mission bio

Advanced Filtering

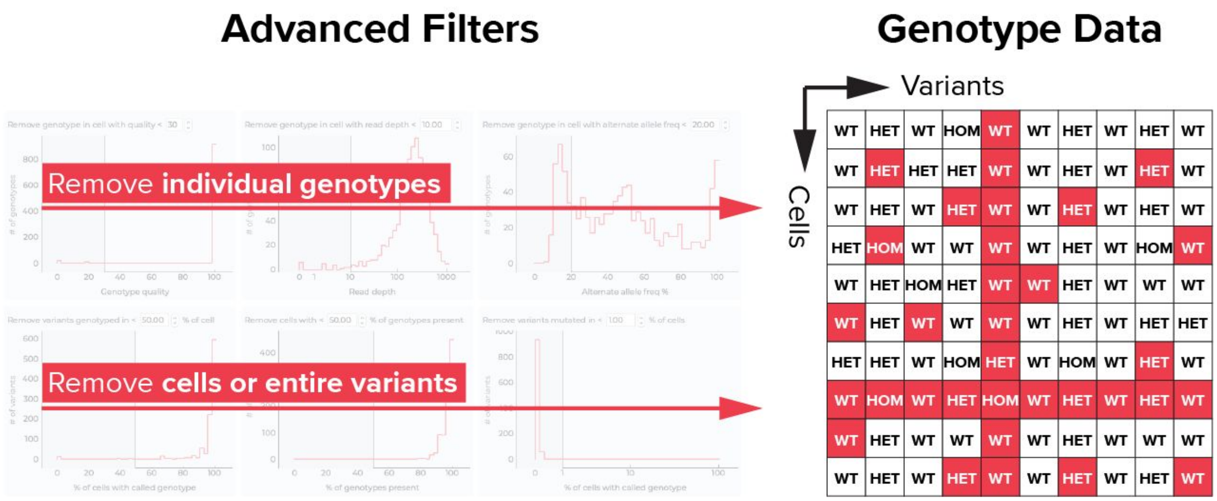
About This Document

This document provides an overview of the **Advanced Filters** that are part of the Tapestri Insights software. For each filter we provide a short description as well as recommendations to adjust filter thresholds in particular use cases.

Introduction

Tapestri Insights software is equipped with a set of quality filters that empowers you to flexibly filter your single cell DNA data to guarantee robust genotype calling and accurate results. Quality filters include *genotype quality*, *read coverage*, *mutant (alternate) allele frequency* as well as *percentage of mutated cells per variant*.

Filters remove either individual genotype information (top row filters in Advanced Filtering window) or remove cells/variants (bottom row filters in Advanced Filtering window).


Advanced Filters
Genotype Data
Remove genotype in cell with quality < 30
Remove genotype in cell with read depth < 10.00
Remove genotype in cell with alternate allele freq < 20.00
Remove individual genotypes
Genotype quality
Read depth
Alternate allele freq %
Remove variants genotyped in < 50.00 % of cell
Remove cells with < 50.00 % of genotypes present
Remove variants mutated in < 1.00 % of cells
Remove cells or entire variants
% of cells with called genotype
% of genotypes present
% of cells with called genotype
Variants
Cells

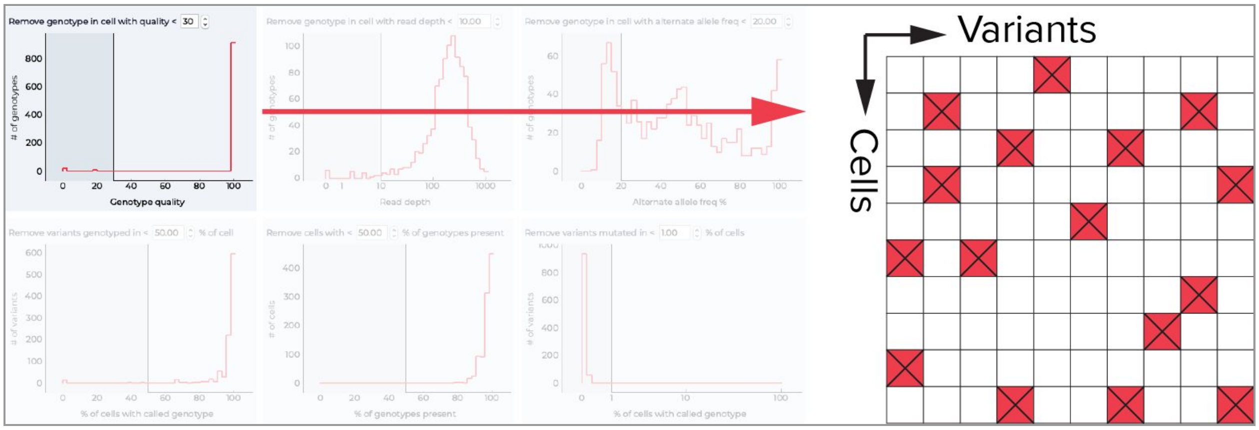
WT	HET	WT	HOM	WT	WT	HET	WT	HET	WT
WT	HET	HET	HET	WT	WT	HET	WT	HET	WT
WT	HET	WT	HET	WT	WT	HET	WT	HET	WT
HET	HOM	WT	WT	WT	WT	HET	WT	HOM	WT
WT	HET	HOM	HET	WT	WT	HET	WT	WT	WT
WT	HET	WT	WT	WT	WT	HET	WT	HET	HET
HET	HET	WT	HOM	HET	WT	HOM	WT	HET	WT
WT	HOM	WT	HET	HOM	WT	HET	WT	HET	WT
WT	HET	WT	WT	WT	WT	HET	WT	WT	WT
WT	HET	WT	HET	WT	WT	HET	WT	HET	WT

Default threshold values for each filter have been defined based on analyzing single-cell DNA data derived from various cell line experiments as well as select primary sample studies.

Note, if multiple samples are loaded in Tapestri Insights (e.g., multiple samples from longitudinal study), genotypes, cells, or variants are only removed from downstream analysis if they do not meet the filter criteria in **<u>all</u>** samples. If in at least one loaded sample the quality metrics meet all filter criteria, genotype/variant/cell information is available for all other loaded samples as well.

All Six Filters Explained

Remove genotype in cell with quality < X



Remove genotype in cell with quality < 30

Remove genotype in cell with read depth < 10.00

Remove genotype in cell with alternate allele freq < 20.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

Data: Genotype Quality (GQ)

Data:	Genotype Quality (GQ)
Range:	0 - 99
Default:	30
What is removed:	Cell-specific genotypes based on genotype quality

Definition:

The Genotype Quality (GQ) metric is derived from the PL value of the VCF file.
PL is a sample-level annotation calculated by GATK variant callers such as HaplotypeCaller, recorded in the FORMAT/sample columns of variant records in VCF files. This annotation represents the normalized Phred-scaled likelihoods of the genotypes considered in the variant record for each sample, as described here.

This article clarifies how the PL values are calculated and how this relates to the value of the GQ field.

The basic formula for calculating PL is:

$$ PL = -10 * \log{P(Data | Genotype)} $$

where P(Data | Genotype) is the conditional probability of the Genotype given the sequence Data that we have observed. The process by which we determine the value of P(Data | Genotype) is described in the genotyping section of the Haplotype Caller documentation.

Once we have that probability, we simply take the log of it and multiply it by -10 to put it into Phred scale. Then we normalize the values across all genotypes so that the PL value of the most likely genotype is 0, which we do simply by subtracting the value of the lowest PL from all the values.

NOTE*: The reason we like to work in Phred scale is because it makes it much easier to work with the very small numbers involved in these calculations. One thing to keep in mind of course is that Phred is a log scale, so whenever we need to do a division or multiplication operation (e.g. multiplying probabilities), in Phred scale this will be done as a subtraction or addition.*

Here's a worked-out example to illustrate this process. Suppose we have a site where the reference allele is A, we observed one read that has a non-reference allele T at the position of interest, and we have in hand the conditional probabilities calculated by HaplotypeCaller based on that one read (if we had more reads, their contributions would be multiplied -- or in log space, added).

NOTE*: Please note that the values chosen for this example have been simplified and may not be reflective of actual probabilities calculated by Haplotype Caller.*

```
# Alleles
Reference: A
Read: T

# Conditional probabilities calculated by HC
P(AA | Data) = 0.000001
P(AT | Data) = 0.000100
P(TT | Data) = 0.010000
```

Calculate the raw PL values

We want to determine the PLs of the genotype being 0/0, 0/1, and 1/1, respectively. So we apply the formula given earlier, which yields the following values:

Genotype	**A/A**	**A/T**	**T/T**
Raw PL	-10 * log(0.000001) = 60	-10 * log(0.000100) = 40	-10 * log(0.010000) = 20

Our first observation here is that the genotype for which the conditional probability was the highest turns out to get the lowest PL value. This is expected because, as described in the VCF FAQ, the PL is the *likelihood* of the genotype, which means (rather unintuitively if you're not a stats buff) it is the probability that the genotype is **not** correct. So, low values mean a genotype is more likely, and high values means it's less likely.

Normalize

At this point we have one more small transformation to make before we emit the final PL values to the VCF: we are going to **normalize** the values so that the lowest PL value is zero, and the rest are scaled relative to that. Since we're in log space, we do this simply by subtracting the lowest value, 20, from the others, yielding the following final PL values:

Genotype	**A/A**	**A/T**	**T/T**
Normalized PL	60 - 20 = 40	40 - 20 = 20	20 - 20 = 0

We see that there is a direct relationship between the scaling of the PLs and the original probabilities: we had chosen probabilities that were each 100 times more or less likely than the next, and in the final PLs we see that the values are spaced out by a factor of 20, which is the Phred-scale equivalent of 100. This gives us a very convenient way to estimate how the numbers relate to each other -- and how reliable the genotype assignment is -- with just a glance at the PL field in the VCF record.

Genotype quality

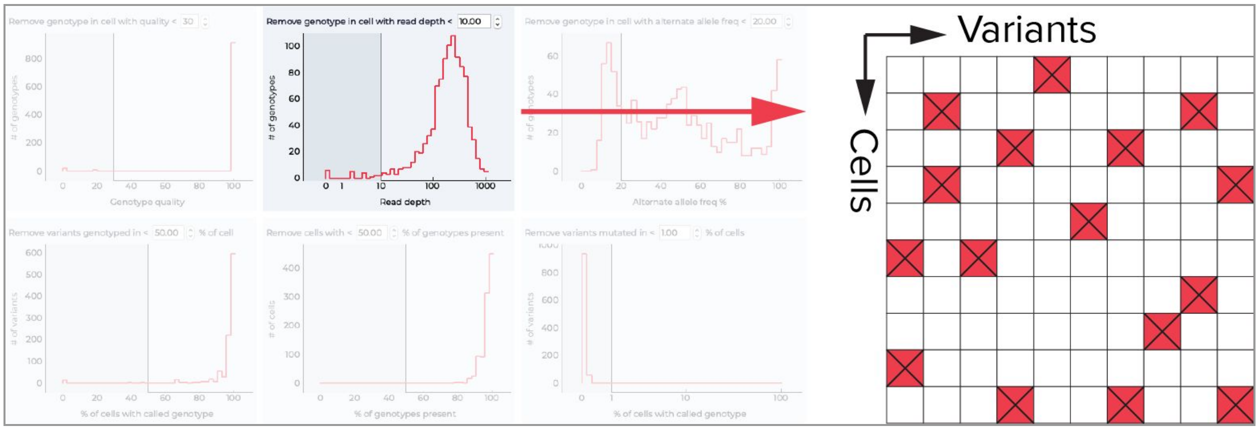
We actually formalize this assessment of genotype quality in the **GQ annotation**, as described also in the VCF FAQ. The value of GQ is simply the difference between the second lowest PL and the lowest PL (which is always 0). So, in our example GQ = 20 - 0 = 20. Note that the value of GQ is capped at 99 for practical reasons, so even if the calculated GQ is higher, the value emitted to the VCF will be 99.

Source: https://gatkforums.broadinstitute.org/gatk/discussion/5913/math-notes-how-pl-is-calculated-in-haplotypecaller

Exceptions:

Lowering the threshold value	**Raising** the threshold value
Not recommended.	If "Remove variants mutated in < X % of cells" filter is lowered < 1 % (see below). In this case only highest-quality genotypes are recommended to be considered.

Remove genotype in cell with read depth < X



Remove genotype in cell with quality < 30

Remove genotype in cell with read depth < 10.00

Remove genotype in cell with alternate allele freq < 20.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

Data: Read Depth

Range: 0 - maximum sequencing reads (sequencer dependent)

Default: 10

What is removed: Cell-specific genotypes based on read depth

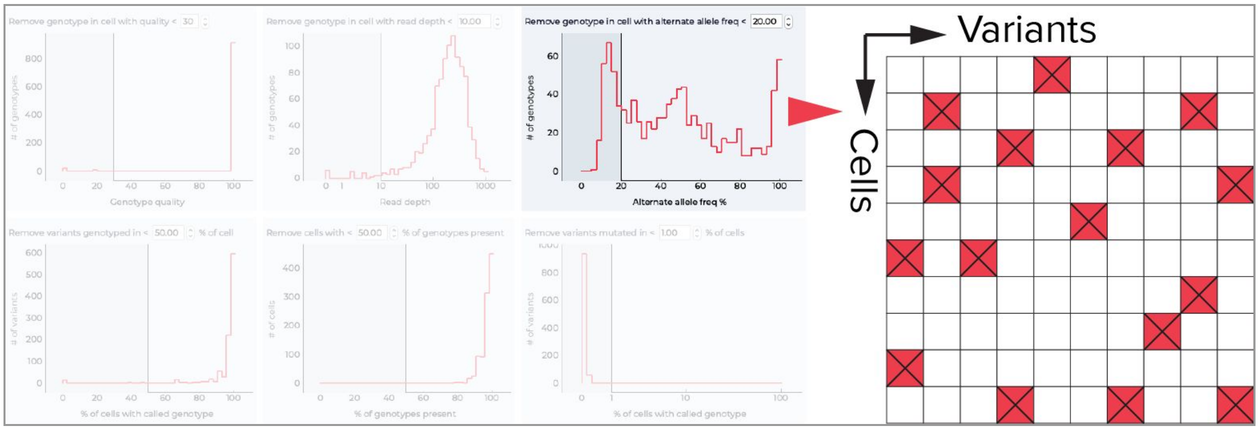
Definition:

This filter removes the genotype from all the cells that have read depth less than the given value. The read depth per variant metric is the filtered depth at the cell level. This shows the number of filtered reads that support each of the reported alleles.

Exceptions:

Lowering the threshold value	**Raising** the threshold value
If sample(s) is undersequenced (e.g., average reads/cell/variant < 30)	Not recommended.

Remove genotype in cell with alternate allele freq < X



Remove genotype in cell with quality < 30

Remove genotype in cell with read depth < 10.00

Remove genotype in cell with alternate allele freq < 20.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

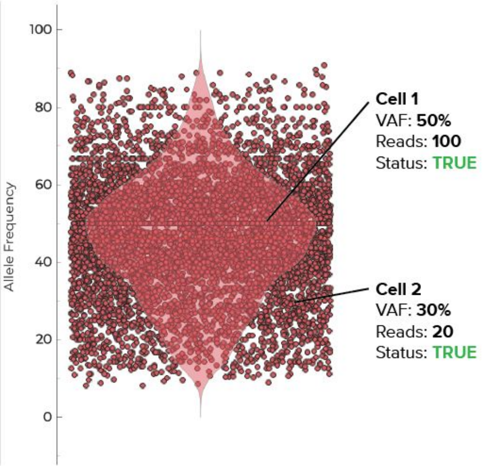
Data: Alternate Allele Frequency
Range: 0 - 100
Default: 20
What is removed: Cell-specific genotypes based on alternate allele frequencies

---

Definition:

This filter removes the genotype from all the cells that have alternate (mutant) variant allele frequency (aVAF) percentages less than the given value. Only non-reference genotype calls (heterozygous and homozygous) are included and displayed in the histogram (e.g., reference [wildtype] genotype calls are removed to help assess the mutant genotype distribution; the majority of genotypes are reference [wildtype] and would result in a large peak at 0 %).

Ideal VAF distribution of **heterozygous** mutant variant



Cell 1
VAF: **50%**
Reads: **100**
Status: **TRUE**

Cell 2
VAF: **30%**
Reads: **20**
Status: **TRUE**

The filter aims to remove low-quality mutant genotypes from the data that are associated with ‘out of range’ variant allele frequency information. Assuming diploidy and no copy number aberrations (CNA) all single-cell VAFs associated to heterozygous mutant calls are expected to follow a normal distribution with a peak at 50 % and symmetric tails on both sides. Due to sampling noise, typically cells with VAFs closer to 50 % are linked to a higher number of reads that support the VAF (e.g., 100 reads) compared to cells with VAFs closer to the tail region (e.g., 30 % VAF with 20 reads). Single-cell variants that are genotyped mutant with less than 20% VAF are considered low-quality and are filtered out.

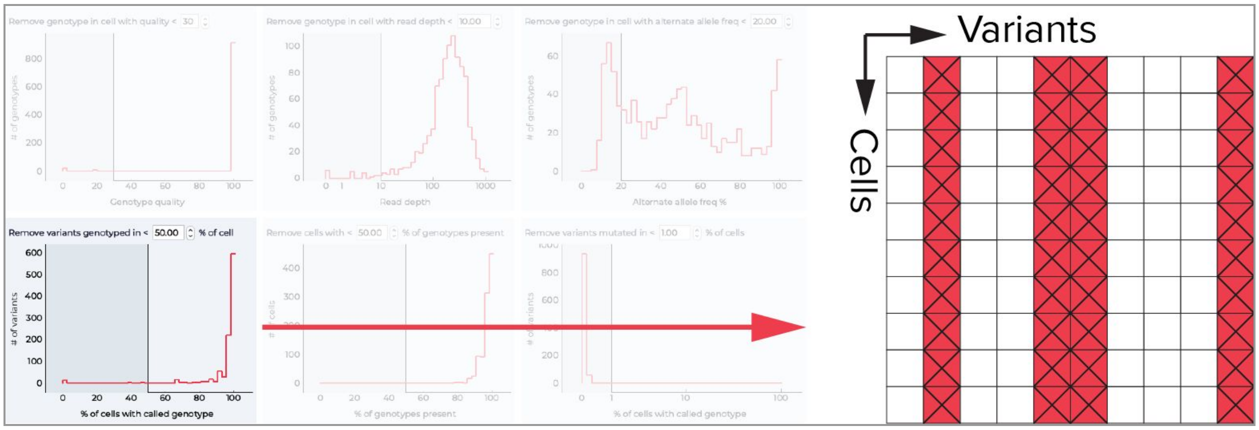
Factors that may contribute to skewed VAFs include (but are not limited to) sequencing error, PCR amplification error, genotype error.

Note, that this filter does not remove potential variants that are implicated in copy number variation, where the wildtype allele is lost (LOH, loss of heterozygosity) or the mutant allele amplifies (VAF % above 50 %).

Exceptions:

Lowering the threshold value	**Raising** the threshold value
In rare occasions to investigate potential false-positive variants. Please contact Mission Bio Support at support@missionbio.com for additional information.	Not recommended.

Remove variants genotyped in < X % of cells



Remove genotype in cell with quality < 30

Remove genotype in cell with read depth < 10.00

Remove genotype in cell with alternate allele freq < 20.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

Data:	Variants
Range:	0 - 100
Default:	50
What is removed:	Variants

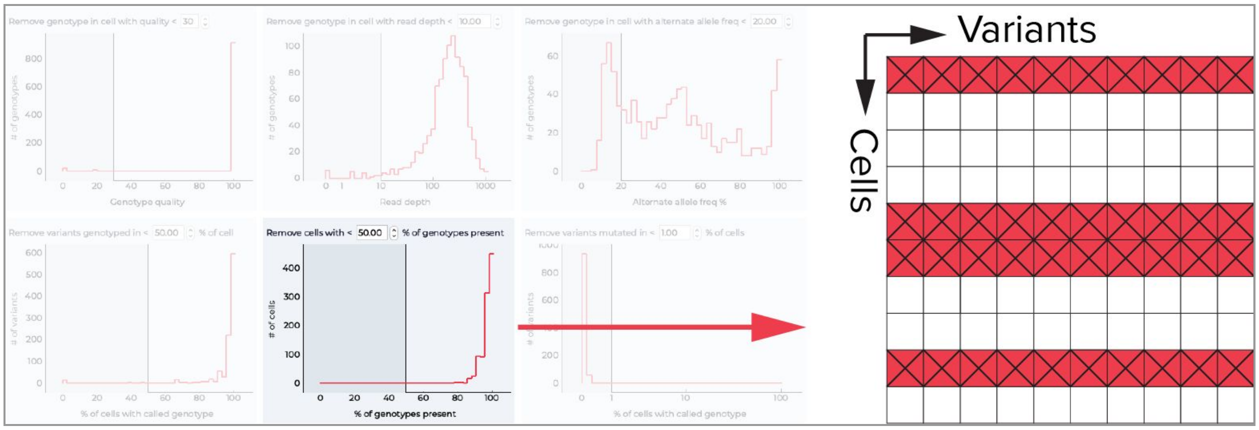
Definition:

This metric is the proportion of cells that have genotype information available. For example, a threshold of 50 % retains only variants for which information is available in at least 50 % of all cells. Variants with information in fewer than 50 % of cells are removed.

Exceptions:

Lowering the threshold value	**Raising** the threshold value
Recommended to recover variants that reside on amplicons with below-average performance (e.g., low average coverage due to GC-rich region). See ***Section 2 - Review Coverage Quality*** in this Knowledge Base article for additional information.	Not recommended.

Remove cells with < X % of genotypes present



Remove genotype in cell with quality < 30

Remove genotype in cell with read depth < 10.00

Remove genotype in cell with alternate allele freq < 20.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

Data:	Cells
Range:	0 - 100
Default:	50
What is removed:	Cells

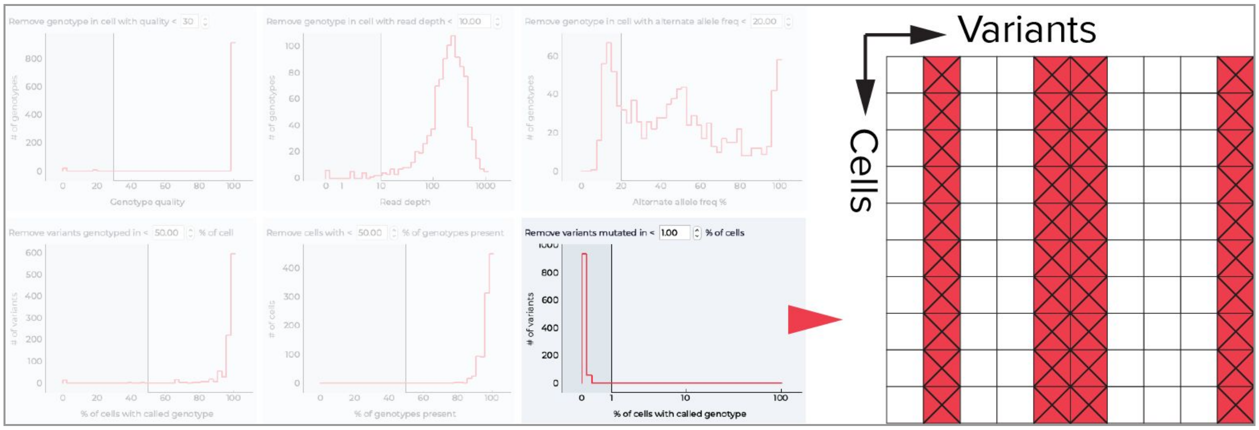
Definition:

This metric is the proportion of genotype information available on a per-cell basis. For example, a threshold of 50 % retains only the cells with at least 50 % of genotype information available for any given variant.

Exceptions:

Lowering the threshold value	**Raising** the threshold value
Not recommended. Please contact Mission Bio Support at support@missionbio.com for additional information.	Not recommended.

Remove variants mutated in < X % of cells



Remove genotype in cell with quality < 30

Remove genotype in cell with read depth < 10.00

Remove genotype in cell with alternate allele freq < 20.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

Data:	Sensitivity
Range:	0 - 100
Default:	1
What is removed:	Variants

Definition:

This metric is the percentage of cells across all cells with a genotype called as a non-reference genotype, e.g., heterozygous or homozygous alternate. Any variant with cells genotyped as HET or HOM ALT in less than the given threshold value would be discarded from the analysis.

The select threshold is dependent on the sample type and the scientific question one intends to address. For example, rare subclone detection requires a lower threshold.

NOTE*: GATK/Haplotypecaller is used at the Tapestri Pipeline level to call genotypes. It is expected that the majority of all called raw variants are at < 1 %. Multi-sample analysis is recommended (e.g., longitudinal time series) to recover rare variants (< 1 %).*

Exceptions:

Lowering the threshold value	**Raising** the threshold value
If targeted variant(s) is undetectable and expected to be present in rare fractions of cells, lower threshold as low as 0.1 %. Note that decreasing this filter threshold may include a high number of potential false-positive variants.	Not needed.

Summary

Metric	Genotype Quality	Read Depth	Alternate Allele Freq	% cells w/ genotype	% of genotypes present	% mutant cells
Unit	N/A	Reads/cell	Percentage	Percentage	Percentage	Percentage
Range	0 - 99	0 - X	0 - 100	0 - 100	0 - 100	0 - 100
Default Value	30	10	20	50	50	1
What is removed	Genotypes	Genotypes	Genotypes	Variants	Cells	Variants

X = maximum sequencing reads/cell

Please note that the term ***Genotypes*** refers to individual genotypes (WT, HET, HOM) of particular variants in particular cells, whereas the term ***Variants*** refers to all genotypes of one particular variant across all cells.

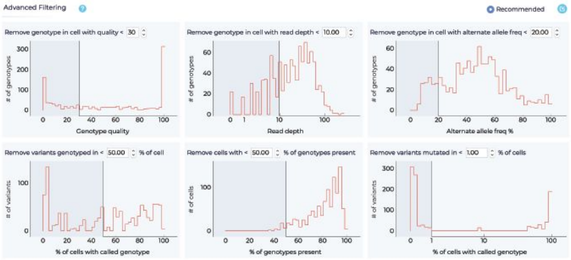
Understanding the relationship between filters

All three top-row filters condition the data displayed in all three bottom-row filters.

For example, if you raise the genotype quality filter from 30 to 80, the read depth filter from 10 to 20, and the alternate allele frequency filter from 20 to 40, an increased number of variants and cells is discarded for downstream analysis, which is reflected by the altered histogram distributions of the bottom-filters.

Note, we do not recommend to lower all filer thresholds to 0 as this will typically include hundreds of variants and result in increased computation time.

1 Changing **Top 3 default** Advanced Filters



Advanced Filtering

● Recommended

Remove genotype in cell with quality < 30

Remove genotype in cell with read depth < 10.00

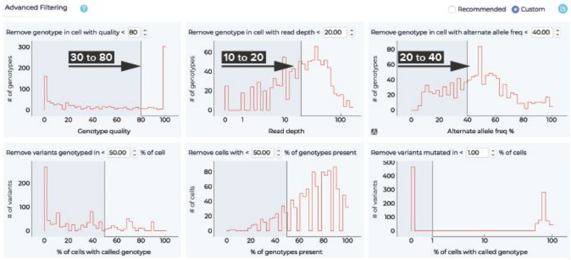
Remove genotype in cell with alternate allele freq < 20.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

2 To **Top 3 custom** Advanced Filters



Advanced Filtering

○ Recommended ● Custom

Remove genotype in cell with quality < 80

Remove genotype in cell with read depth < 20.00

Remove genotype in cell with alternate allele freq < 40.00

Remove variants genotyped in < 50.00 % of cell

Remove cells with < 50.00 % of genotypes present

Remove variants mutated in < 1.00 % of cells

3 Impacts the **Bottom 3 data distributions**