

## Feature Highlight

### Confidence Filter with *Usual Suspects* Option

The Confidence filter finds variants that meet specified criteria for call quality, read depth, and/or additional upstream filters provided by some variant callers.

Notably, the filter now also lets you avoid false positives from *usual suspects*: whole or partial genes that falsely appear enriched with candidate variants in human sequencing studies, regardless of phenotype. We filter these segments by a statistically rigorous method based on apparent nucleotide diversity in healthy public genomes.

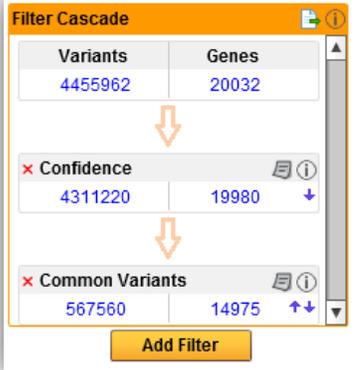
The Confidence filter is applied by default in every analysis, and you can review and adjust settings as needed. Below, we show how to use the Confidence filter to exclude variants that are poorly called, shallowly covered, or in *usual suspects*.

1

Click the  icon to open Confidence filter settings.

2

Review settings and *usual suspect* genes/windows to be excluded. Click Apply to change settings. Variants must satisfy all chosen criteria to pass Confidence filter.



Filter Name	Count
Variants	4455962
Genes	20032
Confidence	4311220
Common Variants	567560

## Glossary

### Call Quality

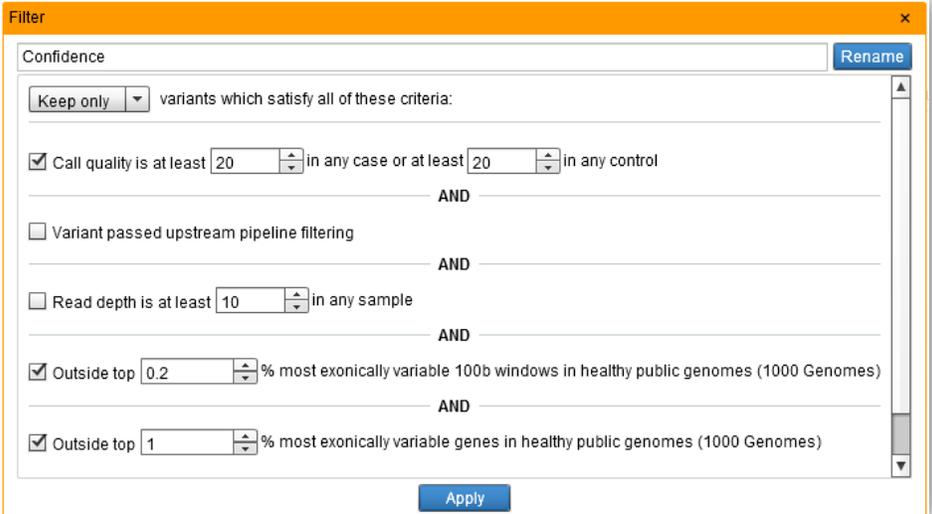
Values are Phred-scaled scores corresponding to the *QUAL* column in a standard *.VCF*, and reflect the probability that, in a given genome, the site carries at least one copy of the reported non-reference variant. The Confidence filter lets you specify minimum passing quality values in case or/and control samples.

### Upstream Pipeline Filtering

Particular variant callers, such as GATK, may further classify calls by their own qualitative filter criteria. In a *.VCF*, a *PASS* value in the *Filter* section means the call passed such criteria. You can filter on this column, to exclude calls that failed such caller-specific criteria.

### Read Depth

Read depth (*DP*) is the number of reads mapped to a given site in a sample or set of samples. Variant Analysis reports per-sample depth (encoded in the *Format* section of a *.VCF*), if available; otherwise it reports total depth (encoded in the *Info* section of a *.VCF*) across all samples. You can filter variants by minimum read depth in a set of analyzed samples.



**Filter**

Confidence Rename

Keep only variants which satisfy all of these criteria:

- Call quality is at least 20 in any case or at least 20 in any control
- AND
- Variant passed upstream pipeline filtering
- AND
- Read depth is at least 10 in any sample
- AND
- Outside top 0.2% most exonically variable 100b windows in healthy public genomes (1000 Genomes)
- AND
- Outside top 1% most exonically variable genes in healthy public genomes (1000 Genomes)

Apply

*Most exonically variable genes/100b windows (usual suspects)*

By default, the Confidence filter excludes variants called in 100b windows that exceed a user-adjustable percentile in the distribution of a measure based on estimated nucleotide diversity (the proportion of sites at which two randomly chosen copies of a chromosome differ) in the exons of generally healthy people published by the 1000 Genomes project. In case-only studies, the default filter also excludes variants called anywhere in whole genes that exceed a user-adjustable percentile in the distribution of an exonic nucleotide diversity-based measure that also reflects gene length (which, in such studies, contributes to noise from usual suspects).

In current human sequencing studies, some whole or partial genes often harbor apparent candidate variants, regardless of studied phenotype(s). Such *usual suspects* include genes that are a) typically poorly sequenced, due to problems with the reference genome; b) dense with real but harmless variation in healthy people; or c) encode very long proteins, so harbor real variants that fit many possible patterns of distribution among people. These variants can add misleading noise to your findings, either directly (as individual false-lead variants called in cases) or indirectly (by skewing case/control counts of such variants in gene or pathway burden tests).

**FAQs**

<p><b>What variant call quality cutoff should I use?</b></p>	<p>By default, we set the minimum Phred-scaled variant call quality at 20 in cases and controls. This typically balances sensitivity and specificity well enough to efficiently find causal variants. But if default settings yield too many/few apparently plausible candidates, try re-filtering with more/less stringent cutoff(s). And in studies of kindreds – where variants called confidently in one genome tend to be harbored in others – consider setting a lower cutoff in controls than cases, to reduce noise from variants that are called confidently in cases and are also evident in controls, but by chance are poorly covered.</p>
<p><b>Should I use the upstream filter criterion?</b></p>	<p>If the genomes that you uploaded into Variant Analysis include qualitative pass/fail values from upstream variant-calling, we recommend filtering on those values, to help exclude spurious calls. If doing so yields too few apparently plausible candidates, then try re-filtering without the upstream filter criterion.</p>
<p><b>Should I filter on read depth?</b></p>	<p>Filtering on read depth is rarely necessary, as variant call qualities already reflect read depth. But read depth can be independently informative in tumors (in which causally relevant aneuploidy and cellular heterogeneity often complicate genotype calling) and other studies when copy number is not well called, but may importantly vary between people or tissues. By setting one or more Confidence filters to <i>keep</i> or <i>exclude</i> variants covered at particular minimum depth, you can flexibly tailor this criterion to such studies.</p>
<p><b>Should I filter out usual suspect genes and/or windows?</b></p>	<p>Filtering out usual suspects typically cuts down on spurious leads, but may occasionally exclude a legitimate causal variant. In every study, the Confidence filter is set by default to exclude variants in the top 0.2% most exonically hypervariable tandem (end-to-end) 100b windows in healthy public genomes. Such filtering is more spatially precise than filtering out whole usual suspect genes, but incurs slight noise from windowing.</p> <p>In studies without controls, the Confidence filter is set by default to also exclude variants in the top 1% most exonically variable genes in healthy public genomes. Such filtering helps cut down on noise from genes that encode extremely long proteins, so may by chance harbor variants that fit many possible patterns of distribution. Note, however, that some such long usual suspect genes, such as <i>TTN</i> and <i>DMD</i>, are actually implicated in some diseases and phenotypes.</p> <p>Thus we recommend reviewing the list of windows and/or genes to be excluded by the Confidence filter, and adjusting settings to avoid filtering any genes deemed strong prior candidates in your study.</p>