
VCF Filter Documentation

Release 1.0

Yichao Shen; Carolyn Caron; Lacey-Anne Sanderson

Apr 21, 2023

Contents:

1	Features	3
1.1	Various Filter Options	3
1.2	Configuration Options	5
1.3	Restrict Access by Permissions	7
2	Installation	9
2.1	Download VCF Filter	9
2.2	Dependencies	9
2.3	Enable VCF Filter	10
3	Configuration	11
3.1	Required information for Adding a file	11
3.2	Optional information for Adding a file	12
3.3	Test before Publication	13

This module provides a form interface so users can custom filter existing VCF files and export in a variety of formats. The form simply provides an interface to VCFtools and uses the Tripal Download API to provide the filtered file to the user.

- **User “Filter VCF” form providing well documented filter options (includes examples) and a variety of formats.**
 - Basic filter options include: Only bi-allelic SNPs, Minimum SNP Call Read Depth, Minor Allele Frequency, Maximum Missing Count, Maximum Missing Frequency.
 - More filter options include: regions and germplasm.
 - Export Formats include: VCF, Quality Matrix (read depth only), A/B Biparental Matrix, Hapmap, Bgzipped VCF.
- All filtering and format conversion is done within a Tripal Job to support large files.
- **Administrative interface for exposing VCF files to users. Extensive configuration options allow comprehensive description**
 - In addition to specifying the path to the VCF file to expose, record helpful information like a friendly name, assembly aligned to, number of SNPs.
 - The information of the methods used in generating each VCF file, a statistic summary and more description can be included.
 - All germplasm names and Chromosome name format can be included as more helpful information.
- Per VCF file permissions allowing you to restrict access to a given file to specific users or roles.

1.1 Various Filter Options

Many filter options are available in this module. Each filter option is well documented with description, example, or even warning as users may not be familiar with all filter options.

1.1.1 Restrict dataset to specific germplasm or regions

- This section will be collapsed if no file is selected.

- Germplasm names from the file are provided to the user, who can then make changes and copy those they want to the textarea below.
- Users can follow the example format provided to keep only sites in one specific region or multiple regions.
- Help information can be configured to improve user experience.

Specify filter criteria.

Restrict dataset to specific germplasm or regions

Select a VCF file from above to see specific germplasm.

Regions


TestChr1:111111..222222

Only include sites in specific regions. For example, if you want to include all genotypic data for a QTL on Chr1 from 5555 to 6666 then you would enter Chr1:5555..6666
Region input format: Chrom:position-lower-bound..position-upper-bound (e.g.,TestChr1:111111..222222).

Germplasm from selected file

1001
1002
1003
1004
1005
1006

All germplasm (individuals) in selected file. **Please copy the germplasm that need to be kept to textarea under.**


You must **copy** the germplasm you want to keep into the **Keep these Germplasm** or your file will contain all germplasm.

Keep these Germplasm

Only include these germplasm (individuals) in export file. Each germplasm name should be on it's own line and must match exactly the names in the file chosen (names are shown above).

1.1.2 Basic Filtering Options

Basic filter options include:

- Bi-allelic
- Read Depth
- Minor Allele Frequency

- Site Missing Count
- Site Missing Frequency

☐ **Only include Bi-allelic SNPs**

If you check this checkbox, only SNPs with 2 alleles across all individuals will be kept. For example, in the example data below, SNP Chr3p34567 would be removed.

Minimum SNP Call Read Depth

Only include SNP calls that have at least the specified number of reads to support the call. For example, if you specify 5 for this filter then for SNP Chr2p25678 in the example table below, only the call for Germplasm4 will be set to missing data.

Minor Allele Frequency

Only include SNP positions with a minor allele frequency greater than or equal to this value. Allele frequency is defined as the number of times an allele appears over all individuals at that site, divided by the total number of non-missing alleles at that site. For example, if you enter 45% in this filter then SNPs with a minor allele frequency lower than 45% could be removed (SNP Chr1p12344 in the example data below).

Maximum Missing Count

Exclude SNPs with more than this number of missing genotypes over all individuals/germplasm. For example, if you enter 1 for this filter then SNPs with more than 1 missing genotype would be removed (SNP Chr4p48765 in the example data below).

Maximum Missing Frequency

Exclude SNPs based on the proportion of missing data. For example, if you enter 25% for this filter then SNPs with a missing data frequency higher than 25% would be removed (SNP Chr4p48765 in the example data below).

Example Table: Example Data for Filter Explanation.

SNP Name	SNP Backbone	SNP Position	Germ1	Germ2	Germ3	Germ4	Germ5	Germ6
Chr1p12344	Chr1	12344	AA:5	TT:12		TT:15	AT:19	TT:15
Chr2p25678	Chr2	25678	GG:7	GG:13	GG:5	TT:2	GG:22	GT:24
Chr3p34567	Chr3	34567	AA:5	CC:12	AC:7	TT:15	CC:19	TC:23
Chr4p48765	Chr4	48765		CC:12	AC:7		CC:19	AA:23

* The above example will be referred to in the description of each filter criteria to aid in the explanation of how it will affect your data. **NOTE: the cell for each SNP by germplasm combination contains the call and the read depth separated by a colon (:).** For example, AA:5 means a call of AA with a read depth of 5.

Note: Filter of VCF files is achieved by using bioinformatic tool [VCFtools](#).

1.2 Configuration Options

As shown in the screenshot below, a particular description is given to a file to help users. It is achieved by the configuration option

- name of the file, assembly it was aligned to and the number of SNPs

- a description which could include a basic introduction, but also details of the file
- a statistic summary could be included to give user a intuitive expression for choosing filter criterias
- chromosome name format can be provided for filter with regions
- germplasm names are provided for filter with specific germplasm

	Name	Assembly	Number of SNPs
<input checked="" type="radio"/>	Test File number 1	Genome2.0	506
<input type="radio"/>	Same Test File number 2	Genome3.0	506

More information on Test File number 1: This is a subset of the GBS diversity panel processed by on researcher in university of S. There are 123 accessions, and it has been filtered for a **minimum mean read depth of 3 for each site across all individuals**. This set is still rough and meant for a "look-see" - further filtering is **strongly recommended!**

Statistical Summary	Min	Max	Median	Average	SD
Read Depth	0	262	4	5.1099	5.6614
Minor Allele Frequency	0	0.5	0.003	0.0728	0.1277
Missing Frequency	0.002	0.823	0.128	0.1476	0.0878

Specify filter criteria.

▼ Restrict dataset to specific germplasm or regions

Select a VCF file from above to see specific germplasm.

Regions

TestChr1:111111..222222

Only include sites in specific regions. For example, if you want to include all genotypic data for a QTL on Chr1 from 5555 to 6666 then you would enter Chr1:5555..6666

Region input format: Chrom:position-lower-bound..position-upper-bound (e.g., TestChr1:111111..222222).

Germplasm from selected file

1001
1002
1003
1004
1005
1006

All germplasm (individuals) in selected file. **Please copy the germplasm that need to be kept to textarea under.**

1.3 Restrict Access by Permissions

Per file access can be managed in Home » Administration » Tripal » Extensions » VCF Filter.

Role Permissions: All users with the following roles have access to *Test File number 1*.

NAME	OPERATIONS
administrator	remove

Role to give permission to

- None -

Add Role

User Permissions: The following users have access to *Test File number 1*.

NAME	OPERATIONS
None.	

User to give permission to

Add User

Note: It is recommended to clear caches regularly in this installation processes.

2.1 Download VCF Filter

The module is available as one repository for [Pulse Bioinformatics, University of Saskatchewan](#) on GitHub. Recommended method of downloading and installation is using git:

```
cd [your drupal root]/sites/all/modules  
git clone https://github.com/UofS-Pulse-Binfo/vcf_filter.git
```

2.2 Dependencies

Required dependencies for VCF Filter

- Tripal Core (utilizes the Tripal API)
- Tripal Download API

We can check status of modules in “Home » Administration » Tripal » Modules”.

TRIPAL EXTENSIONS				
ENABLED	NAME	VERSION	DESCRIPTION	OPERATIONS
<input checked="" type="checkbox"/>	Breeding API		Breeding API is an extension module for Tripal that provides REST API for plant breeding. Requires: Breeding API site entity (enabled), Entity API (enabled)	Help Permissions Configure
<input checked="" type="checkbox"/>	Breeding API site entity		Provides a Breeding API site entity type. Requires: Entity API (enabled) Required by: Breeding API (enabled)	Permissions
<input checked="" type="checkbox"/>	Tripal Jobs Daemon		Creates a Daemon to run Tripal Jobs as they are submitted. Requires: Drush Daemon API (enabled), Libraries (enabled), System (enabled), Tripal (enabled), Views (enabled), Chaos tools (enabled), Path (enabled), Search (enabled), PHP filter (enabled), Entity API (enabled), Redirect (enabled)	
<input type="checkbox"/>	VCF Filter		Form interface so users can custom filter existing VCF files and export in a variety of formats. Requires: Trpdownload_api (missing)	

In this example, it is clear that Trpdownload_api is required but not available in system. Trpdownload_api is available on GitHub, and can be installed with following commands:

```
cd [your drupal root]/sites/all/modules
git clone https://github.com/tripal/trpdownload_api.git
drush pm-enable trpdownload_api
```

Note: VCFtools is required for VCF Filter.

2.3 Enable VCF Filter

After all dependencies are installed and enabled, VCF Filter can be enabled to use in “Home » Administration » Tripal » Modules” of your site.

Also, VCF Filter can be enabled by drush command:

```
drush pm-enable vcf_filter
```

This command will enable the module after which we should able to find it in Home » Administration » Tripal » Extensions.

Home » Administration » Tripal » Extensions			
VCF Filter o			
This module provides a form so users can custom filter existing VCF files and export in a variety of formats. In order to make a VCF file available to users, it must first be included below:			
Add			
HUMAN-READABLE NAME	BACKBONE	FILE	OPERATIONS
None.			


The module can be configured in Home » Administration » Tripal » Extensions » VCF Filter by edit a file.

3.1 Required information for Adding a file

Only site admins can configure VCF Filter in Home » Administration » Tripal » Extensions » VCF Filter. The following information is required:

- Absolute path of the file
- Human-readable Name
- Number of SNPs (sites) of the file
- Backbone

[Home](#) » [Administration](#) » [Tripal](#) » [Extensions](#) » [VCF Filter](#)

Add VCF File 

VCF File (absolute path)

The absolute path to your VCF file. This file must already exist.

Human-readable Name

This is the name shown to your users so make sure it is descriptive and uniquely identifies the VCF file

Number of SNPs

The number of SNPs in the file.

Backbone (e.g. Assembly)

The name of the sequence assembly the SNPs were called on.

3.2 Optional information for Adding a file

The module can work without optional configuration, but it is highly recommended to provide it for better user experience. Instructions are provided for each configuration option.

The following screenshot is an example:

Description

This is a subset of the GBS diversity panel processed by on researcher in university of S. There are 123 accessions, and it has been filtered for a **minimum mean read depth of 3 for each site across all individuals**. This set is still rough and meant for a "look-see" - further filtering is **strongly recommended**

Statistical Summary
Min-Max-Median-Average-SD-Read Depth
0-262-4-5.1099-5.6614
Minor Allele Frequency
0-0.5-0.003-0.0728-0.1277
Missing Frequency
0.002-0.823-0.128
0.1476-0.0878

This should include the method used to generate the file and any filtering that has already been done. It may also be helpful to include some stats about the file such as average read depth to give users some context when filtering.

Germplasm From Header

1001
1002
1003
1004
1005
1006

This should include all names of germplasm (individuals) from the header of VCF file. If this textarea is not filled, shell command will be used to find the header and extract all germplasm names but it may slow down this module a lot.

Chromosome format

TestChr1

This should indicate how chromosome names are formatted in this VCF file. For example, Chr1, chr1, Chr01 or name with a prefix like Lc, Ca, Gm, Pv or Mt.

3.2.1 Description

What we could include in description:

- Background information about project/experiment and researchers/institution could help for better understanding of the file
- Bioinformatic tools and correlated parameters that have been applied in generating the VCF file
- Number of germplasm (individuals) included in the file, and names for maternal parent and paternal parent
- A filter criteria related statistic summary (the summary in example can be generated by a [PHP script](#))

3.2.2 Germplasm From Header

The names of all germplasm (individuals) in this vcf file. The germplasm list must be new line separated without any header or empty lines.

Note: If this textarea is not filled, the module is able to find the list from selected VCF fiels. However, waiting time of extracting germplasm list from a selected file can be significant for large VCF files. Loading time for a 10G VCF file will be about 3 seconds.

Since the germplasm list can be generated, it's not necessary to generate such a list for configuration otherwise. We can leave this section blank, select this file and copy generated list back to configuration.

3.2.3 Chromosome format

- Chromosome name can have various format, for example, chromosome 1 for one lentil cultivar could be chr1, Chr1, CHR1, LcChr1, Lcchr, and so on. Therefore, it is important to provide this information so users can filter vcf file by regions properly.

3.3 Test before Publication

An comprehensive test of your configuration is recommended before making this module public to users. Some good things to check

- test if all files added are downloadable
- test if download files have proper contents
- test if accesses are given to proper groups and/or individuals

Note: It is recommended to give permissions to site admins for testing before release.

Note: We appreciate if you can report issues found while using this module. You can reach us at knowpulse@usask.ca or report the issue on [GitHub](#). It will be more appreciated if you can include screenshots and an informative description of the issue.

Thank you for using VCF Filter!

Have a wonderful day!

After configuration, description of one file can be very informative and helpful for filtering options.

	Name	Assembly	Number of SNPs
<input checked="" type="radio"/>	Test File number 1	Genome2.0	506
<input type="radio"/>	Same Test File number 2	Genome3.0	506

More information on Test File number 1: This is a subset of the GBS diversity panel processed by on researcher in university of S. There are 123 accessions, and it has been filtered for a **minimum mean read depth of 3 for each site across all individuals**. This set is still rough and meant for a "look-see" - further filtering is **strongly recommended!**

Statistical Summary	Min	Max	Median	Average	SD
Read Depth	0	262	4	5.1099	5.6614
Minor Allele Frequency	0	0.5	0.003	0.0728	0.1277
Missing Frequency	0.002	0.823	0.128	0.1476	0.0878

Specify filter criteria.

▼ Restrict dataset to specific germplasm or regions

Select a VCF file from above to see specific germplasm.

Regions

TestChr1:111111..222222

Only include sites in specific regions. For example, if you want to include all genotypic data for a QTL on Chr1 from 5555 to 6666 then you would enter Chr1:5555..6666

Region input format: Chrom:position-lower-bound..position-upper-bound (e.g., TestChr1:111111..222222).

Germplasm from selected file

1001
1002
1003
1004
1005
1006

All germplasm (individuals) in selected file. Please copy the germplasm that need to be kept to textarea under.