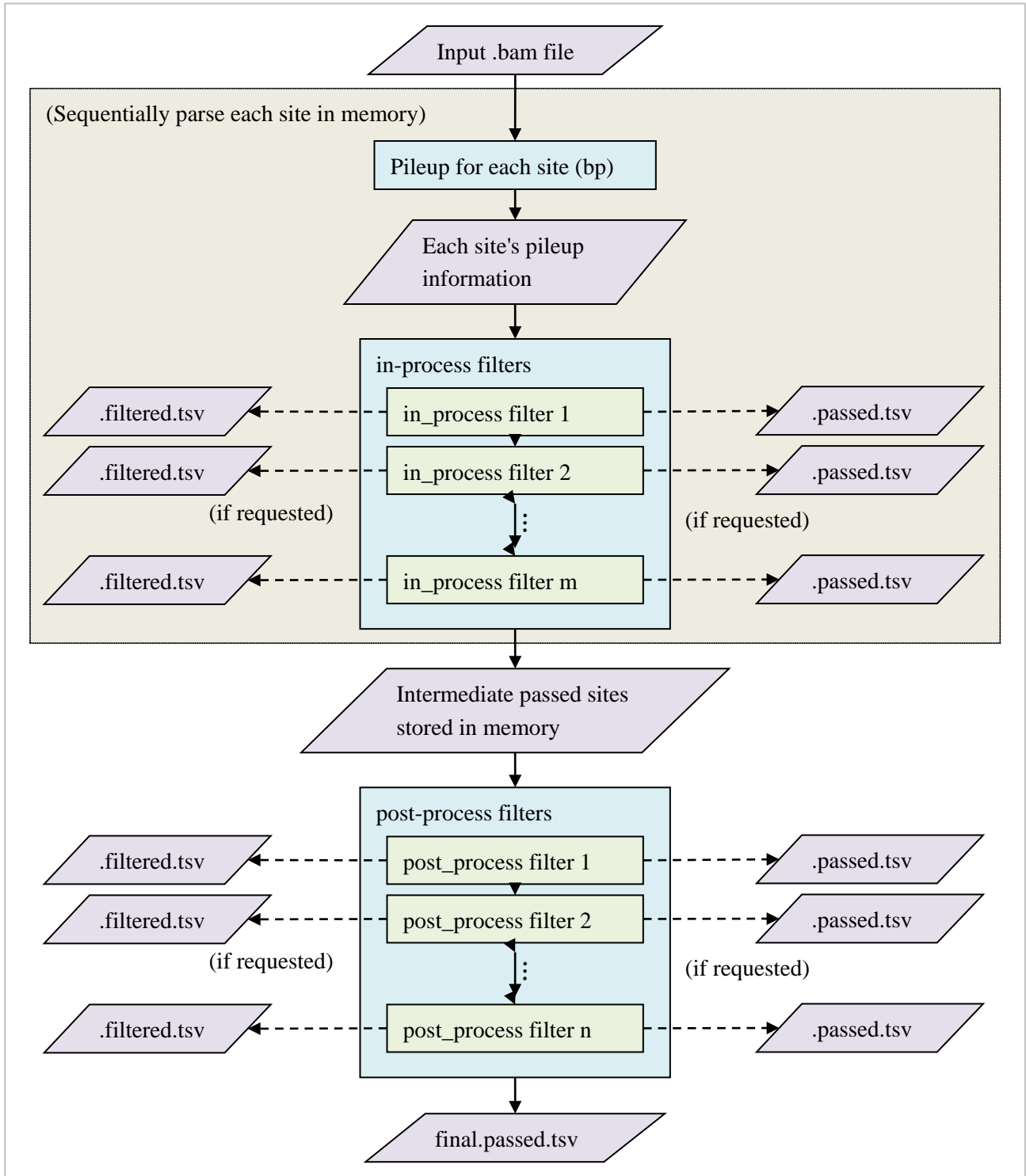
The output files will be put into the path set by the top parameter `output_dir`.

(Details see also I.2 and II.3)

## I.2 Framework and Best Practice

### I.2.a Program framework and work flow

The main framework and work flow of MosaicHunter is demonstrated in the figure below.



For running efficiency, by default, the **in\_process** filters will be applied first to each sites and then the **post\_process** filters will be applied sequentially. Any filters which are independently to adjacent

sites can be in\_process filters, while some filters such as clustered\_filter that require the information of adjacent sites can only be post\_process filters. The much slower misaligned\_reads\_filter (which calls UCSC blat) is recommended to be run at the last step.

Details about filters see also II.2.

The pipeline of filters with their parameters (and the 'top parameters') are usually specified in the config.properties file (-C <path> option), with a few command-line overwritten parameters (-P param=value option). Note: if the value is a string with any space character, quotes ( ' or " ) are not needed in the config.properties file, but may be required in the command-line, because bash will separate options by space character.

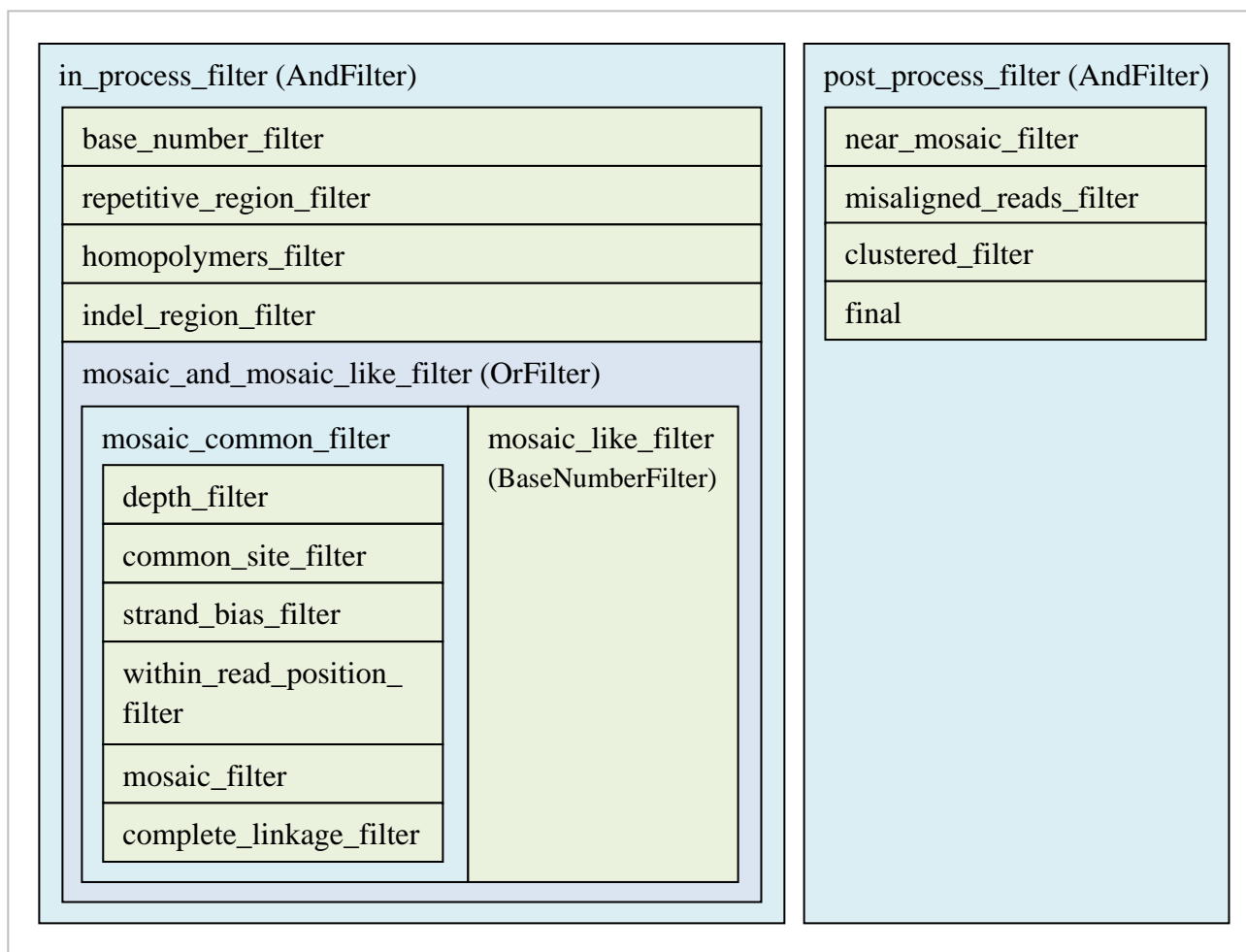
### **I.2.b Best practice pipelines**

Here we describe the suggested config properties for analyzing whole-genome sequencing (WGS) and whole-exome sequencing (WES) data obtained from either normal (non-cancer) samples or cancer samples. You can also find them in the conf / directory.

#### **1) Whole genome sequencing for normal (non-cancer) samples**

(see also conf/default.properties for the list of parameters with default values)

The flowchart and structure of filters applied to WGS data is shown in this figure. (see also I.2.a)



In brief, by masking the error-prone genomic regions (`repetitive_region_filter`, `homopolymers_filter`, `indel_region_filter`) and sites with extremely high or low sequencing depth (`base_number_filter`), we identify candidate mosaic sites (`mosaic_common_filter`) and sites with strange allele fraction (AF) – 'mosaic\_like' sites (`mosaic_like_filter`). We then use `near_mosaic_filter` to preserve the 'mosaic\_like' sites which are closed to candidate mosaic sites only, then check misalignment by calling UCSC blat (`misaligned_reads_filter`), and remove clustered mosaic candidate sites (which are more likely due to CNVs or unannotated repetitive elements).

Note: `mosaic_and_mosaic_like_filter`, `mosaic_like_filter` and `near_mosaic_filter` are designed to work coupling with `clustered_filter`, in which we focused on candidate 'mosaic' sites as well as some auxiliary nearby 'mosaic\_like' sites (see also II.2). Such filters are suitable for WGS but not WES analysis.

You can run this pipeline by:

```
java -jar mosaichunter.jar genome [-P param=value [-P ...]]
```

or

```
java -jar mosaichunter.jar -C conf/genome.properties [-P param=value [-P ...]]
```



Here are some required parameters:

```
input_file=<path>
reference_file=<path>
output_dir=<path>
mosaic_filter.sex=<M|F>
mosaic_filter.mode=<single|trio|paired_naive|paired_fisher>
```

Some path parameters of resource files:

```
mosaic_filter.dbsnp_file=<path>
repetitive_region_filter.bed_file=<path>
indel_region_filter.bed_file=<path>
common_site_filter.bed_file=<path>
```

Some parameters we suggest to modify according to your data:

```
max_depth=500
depth_filter.min_depth=25
depth_filter.max_depth=150
```

We suggest to set `depth_filter.min_depth` and `depth_filter.max_depth` to be Q10 (10% quantile) and Q90 (90% quantile) of depth in your data, and calling mosaic sites with depth less than 25 may be unreliable.

The final output will be `output_dir/final.passed.tsv`.

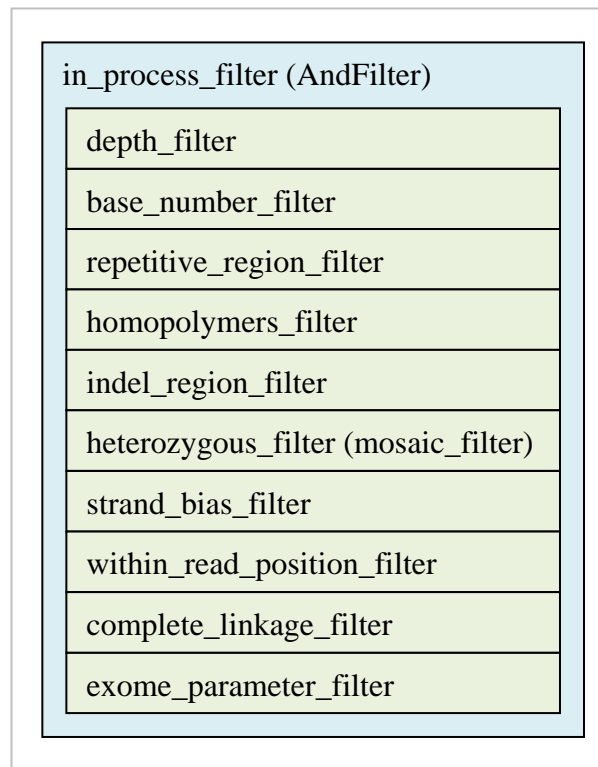
See also II.3 for interpretation of the format.

## 2) Whole exome sequencing for normal (non-cancer) samples

**1st step: estimate the shape parameters ( $\alpha$ ,  $\beta$ ) for the beta prior of heterozygous sites**

(see also `conf/exome_parameters.properties` for the list of parameters with default values)

The flowchart and structure of filters applied to this step is shown in this figure. (see also I.3.a)



In brief, after applying filters to remove technical artifacts, we estimate the distribution of depths and AFs among heterozygous sites, which will be used in the 2nd step for distinguishing mosaic sites from heterozygous sites. The final output of the average depth and estimated shape parameters ( $\alpha$ ,  $\beta$ ) will be reported to `output_dir/stdout_*.log`.

You can run this step by:

```
java -jar mosaichunter.jar exome_parameters [-P param=value [-P ...]]
```

or

```
java -jar mosaichunter.jar -C conf/exome_parameters.properties [-P param=value [-P ...]]
```

Here are some required parameters:

`input_file=<path>`

`reference_file=<path>`

`output_dir=<path>`

`heterozygous_filter.sex=<M|F>`

Some path parameters of resource files:

```
repetitive_region_filter.bed_file=<path>  
indel_region_filter.bed_file=<path>  
common_site_filter.bed_file=<path>  
mosaic_filter.dbsnp_file=<path>
```

Some parameters we suggest to modify according to your data:

```
max_depth=5001  
depth_filter.min_depth=25  
depth_filter.max_depth=5000
```

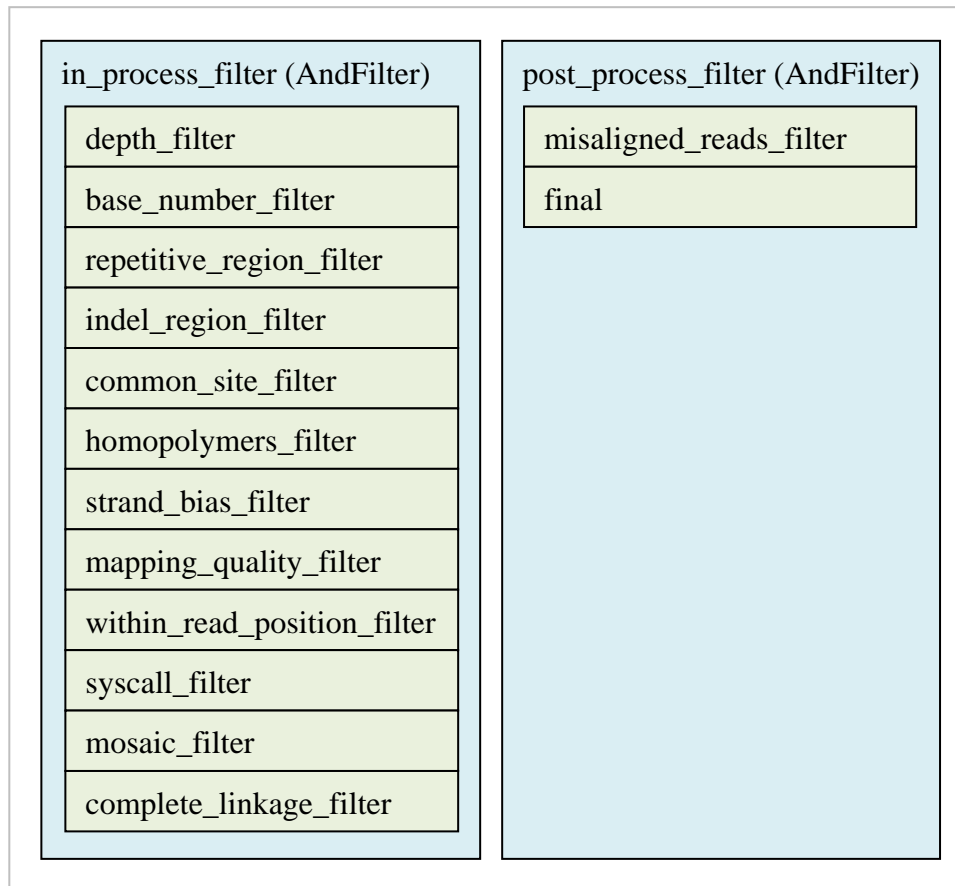
For efficiency issue, we suggest to limit the parameter `max_depth`  $\leq 5000$ .

The final output will be `output_dir/stdout_*.log`.

## **2nd step: actually call mosaic sites**

(see also `conf/exome.properties` for the list of parameters with default values)

The flowchart and structure of filters applied to this step is shown in this figure. (see also I.3.a)



Compared to the WGS flowchart, we add `mapping_quality_filter` and `syscall_filter` for the WES analysis, and remove `clustered_filter` as well as its related `near_mosaic_filter` and `mosaic_like_filter`. Remember to specify `mosaic_filter.alpha_param`, `mosaic_filter.beta_param` and `syscall_filter.depth` to  $\alpha$ ,  $\beta$  and the average depth, which are estimated in the 1st step (check `output_dir/stdout_*.log`). We recommend to set a different `output_dir` from the one in the 1st step.

You can run this step by:

```
java -jar mosaichunter.jar exome [-P param=value [-P ...]]
```

or

```
java -jar mosaichunter.jar -C conf/exome.properties [-P param=value [-P ...]]
```

Here are some required parameters:

`input_file=<path>`

`reference_file=<path>`

`output_dir=<path>`

`misaligned_reads_filter.reference_file=<path>`

```
mosaic_filter.sex=<M|F>
mosaic_filter.alpha_param=<int>
mosaic_filter.beta_param=<int>
```

Some path parameters of resource files:

```
repetitive_region_filter.bed_file=<path>
indel_region_filter.bed_file=<path>
common_site_filter.bed_file=<path>
mosaic_filter.dbsnp_file=<path>
```

Some parameters we suggest to modify according to your data:

```
max_depth=500
depth_filter.min_depth=25
depth_filter.max_depth=150
syscall_filter.depth=66
```

The final output will be `output_dir/final.passed.tsv`.

See also II.3 for interpretation of the format.

### 3) Whole genome sequencing for cancer samples

(see also `conf/cancer.properties` for the list of parameters with default values)

The pipeline for cancer sample is almost identical to the pipeline for normal (non-cancer) sample, except that `mosaic_rate` in the `mosaic_filter` is changed to `1e-6` and `inner_distance` and `outer_distance` in the `clustered_filter` are changed to `2000`, in order to increase the sensitivity to identify SNM in cancer samples.

You can run this pipeline by:

```
java -jar mosaichunter.jar -C conf/cancer.properties [-P param=value
[-P ...]]
```

## I.3 Bayesian Models and Program Modes

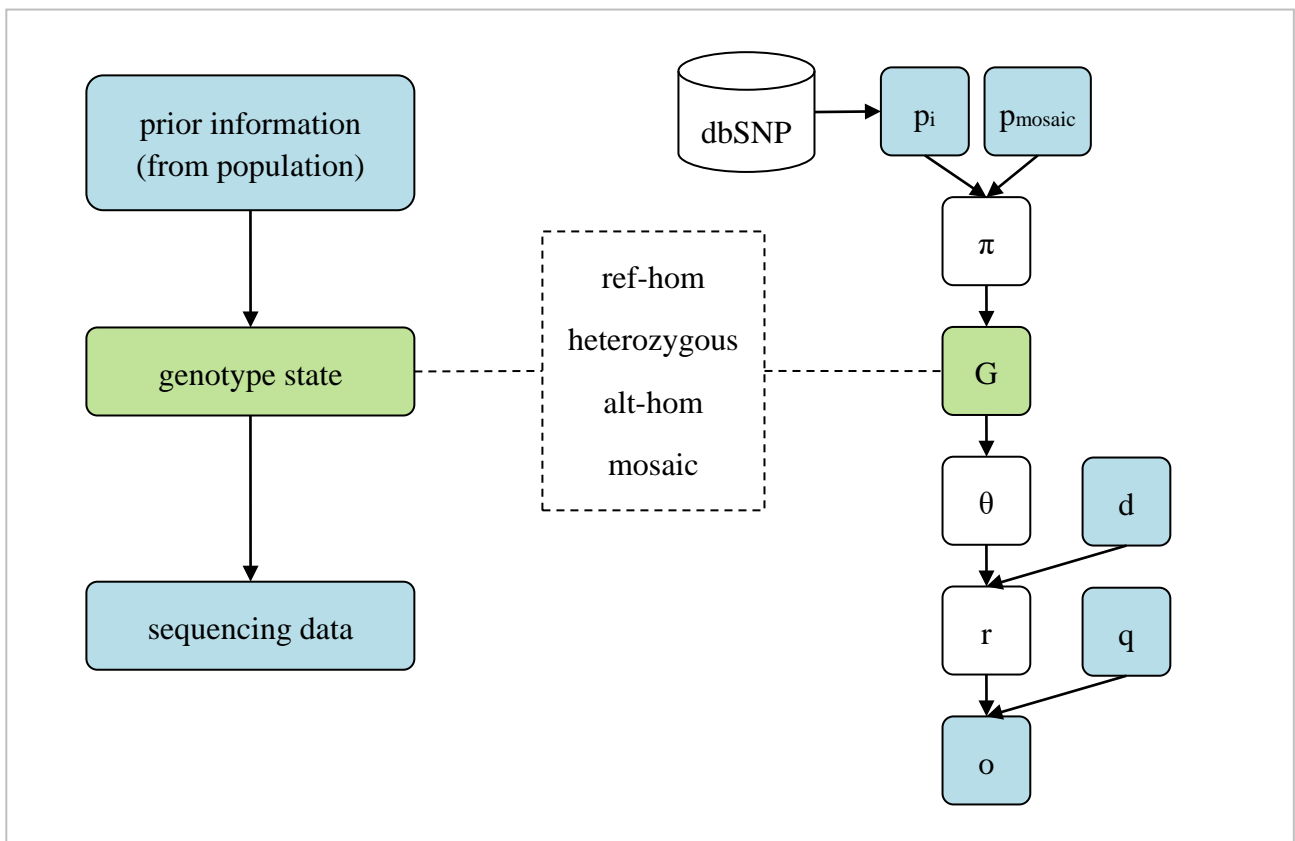
### I.3.a Single mode, paired mode & trio mode

Because matched control samples are sometimes unavailable in practice, MosaicHunter implements a Bayesian genotyper to identify SNMs from the unpaired samples ('single' mode). In addition, MosaicHunter is able to utilize the parental sequencing data ('trio' mode) to achieve a better performance, especially for calling SNMs with AF close to 0.5. For the practice of calling somatic mutation in cancer studies, MosaicHunter can also utilize the sequencing data of matched control samples obtained from the same individual ('paired' mode).

#### 1) Single mode

When no related samples are available, after users prepared its bam file, it will be easy to call SNMs with the 'single' mode of MosaicHunter: set the bam file as `input_file`, and set `mosaic_filter`'s parameters `mode=single` (see also II.2).

The probabilistic graphical model of the Bayesian genotyper (implemented in `mosaic_filter`) for unpaired sample can be depicted in this figure.



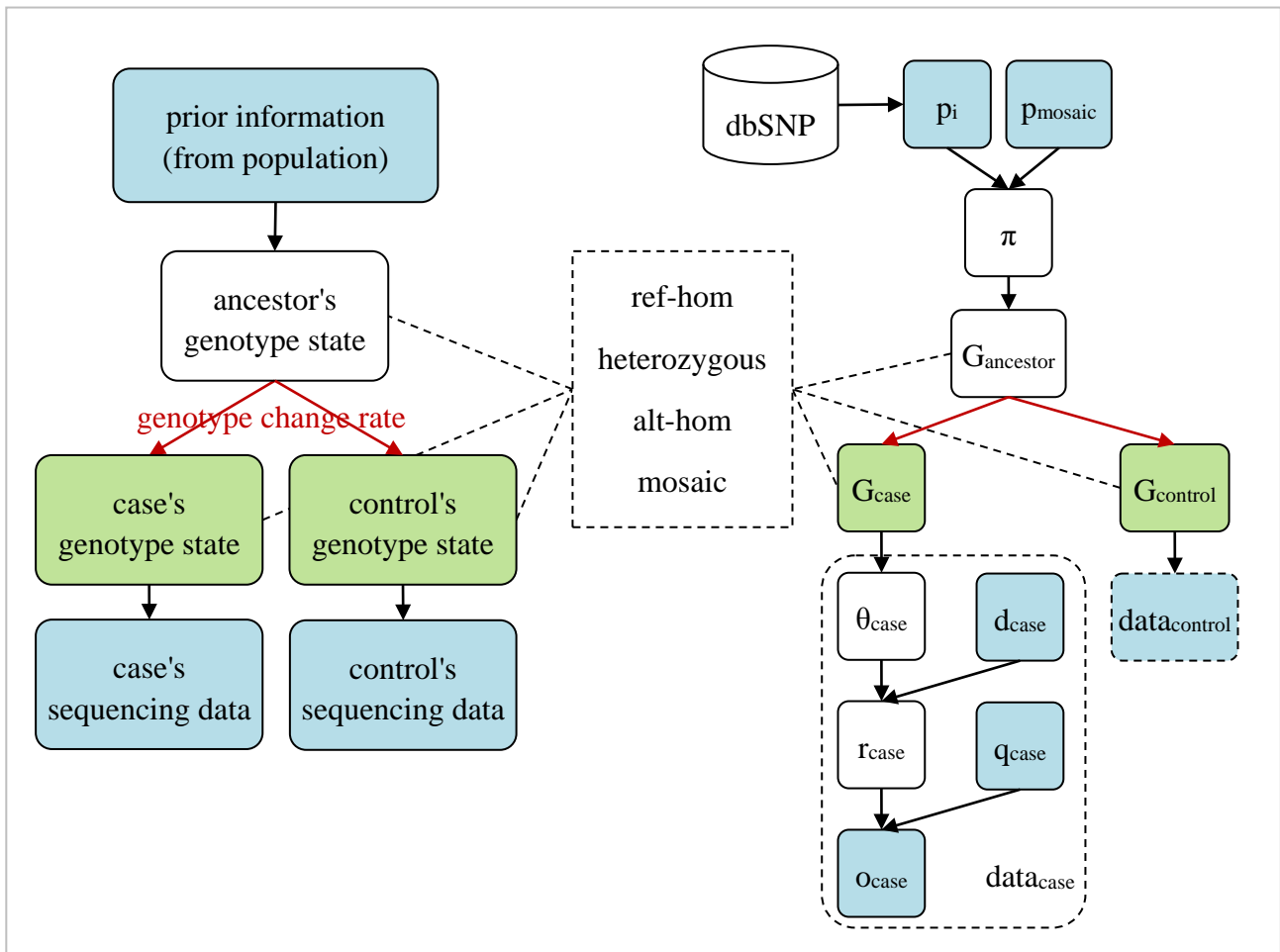
The blue boxes in the figure represent pre-specified prior information or the observed sequencing data (specified as `mosaic_filter`'s parameter or data files), and the green box of genotype state `G` is

what we try to infer. We use Bayesian inference to estimate the genotype state  $G$ . The mosaic\_filter will pass the sites if its mosaic posterior probability  $P(G=\text{mosaic}|\text{data}) > \text{mosaic\_filter's parameter mosaic\_threshold}$ .

## 2) Paired mode

We implemented two types of 'paired' mode, namely 'naive' paired mode and 'fisher' paired mode. Both additionally need to specify the parameter `control_bam_file=<path>` in mosaic\_filter.

In the 'naive paired' mode (mosaic\_filter's `mode=paired_naive`), we extended the Bayesian model and incorporated a latent variable – ancestor's genotype, and a genotype change rate matrix. We inferred the joint posterior probability of the genotype states of both case and control samples, and summed up the posterior probability that their genotype states were different, and will pass the sites if this probability  $P(G_{\text{case}} \neq G_{\text{control}}|\text{data}) > \text{mosaic\_filter's parameter mosaic\_threshold}$ . The probabilistic graphical model for 'naive paired' mode can be shown in this figure.

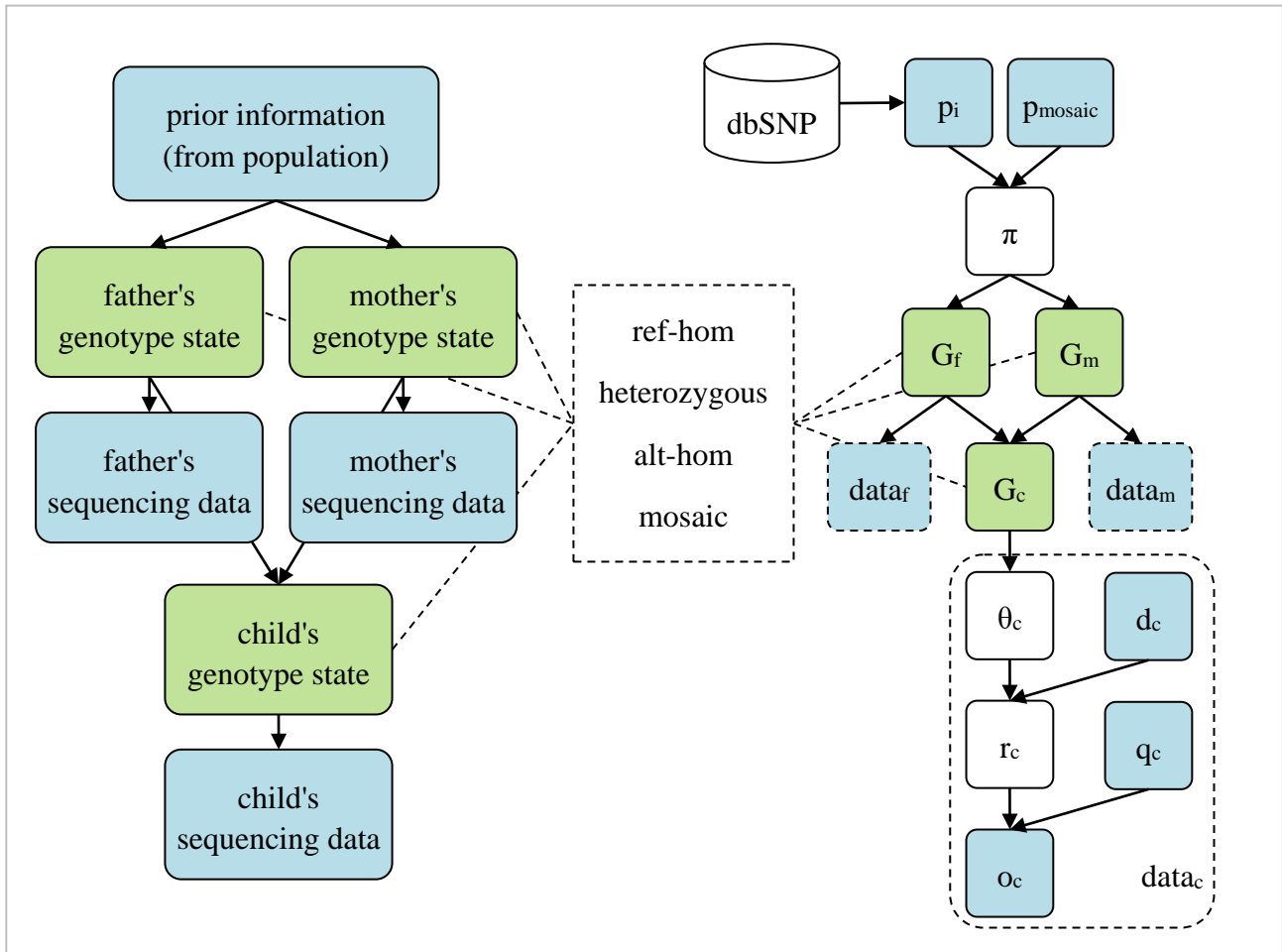


'Fisher paired' mode (mosaic\_filter's `mode=paired_fisher`) does not use the Bayesian model, but just compares the refDepth and altDepth in case and control samples using Fisher's exact test for 2\*2 contingency table. The mosaic\_filter will pass the sites with p-value  $< \text{fisher\_threshold}$ .

### 3) Trio mode

If you have sequenced the individual (child) with his/her either or both parents, you can try trio mode (mosaic\_filter's `mode=trio`, and set the corresponding `father_bam_file=<path>`, `mother_bam_file=<path>`).

The probabilistic graphical model for 'trio' mode can be shown in this figure.



The Bayesian model will pass the sites if the child has SNM (child's mosaic posterior probability  $>$  mosaic\_filter's parameter `mosaic_threshold`), or the Mendelian inheritance is violated (such as *de novo* mutation). Therefore, in 'trio' mode, you may need to further check the child's likelihood to judge whether the candidate site is a SNM (child's largest likelihood is mosaic state) or a *de novo* SNV (child's largest likelihood is heterozygous state).

#### I.3.b Whole-genome sequencing vs. whole-exome sequencing

(mean-shift & over-dispersion of allele fraction among heterozygous sites)

Another thing need to be taken into consideration is whether your data is WGS or WES. It has been noticed that the heterozygous sites follow the binomial distribution well in WGS, but poor in WES.



To address this issue, we replace the prior of theoretical AF  $\theta$   $P(\theta|G=\text{heterozygous})$  from an impulse at 0.5 to a beta prior distribution fitted from the WES data. This adjusting method has two steps: (1) traverse the bam file once, summarize the AF information of heterozygous sites, then output the estimated shape parameters ( $\alpha$ ,  $\beta$ ) for the beta prior.

In the 1st step, we applied several error filters to get a list of high-confidence heterozygous sites (mosaic\_filter's `mode=heterozygous`), and then used `exome_parameter_filter` to summarize the AF information, calculate statistics, and report the optimized shape parameters ( $\alpha$ ,  $\beta$ ) estimates in file `output_dir/stdout_*.log`.

In the 2nd step, MosaicHunter will call SNMs from the WES data with mosaic\_filter's parameters `alpha_param=<int>` and `beta_param=<int>`, which are estimated in the 1st step. If they are kept to their default value 0, then the original prior – impulse at 0.5 – will be used. If the estimated shape parameters ( $\alpha$ ,  $\beta$ ) are both very large ( $>1000$ ), and  $\alpha/(\alpha+\beta)$  is close to 0.5, which means that the beta prior is quite similar to the original impulse at 0.5 (no over-dispersion and mean-shift), we suggest to keep these parameters to the default 0, for running efficiency.

Except for the change in mosaic\_filter, we have also modified and added some other filters for WES, such as `mapping_quality_filter` and `syscall_filter` (see also I.3.b and II.2). For the `syscall_filter`, we adapt the parameters of the SysCall logistic model with different average depth (Meacham *et al. BMC Bioinformatics* 2011). You can check the output file `output_dir/stdout_*.log` of the 1st step to obtain the average depth.

## II Manual for Parameters and Filters

### II.1 Top Parameters

There are some parameters in the top level ('top parameters'). Here is the list with explanation.

`input_file=<path>`

The path of input .bam file

[Required]

`reference_file=<path>`

The path of reference file

[Required]

`output_dir=<path>`

The working directory for output files

[Required]

`valid_references=<str>`

The contig name list for calling mosaicisms (delimited by ','); ie. the contigs not in this list will be ignored.

[Default: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y]

`chr_X_name=<str>`

`chr_Y_name=<str>`

The contig name for sex chromosome X and Y

[Default: X]    [Default: Y]

`max_depth=<int>`

For efficiency, if the site has depth > this\_value, it will be trimmed (or down-sampled) to this\_value. (see also `depth_sampling=<bool>`)

[Default: 5001]

`base_quality=<int>`

The Phred shift of baseQ in the bam file

[Default: 33]

`min_mapping_quality=<int>`

The reads with mapQ < this\_value will be immediately discarded.

[Default: 20]

`min_read_quality=<int>`

The bases with baseQ < this\_value will be immediately discarded.

[Default: 20]

`remove_duplicates=<bool>`

If the reads marked duplicate (with flag 0x400) should be discarded.

[Default: true]

`remove_flag=<int>`

The reads with any bit on of this specified flag will be discarded. For example, to remove reads indicating secondary alignment, you can set `remove_flag=0x100`.

[Default: 0x0]

`seed=<int>`

The seed for random number generator; set a fixed number for replicability.

[Default: 0]

`depth_sampling=<bool>`

When the site has `depth > max_depth`, should I down-sampled it (true) or just trimmed it (false). (see also `max_depth=<int>`)

[Default: false]

`chr=<str>`

If you just want to call SNMs on one contig (chromosome) instead of all contigs, set this.

[Default: (empty)]

`start_position=<int>`

`end_position=<int>`

If you just focused on one block on one contig, you can set these two parameters.

(1-based, end-included)

[Default: 1] [Default: 300000000]

`read_buffer_size=<int>`

The number of reads which are read in buffer.

[Default: 100000]

`in_process_filter_name=<str>`

Overall in\_process filter's name (not recommend to change)

[Default: in\_process\_filter]

`post_process_filter_name=<str>`

Overall post\_process filter's name (not recommend to change)

[Default: post\_process\_filter]

## II.2 Filter Introduction and Parameters

We develop several stringent filters to remove artifacts due to technical errors as well as other types of genomic variations. For convenience, we call the Bayesian genotyper as mosaic\_filter in the java implementation.

No matter the filters are put into the categories of in\_process or post\_process, they will be checked sequentially. The only difference is that the in\_process filters work on each site independently whereas the post\_process filters can simultaneously handle all candidate sites altogether. The order of some post\_process filters may affect the final results. Any filters that can be put in the in\_process can also be put in the post\_process, although it may increase the memory burden and slow down the speed.

Each filter has a name and a class in our implementation, as some filters can actually apply the same algorithm arranged in the same class. Class names are fixed in the source code, while the filter names can be changed in the config.properties file. Filter names will be used as prefix in the output filenames. The names of passed filters for each site will be recorded in the memory, which may be subsequently used, especially for OrFilter (eg. mosaic\_and\_mosaic\_like\_filter), clustered\_filter, and near\_mosaic\_filter in the WGS analysis pipeline (see also I.3.b).

Each filter has two general parameters and some specific parameters for the filter class (see also II.2).

Here is a table of our developed filter classes with filter names.

Filter Class	Filter Name	Comment
<a href="#">AndFilter</a>	in_process_filter, post_process_filter, mosaic_common_filter	
<a href="#">OrFilter</a>	mosaic_and_mosaic_like_filter	

<a href="#">DepthFilter</a>	depth_filter	in_process
<a href="#">BaseNumberFilter</a>	base_number_filter, mosaic_like_filter	in_process
<a href="#">RegionFilter</a>	repetitive_region_filter, indel_region_filter, common_site_filter	in_process
<a href="#">HomopolymersFilter</a>	homopolymers_filter	in_process
<a href="#">MosaicFilter</a>	mosaic_filter, heterozygous_filter	in_process
<a href="#">StrandBiasFilter</a>	strand_bias_filter	in_process
<a href="#">WithinReadPositionFilter</a>	within_read_position_filter	in_process
<a href="#">CompleteLinkageFilter</a>	complete_linkage_filter	in_process
<a href="#">Nearmosaic_filter</a>	near_mosaic_filter	post_process
<a href="#">MisalignedReadsFilter</a>	misaligned_reads_filter	post_process
<a href="#">ClusteredFilter</a>	clustered_filter	post_process
<a href="#">OutputFilter</a>	final	post_process
<a href="#">ExomeParameterFilter</a>	exome_parameter_filter	in_process
<a href="#">NullFilter</a>	null_filter	in_process
<a href="#">MappingQualityFilter</a>	mapping_quality_filter	in_process
<a href="#">SysCallFilter</a>	syscall_filter	in_process

Below are details for general parameters and each filter class.

## General parameters

`output_filtered=<bool>`

If the corresponding output <filter\_name>.filtered.tsv is generated.

[Default: false]

`output_passed=<bool>`

If the corresponding output <filter\_name>.passed.tsv is generated.

[Default: false]

### 1) AndFilter

[in\_process\_filter, post\_process\_filter, mosaic\_common\_filter]

This logic filter registers some more filters. Only the sites passed every the registered filters will be

passed by this filter. The registered filters will be checked sequentially.

```
filters=<str>
```

The registered filter names, separated by ',', and no space ' ' is allowed.

[Default: (empty)]

## 2) OrFilter

[mosaic\_and\_mosaic\_like\_filter]

This logic filter registers some more filters. The sites passed any of the registered filters will be passed by this filter. The registered filters will be checked sequentially.

The passed filter names for each site will be recorded in the memory, which may be further used.

```
filters=<str>
```

The registered filter names, separated by ',', and no space ' ' is allowed.

[Default: (empty)]

## 3) DepthFilter

[depth\_filter]

This filter will pass sites with depth in the specified range.

```
min_depth=<int>
```

[Default: 25]

```
max_depth=<int>
```

[Default: 150]

## 4) BaseNumberFilter

[base\_number\_filter, mosaic\_like\_filter]

This filter checked the observed minor allele count and minor AF. The sites with minor allele in specified range will be passed.

`min_minor_allele_number=<int>`

[Default: 5]

`min_minor_allele_percentage=<int>`

[Default: 5]

`max_minor_allele_percentage=<int>`

[Default: 100]

## 5) **RegionFilter**

[repetitive\_region\_filter, indel\_region\_filter, common\_site\_filter]

This filter will filter out sites inside or outside the specified region blocks.

Output tsv (filtered.tsv) 11+ columns (see also II.3):

11-13: chr, start, end of the region block (1-based, end-included)

`bed_file=<path>`

The specified regions .bed file.

[Required]

`expansion=<int>`

The number of bp to expand for each block.

[Default: 5]

`include=<bool>`

If you want to filter out sites within the region blocks (false) or outside the region blocks (true).

[Default: false]

## 6) HomopolymersFilter

[homopolymers\_filter]

This filter will filter out sites near homopolymers on the reference sequence.

`short_homopolymer_length=<int>`

Define the length of 'short' homopolymer.

[Default: 4]

`short_homopolymer_expansion=<int>`

The number of bp for expanding each 'short' homopolymer, ie. filter out those sites close to a 'short' homopolymer.

[Default: 2]

`long_homopolymer_length=<int>`

Define the length of 'long' homopolymer

[Default: 6]

`long_homopolymer_expansion=<int>`

The number of bp for expanding each 'long' homopolymer, ie. filter out those sites close to a 'long' homopolymer.

[Default: 3]

## 7) MosaicFilter

[mosaic\_filter]

The main Bayesian genotyper for SNMs calling (the mosaic caller) (see also I.2)

Output tsv (filtered.tsv, passed.tsv) will be different for different mode (single, paired, trio mode), see also II.3.

`min_read_quality=<int>`

Ignore bases with baseQ < this\_value on the site



[Default: 20]

`min_mapping_quality=<int>`

Ignore bases with mapQ < this\_value on the site

[Default: 20]

`dbSNP_file=<path>`

A tab-separated file providing information in dbSNP, which will provide reference/alternative allele frequency (prior information). It should have six columns:

1: chr

2: pos

3: rsId

4: reference allele (refBase)

5: alternative allele (altBase)

6: alternative allele frequency; -1 means that the variant is recorded in dbSNP but has no allele frequency information.

[Required]

`unknown_af=<double>`

The specified population allele frequency for those variants that are recorded in dbSNP but no allele frequency information.

[Default: 0.002]

`novel_af=<double>`

The specified population allele frequency for those variants without records in dbSNP.

[Default: 1e-8]

`mosaic_rate=<double>`

The mosaic prior probability.

[Default: 1e-7]

`de_novo_rate=<double>`

The probability for a de novo mutation occurring.

[Default: 1e-8]

`sex=<str>`

'M' for male, and 'F' for female. This affects mosaic calling for sex chromosomes.

`alpha_param=<int>`

`beta_param=<int>`

For exome sequencing, we observed over-dispersion on AF of heterozygous sites, so we change the prior of theoretically AF  $\theta$  from impulse at 0.5 to a beta distribution, with two shape parameters ( $\alpha$ ,  $\beta$ ) specified here. The two shape parameter ( $\alpha$ ,  $\beta$ ) can be estimated by running MosaicHunter with `exome_parameters.properties` (details see also I.2.b).

The default value (0, 0) means trigger off this part, and just use the impulse at  $\theta = 0.5$ . In fact, to trigger on the heterozygous beta prior, both parameters should be  $> 0$  and their sum should be  $> 2$ .

For efficiency issue, we suggest that both parameters should be  $\leq 1000$ . For larger estimated parameters, which indicates closeness of the beta prior to the  $\theta = 0.5$  impulse, we suggest to trigger off this part by setting the default (0, 0), and turn back to the  $\theta = 0.5$  impulse prior.

[Default: 0]     [Default: 0]

`base_change_rate=<16*double>`

If you want to specify a prior on mosaic base change frequency, you can specify it here. It contains 16 double values(separated by ',') corresponding to A->A,C,G,T; C->A,C,G,T; G->A,C,G,T; T->A,C,G,T. You can specify the relative rate, as they will later be row-wisely normalized to sum to 1.

[Default: 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]

`mosaic_threshold=<double>`

The sites with mosaic posterior (or heterozygous posterior if `heterozygous=true`)  $>$  `this_value` will be passed.

[Default: 0.05]

`mode=<single | trio | paired_naive | paired_fisher | heterozygous>`

Specify the Bayesian mode of this filter, which can be 'single', 'trio', 'paired\_naive', 'paired\_fisher' or 'heterozygous'. (details see also I.3)

[Default: single]

`father_bam_file=<path>`

`mother_bam_file=<path>`

If you specify the 'trio' mode (`mode=trio`), you also need to specify these parameters. Then the mosaic calling in the child will also consider parents' sequencing data.

[Default: (empty)]

`control_bam_file=<path>`

If you specify the 'paired' mode (`mode=paired_naive` or `mode=paired_fisher`), you also need to specify these parameters.

[Default: (empty)]

`fisher_threshold=<double>`

If you specify the 'Fisher paired' mode (`mode=paired_fisher`), then you need to specify this parameter. The filter will pass sites with the p-value of Fisher's exact test  $< \text{this\_value}$ .

## 8) StrandBiasFilter

[strand\_bias\_filter]

This filter will apply Fisher's exact test on strand bias (major strand+, major strand-, minor strand+, minor strand-), and filters out site with small p-value.

Output tsv (filtered.tsv, passed.tsv) 11+ columns (see also II.3):

11-14: count of major +, minor + (forward strand), major -, minor - (reverse strand)

15: p-value of Fisher's exact test

`min_p_value=<double>`

The sites with p-value  $\geq \text{this\_value}$  will be passed.

[Default: 0.05]

## 9) WithinReadPositionFilter

[within\_read\_position\_filter]

This filter will apply Wilcoxon rank-sum test (ie. Mann-Whitney U test) on positions within reads according to the reference orientation (major pos, minor pos), and filters out sites with small p-value.

Output tsv (filtered.tsv, passed.tsv) 11+ columns (see also II.3):

11: major read position

12: minor read position

13: p-value of Wilcoxon rank-sum test (ie. Mann-Whitney U test)

`min_p_value=<double>`

The sites with p-value  $\geq$  this\_value will be passed.

[Default: 0.05]

## 10) CompleteLinkageFilter

[complete\_linkage\_filter]

If there are more SNV(s) on the same read with mosaic sites, we would expect that the mosaic site should not be randomly linked or completely linked to the SNV site, which may be filtered by this filter.

Output tsv (filtered.tsv) 11+ columns (see also II.3):

11: the close position with a possible SNV

12-16: count of four kinds of haplotypes

17: p-value of Fisher's exact test

`max_p_value=<double>`

We do Fisher's exact test on the 2\*2 contingency table of count of four kinds of haplotypes. The sites with p-value  $>$  this\_value will be filtered.

[Default: 0.01]

## 11) NearMosaicFilter

[near\_mosaic\_filter]

This filter drops auxiliary ('mosaic-like') sites far away from focused ('mosaic candidate') sites.

`distance=<int>`

The maximum distance allowed between passed mosaic-like sites and a mosaic candidate site.

[Default: 10000]

`auxiliary_filter_name=<str>`

The passed filter name marking the auxiliary sites.

[Default: mosaic\_like\_filter]

## 12) MisalignedReadsFilter

[misaligned\_reads\_filter]

Call blat to check read misalignment (very slow).

Passed sites require a enough large proportion of reads with high-confidence alignment (see also `max_misalignment_percentage=<double>`): 1) blat mapping result is unique, and agrees with .bam mapping result (eg. by bwa) (see also `min_overlap_percentage=<double>`); 2) the candidate site is not near two ends (see also `min_side_distance=<int>`) or gaps (see also `min_gap_distance=<int>`) of the read.

Output tsv (filtered.tsv, passed.tsv) 11+ columns (see also II.3):

11: ALIGNMENT\_OK

12: ALIGNMENT\_MISSING

13: MULTIPLE\_ALIGNMENTS

14: CHROM\_MISMATCH

15: ALIGNMENT\_MISMATCH

16: NEAR\_SIDE

17: NEAR\_GAP

18: NM

19: misalignment

20: alignment\_ok proportion among major allele

21: alignment\_ok proportion among minor allele

`blat_path=<path>`

The path of runnable blat.

[Default: blat]

`blat_param=<str>`

The additional parameters when calling blat.

[Default: -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0.5 -noHead]

`reference_file=<path>`

The reference fasta file for blat, which can be different from the top parameter `reference_file`. We suggest putting all possible contigs in this file, such as chr6 HLA haplotypes, to reduce the misalignment due to sequence difference.

[Default: Same as top parameter `reference_file`]

`max_misalignment_percentage=<double>`

Passed sites should have proportion of 'misaligned' reads < this\_value ('good' reads > (1-this\_value)).

[Default: 0.5]

`min_overlap_percentage=<double>`

Each read will have one mapping block by blat and one from .bam (eg. by bwa). If the proportion of intersection of two mapping blocks on either block is < this\_value, the read will be regarded as 'misaligned'.

[Default: 0.9]

`min_side_distance=<int>`

If the candidate site is near the ends of the read (the end base corresponding to distance = 1), the read will be regarded as 'misaligned'.

[Default: 15]

`min_gap_distance=<int>`

If the candidate site is near the mapping gap of the read (the marginal base corresponding to

distance = 1), the read will be regarded as 'misaligned'.

[Default: 5]

`max_NM=<int>`

If the NM tag of the read in .bam is > this\_value, the read will be regarded as 'misaligned'.

### 13) CluteredFilter

[clustered\_filter]

Because some region on the genome may enrich sequencing and mapping artifacts, we see some false-positive candidate sites clustered (with close position). This filter will remove main ('mosaic') candidate sites which are clustered with other main ('mosaic') or auxiliary ('mosaic-like') sites.

Output tsv (passed.tsv, filtered.tsv) 11+ columns (see also II.3):

11: the positions of other close clustered sites

`inner_distance=<int>`

If three neighbor mosaic candidate sites are with distance < this\_value bp, they are regarded as clustered sites.

[Default: 20000]

`outer_distance=<int>`

For any found clustered sites, we will expand the boundary of filtering region to upstream and downstream this\_value bp, and the mosaic candidate sites falling in the region will be filtered.

[Default: 20000]

`auxiliary_filter_name=<str>`

The passed filter name marking the auxiliary sites.

[Default: mosaic\_like\_filter]

### 14) OutputFilter

[final]

The final 'filter' for outputting the result list of the passed candidate sites.

## 15) ExomeParameterFilter

[exome\_parameter\_filter]

This 'filter' does statistics on passed sites (assuming heterozygous sites), and finally reports to `output_dir/stdout_*.log` the average depth and the estimated (optimized) shape parameters( $\alpha$ ,  $\beta$ ) for exome heterozygous AF prior.

`min_group_size=<int>`

When doing statistics on standard deviation of AF, we need to group together sites with close depth, if the mount of sites with the same depth is not large enough (< this\_value).

[Default: 50]

`optimal_depth=<int>`

When the observed relationship between  $\text{std}(\text{AF}) \sim \text{depth}$  cannot be exactly fitted by the model formula, I will try to optimally fit the two values of  $\text{depth} = \text{this\_value}$ .

[Default: 80]

`r_data_file=<filename>`

The temporary file for counting, which will then be read in again and fitted.

[Default: "r\_het\_data.tsv"]

## 16) NullFilter

[null\_filter]

This filter will discard all the sites, thus to clear up memory. Only useful after the filter doing statistics or estimating parameters (eg. ExomeParameterFilter).

## 17) MappingQualityFilter

[mapping\_quality\_filter]

This filter will apply Wilcoxon rank-sum test (ie. Mann-Whitney U test) on reads' mapQs (major mapQs, minor mapQs), and filters out sites with small p-value.



`min_p_value=<double>`

The sites with p-value  $\geq$  this\_value will be passed.

[Default: 0.05]

## 18) SysCallFilter

[syscall\_filter]

This filter applies the logistic regression model developed and introduced by SysCall (Meacham *et al. BMC Bioinformatics* 2011), to filter out exome false positives. The trained parameters varied across different depth of exome sequencing data, so the average depth should be specified.

`depth=<int>`

The average depth of exome sequencing data, which can be estimated by running MosaicHunter with `exome_parameters.properties` (details see also I.2.b and I.3.b).

[Default: 66]

## II.3 Output Files and Format

All the output files are in the specified directory `output_dir`, including those requested `<filter_name>.filtered.tsv`, `<filter_name>.passed.tsv` files, and some other temporary files, such as blat input and output.

`<filter_name>.filtered.tsv` and `<filter_name>.passed.tsv` files are in tab-separated format, whose columns' meaning is listed below:

1: chr	the contig/chromosome name
2: pos	the position/coordinate on the contig (1-based)
3: refBase	the base of reference allele
4: depth	the sequencing depth of this site
5: bases	pileup sequencing bases at this site
6: baseQs	pileup sequencing baseQs at this site
7: majorBase	the base of major allele
8: majorDepth	the depth of major allele

9: minorBase            the base of minor allele  
10: minorDepth        the depth of minor allele

11+ columns are dependent on filters (see also Output tsv 11+ part for each filter in II.2).  
final.passed.tsv is the same as mosaic\_filter.passed.tsv, which will be different for different mode.

### **1) Single mode**

11: major : minor allele, with dbSNP allele frequency in population shown in parentheses  
12-15: log10 prior of major-homozygous, heterozygous, minor-homozygous, mosaic  
16-19: log10 likelihood of major-homozygous, heterozygous, minor-homozygous, mosaic  
20-23: log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic  
24: mosaic posterior probability

### **2) Naive paired mode**

11: case ( major allele : major depth, minor allele : minor depth )  
12: control ( major allele : major depth, minor allele : minor depth )  
13-16: case log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic  
17-20: control log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic  
21: case sum of posterior probabilities other than control's most probable genotype state  
22: control posterior probability of control's most probable genotype state

### **3) Fisher paired mode**

11: case ( major allele : major depth, minor allele : minor depth )  
12: control ( major allele : major depth, minor allele : minor depth )  
13-16: case log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic  
17-20: control log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic  
21: p-value of Fisher's exact test

#### **4) Trio mode**

11: child major : minor allele, with dbSNP allele frequency in population shown in parentheses

12: father ( major allele : major depth, minor allele : minor depth )

13: mother ( major allele : major depth, minor allele : minor depth )

14-17: child log10 likelihood of major-homozygous, heterozygous, minor-homozygous, mosaic

18-21: father log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic

22-25: mother log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic

26-29: child log10 posterior of major-homozygous, heterozygous, minor-homozygous, mosaic

30: child mosaic posterior probability

31: child log10 likelihood ratio of mosaic vs. heterozygous

(log10 likelihood of mosaic – log10 likelihood of heterozygous)

#### **5) Heterozygous mode**

11: major : minor allele, with dbSNP allele frequency in population shown in parentheses

12-14: log10 prior of major-homozygous, heterozygous, minor-homozygous

15-17: log10 likelihood of major-homozygous, heterozygous, minor-homozygous

18-20: log10 posterior of major-homozygous, heterozygous, minor-homozygous

21: heterozygous posterior probability

### **III Miscellaneous**

#### **III.1 Contact Information**

Team:

August Yue Huang

Adam Yongxin Ye

Zheng Zhang

Yanmei Dou

Liping Wei (PI)

@Center for Bioinformatics, Peking University

@Liping Wei's Lab, National Institute of Biological Science

Contact us:

Room 307, Wangkezhen Building

Center for Bioinformatics

Peking University

Beijing, 100871

P. R. China

Email: [mosaichunter@mail.cbi.pku.edu.cn](mailto:mosaichunter@mail.cbi.pku.edu.cn)

## **III.2 Acknowledgement**

We thank Drs. Yu Shyr, Jinzhu Jia, and Cheng Li for their valuable suggestions and discussions about the statistical models.

Acknowledge to Viktor Persson for the "Indigo" template of the web interface.

We thank Li-Ming Xu for the source codes of the web pages.

We thank Chaozhi Hu for the design of MosaicHunter's logo.

## **III.3 License**

MosaicHunter is licensed under the MIT License.

## **III.4 FAQ**