Annotate and Filter Variants

Author: Greta Peterson, Andrew Jesaitis, Golden Helix, Inc.

Overview

Annotate and filter variants based on information found in Variant, Interval, and Gene sources.

Filters can be chained together and spreadsheets will be output at each level of the filter chain.

Recommended Directory Location

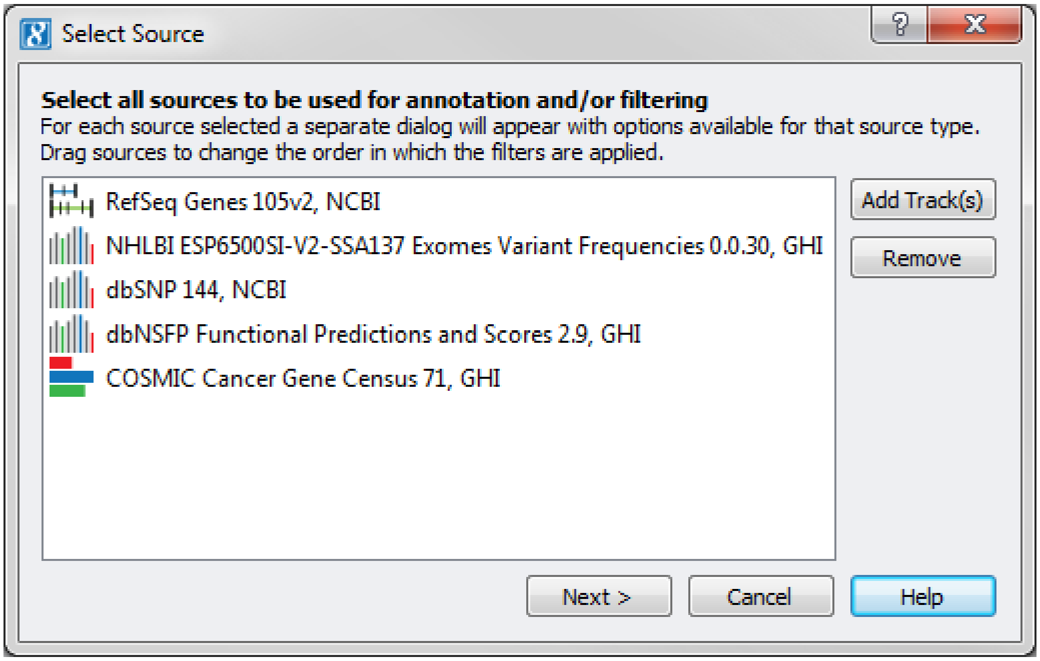
Save the script to the following directory:

*..**\Application Data\Golden Helix SVS\UserScripts\Spreadsheet\DNA_Seq\**

Note: The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between XP and Vista. The easiest way to locate this directory on your computer is to open SVS and go to **Tools > Open Folder > User Scripts Folder**. If saved to the proper folder, this script will be accessible from the spreadsheet's **DNA-Seq**.

Using the Script

From a marker mapped spreadsheet go to **DNA-Seq > Annotate and Filter Variants**. Click the **Add Track(s)** button to select all data sources to be used for annotating and/or filtering variants.

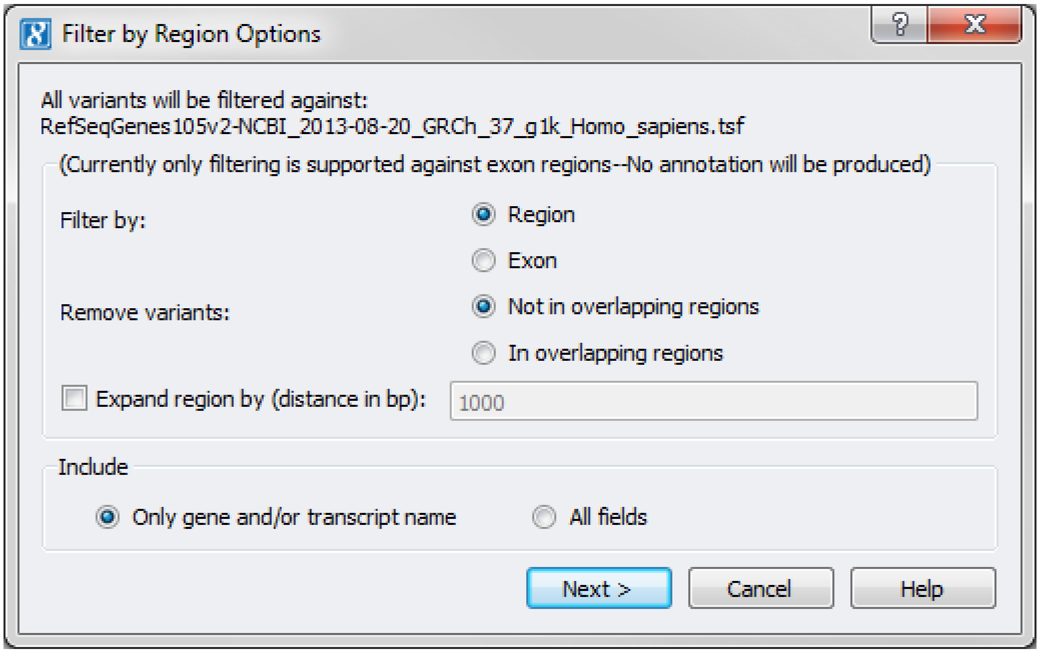


Select all sources to be used for annotation and/or filtering
For each source selected a separate dialog will appear with options available for that source type.
Drag sources to change the order in which the filters are applied.

- RefSeq Genes 105v2, NCBI
- NHLBI ESP6500SI-V2-SSA137 Exomes Variant Frequencies 0.0.30, GHI
- dbSNP 144, NCBI
- dbNSFP Functional Predictions and Scores 2.9, GHI
- COSMIC Cancer Gene Census 71, GHI

Gene Region Membership

When a gene source is selected this tool allows for filtering and annotating by the defined gene regions or for filtering by exon regions. If an interval source is selected this tool will annotate and filter by the defined interval regions in the source.



All variants will be filtered against:
RefSeqGenes105v2-NCBI_2013-08-20_GRCh_37_g1k_Homo_sapiens.tsf

(Currently only filtering is supported against exon regions--No annotation will be produced)

Filter by: ● Region ○ Exon

Remove variants: ● Not in overlapping regions ○ In overlapping regions

□ Expand region by (distance in bp): 1000

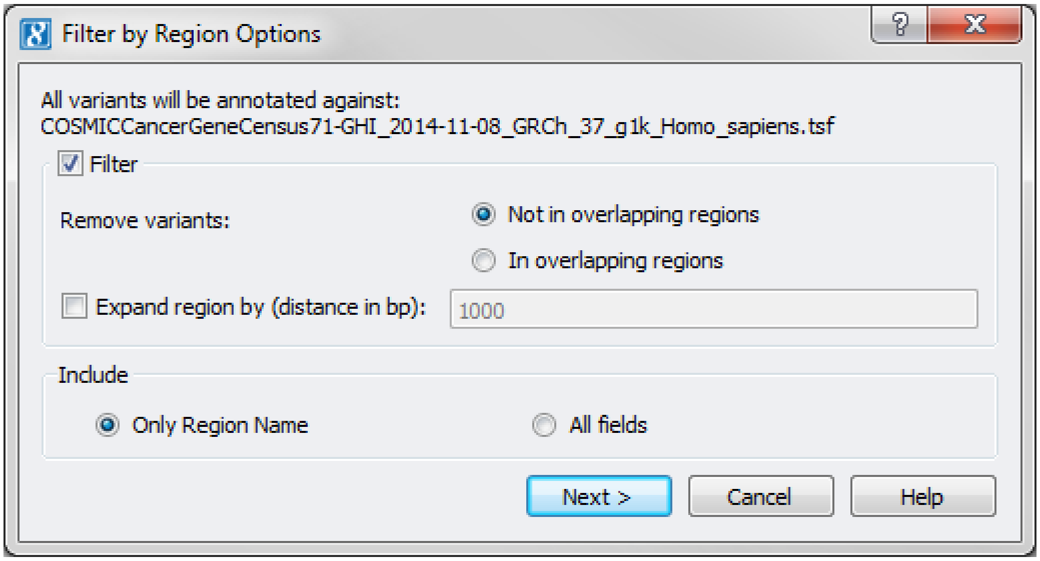
Include

● Only gene and/or transcript name ○ All fields

- Options to specify:
  - Select your filter options then indicate if markers inside or outside the source regions should be removed. Checking the "Expand region by (distance in bp):" option will treat markers that are within a specified distance from region boundaries as part of the region for filtering purpose.
  - When Region filtering is selected you have the option of only including the gene and/or transcript name fields or all fields in the source.
  - If the spreadsheet has a considerable number of variants selecting the **All fields** option can substantially increase the time it takes the tool to finish.
- Results:
  - For annotation results an output spreadsheet is produced with a row for each marker. The first column of the spreadsheet is a binary column to indicate whether the marker was found in the selected source, followed by columns that contain any additional information included in the source for that marker.
  - If filtering is selected, a subset spreadsheet will be created containing those markers left activate in the original spreadsheet based on filter criteria.
  - Currently only filtering is supported against exon regions, no annotation report will be produced.

Interval Region Membership

When an interval source is selected this tool allows for filtering and annotating by the defined interval regions in the source.



Filter by Region Options

All variants will be annotated against:
COSMICCancerGeneCensus71-GHI_2014-11-08_GRCh_37_g1k_Homo_sapiens.tsf

☑ Filter

Remove variants: ● Not in overlapping regions ○ In overlapping regions

□ Expand region by (distance in bp): 1000

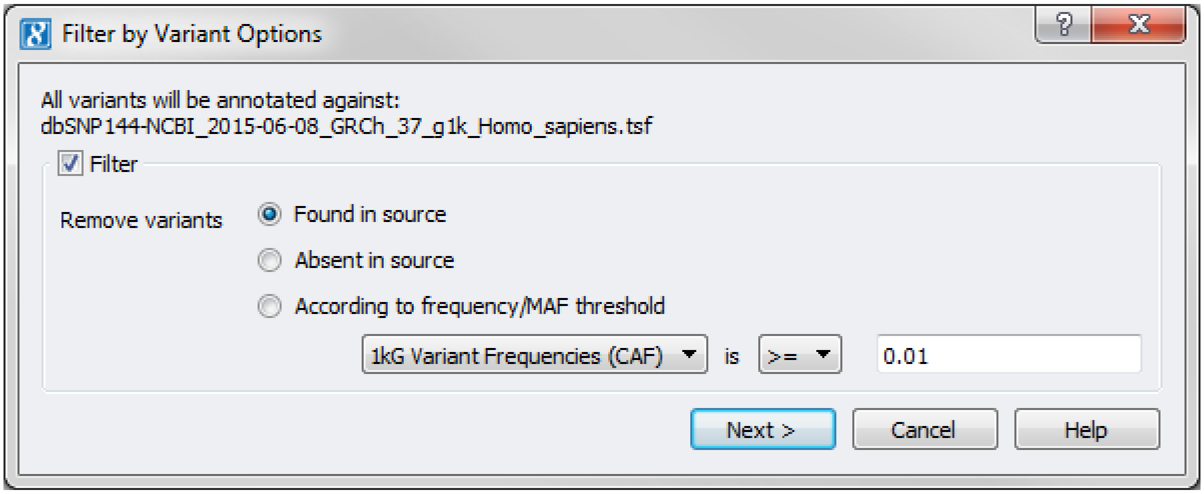
Include

● Only Region Name ○ All fields

- Options to specify:
  - Select your filter options then indicate if markers inside or outside the source regions should be removed. Checking the "Expand region by (distance in bp):" option will treat markers that are within a specified distance from region boundaries as part of the region for filtering purposes.
  - For annotation and filtering results you have the option to include only the region name in the report to include all fields from the selected source.
- Results:
  - For annotation results an output spreadsheet is produced with a row for each marker. The first column of the spreadsheet is a binary column to indicate whether the marker was found in the selected source, followed by columns that contain any additional information included in the source for that marker.
  - If filtering is selected, a subset spreadsheet will be created containing those markers left activate in the original spreadsheet based on filter criteria.

Variant Source Membership

This tool identifies markers in the spreadsheet that are also in a specified variant source. These markers can also be filtered based on presence or absence in the specified source using this feature.



Filter by Variant Options

All variants will be annotated against:
dbSNP144-NCBI_2015-06-08_GRCh_37_g1k_Homo_sapiens.tsf

☑ Filter

Remove variants
● Found in source
○ Absent in source
○ According to frequency/MAF threshold

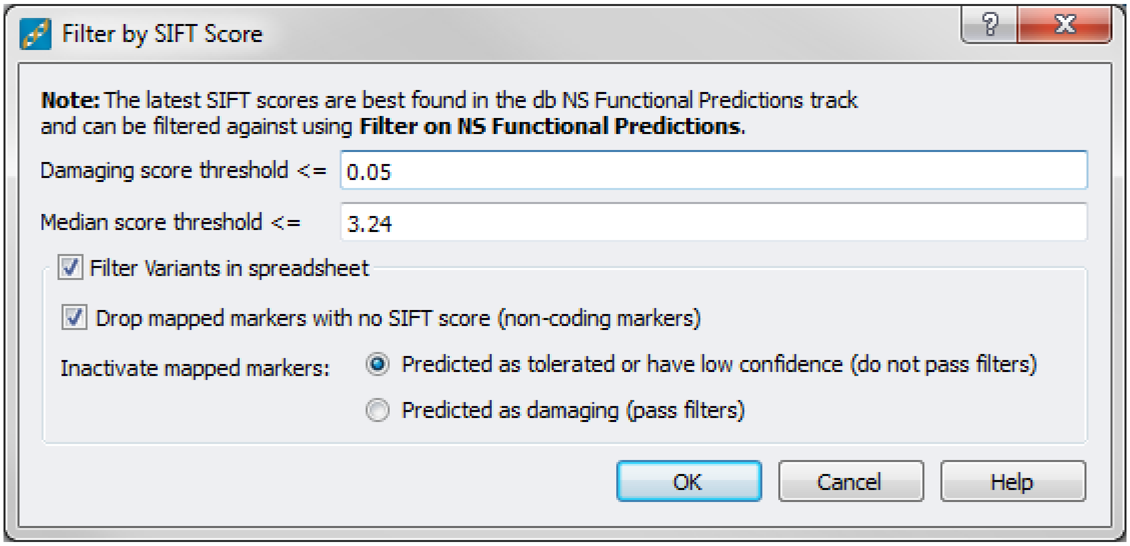
1kG Variant Frequencies (CAF) is >= 0.01

- Options to specify:
  - Select whether to filter variants by selected source and select which variants to remove based on presence or absence in source.
- Results:
  - For annotation results an output spreadsheet is produced with a row for each marker. The first column of the spreadsheet is a binary column to indicate whether the marker was found in the selected source, followed by columns that contain any additional information included in the source for that marker.
  - If filtering is selected, a subset spreadsheet will be created containing those markers left activate in the original spreadsheet based on filter criteria. If filtering is selected, markers will be inactivated according to the criteria specified and a subset spreadsheet of active variants will be produced.

Filter by SIFT Score

Note: The latest SIFT scores are best found in the **dbNSFP Functional Predictions and Scores** source and can be filtered against by selecting that source in the source dialog, shipped in SVS version 7.6.7+. SIFT filtering can now be applied simultaneously with other filtering tools such as PolyPhen2 and MutationTaster.

This function creates a report with several statistics per marker. Optionally the user can also inactivate mapped markers that are either predicted as tolerated or have low confidence (do not pass filters) or are predicted as damaging (pass filters), depending on the inactivation option selected.



Note: The latest SIFT scores are best found in the db NS Functional Predictions track and can be filtered against using **Filter on NS Functional Predictions**.

Damaging score threshold <= 0.05

Median score threshold <= 3.24

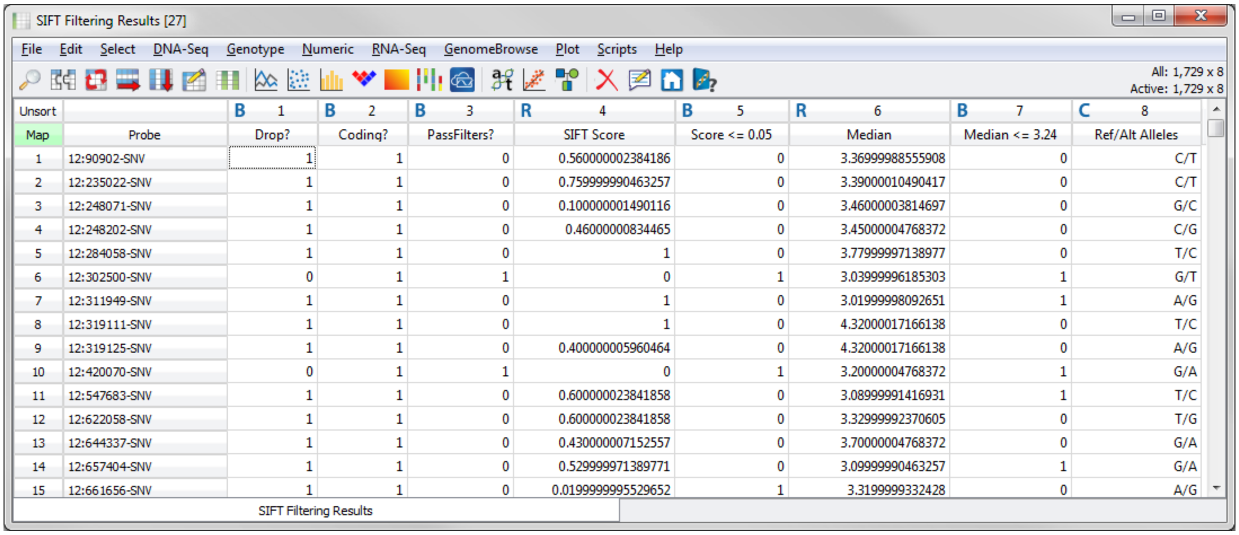
☑ Filter Variants in spreadsheet

☑ Drop mapped markers with no SIFT score (non-coding markers)

Inactivate mapped markers: ● Predicted as tolerated or have low confidence (do not pass filters)

○ Predicted as damaging (pass filters)

- Options to specify:
  - Enter a damaging score threshold. The SIFT score ranges from 0 to 1. If the threshold is set to 0.05, then the amino acid substitution is predicted as damaging if the score is less than or equal to 0.05 and predicted as tolerated if the score is greater than 0.05.
  - Enter a median score threshold. The SIFT median info value ranges from 0 to 4.32. Ideally the median should be between 2.75 and 3.5. This is used to measure the diversity of the sequences used for prediction. Median score values greater than or equal to 3.25 should be considered low quality predictions, as this indicates that the prediction was based on closely related sequences.
  - Optionally choose to filter the variants in the original spreadsheet.
  - Specify whether mapped markers with no SIFT score (non-coding markers) should be inactivated or not.
  - If no SIFT score is available at the loci of a given marker, then it is not in a coding region (exon). These markers can be inactivated as well or left as active.
  - Indicate which markers should be inactivated, either:
    - Those predicted as tolerated or having low confidence (do not pass filters), or
    - those predicted as damaging (pass filters).
- Results:
  - If selected, in the original spreadsheet, the columns that meet the inactivation criteria are inactivated. The other columns are left active and subset spreadsheet is created from those active columns.
  - A marker mapped filtering results spreadsheet is also created as a child of the original spreadsheet. This spreadsheet contains the following columns.



Unsort		B 1	B 2	B 3	R 4	B 5	R 6	B 7	C 8
Map	Probe	Drop?	Coding?	PassFilters?	SIFT Score	Score <= 0.05	Median	Median <= 3.24	Ref/Alt Alleles
1	12:90902-SNV	1	1	0	0.560000002384186	0	3.36999988555908	0	C/T
2	12:235022-SNV	1	1	0	0.759999990463257	0	3.39000010490417	0	C/T
3	12:248071-SNV	1	1	0	0.100000001490116	0	3.46000003814697	0	G/C
4	12:248202-SNV	1	1	0	0.46000000834465	0	3.45000004768372	0	C/G
5	12:284058-SNV	1	1	0	1	0	3.77999997138977	0	T/C
6	12:302500-SNV	0	1	1	0	1	3.03999996185303	1	G/T
7	12:311949-SNV	1	1	0	1	0	3.01999998092651	1	A/G
8	12:319111-SNV	1	1	0	1	0	4.32000017166138	0	T/C
9	12:319125-SNV	1	1	0	0.400000005960464	0	4.32000017166138	0	A/G
10	12:420070-SNV	0	1	1	0	1	3.20000004768372	1	G/A
11	12:547683-SNV	1	1	0	0.600000023841858	0	3.08999991416931	1	T/C
12	12:622058-SNV	1	1	0	0.600000023841858	0	3.32999992370605	0	T/G
13	12:644337-SNV	1	1	0	0.430000007152557	0	3.70000004768372	0	G/A
14	12:657404-SNV	1	1	0	0.529999971389771	0	3.09999990463257	1	G/A
15	12:661656-SNV	1	1	0	0.0199999995529652	1	3.3199999332428	0	A/G

- SIFT Filtering Results
  - Drop? - Indicates whether the marker is dropped based on the specified filtering criteria. Only included if filtering is selected.
  - Coding? - Indicates whether the marker is in a coding region or not.
  - PassFilters? - Indicates whether the marker passes the SIFT filters or not (i.e. damaging or tolerated/have low confidence).
  - SIFT Score - The SIFT score for the marker.
  - Score <= # - Indicates whether the SIFT score is below the specified damaging score threshold for each marker.
  - Median - Median score value for the marker.
  - Median <= # - Indicates whether the median score is below the specified median score threshold.
  - Ref/Alt Alleles - The reference and alternate nucleotide bases used for the calculations of the SIFT and median scores.

Filter by SIFT Synonymous Classification

If needed to reproduce existing workflows, this script can be obtained from the SVS Add-On Scripts repository. http://www.goldenhelix.com/SNP_Variation/scripts/index.html

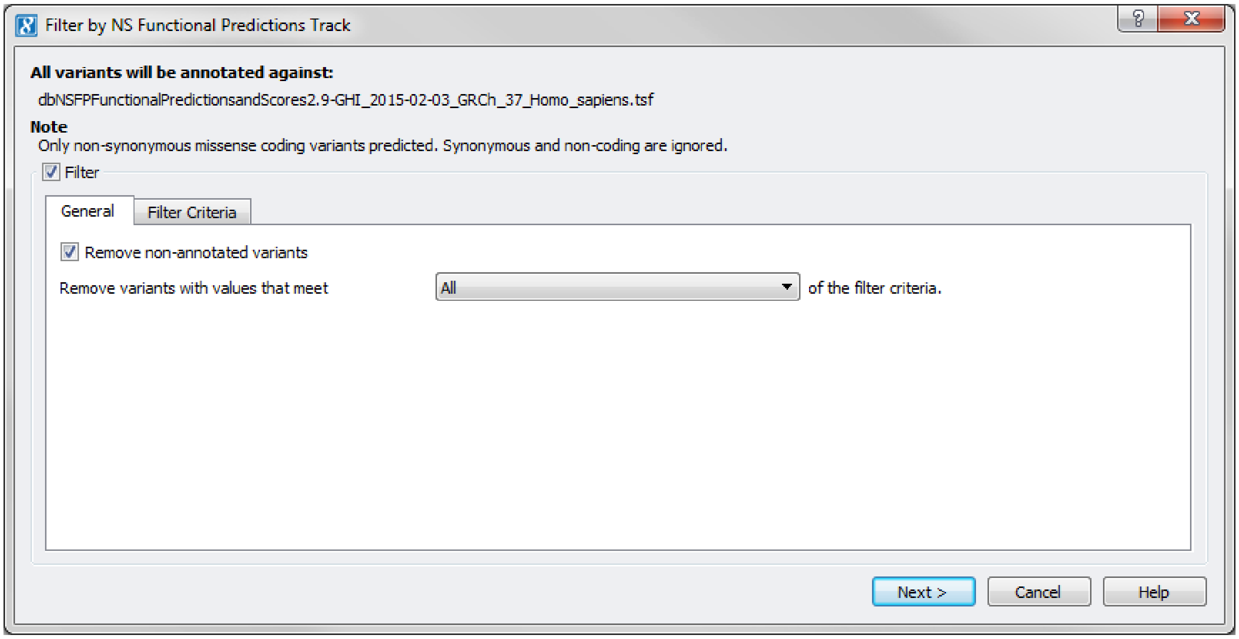
Filter by PolyPhen2 Score

If needed to reproduce existing workflows, this script can be obtained from the SVS Add-On Scripts repository. http://www.goldenhelix.com/SNP_Variation/scripts/index.html

Filter Variants on Non-Synonymous Functional Predictions

This tool allows the user to filter against functional prediction information available in either the full dbNSFP Functional Predictions and Scores or the subset dbNSFP Functional Predictions annotation sources. These sources are available for the three human builds, NCBI_36 ,GRCh_37, and GRCh_38.

This tool requires the most recent dbNSFP annotation source be used for annotation and filtering.



Filter by NS Functional Predictions Track

All variants will be annotated against:
dbNSFPFunctionalPredictionsandScores2.9-GHI_2015-02-03_GRCh_37_Homo_sapiens.tsf

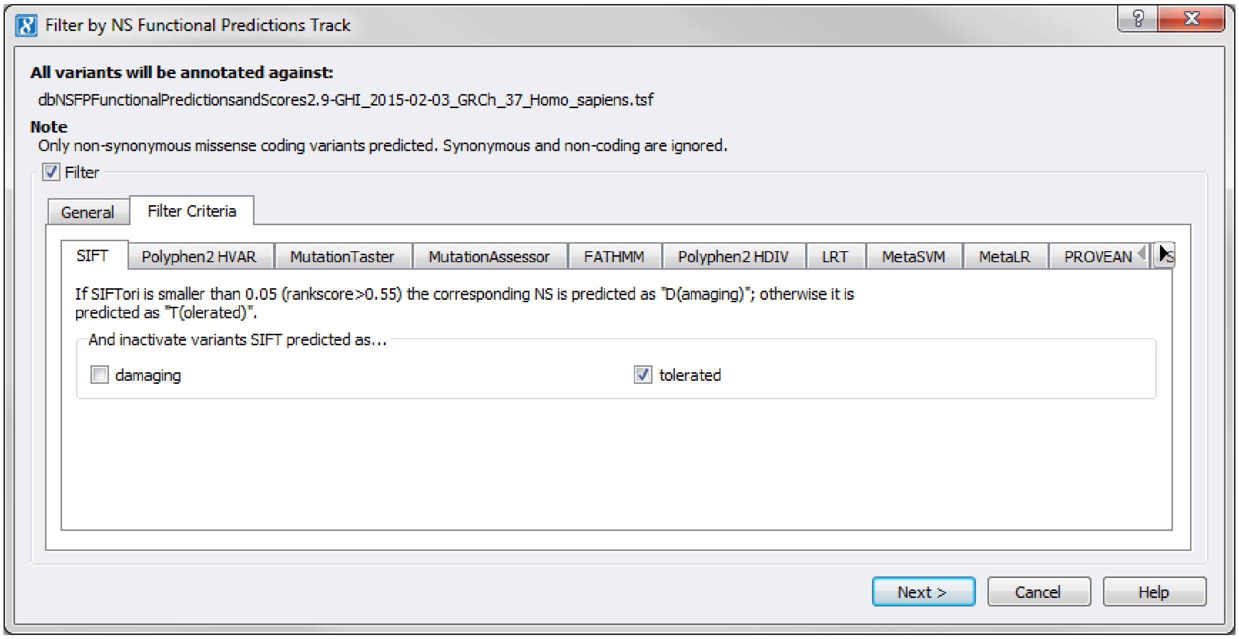
Note
Only non-synonymous missense coding variants predicted. Synonymous and non-coding are ignored.

☑ Filter

☑ Remove non-annotated variants

Remove variants with values that meet All ▾ of the filter criteria.

With this tool, all variants present in the selected source will be annotated. If the filtering option is selected, the user can choose a number of characteristics to filter against and may also choose to inactivate non-annotated variants, or those not found in the dbNSFP Functional Predictions source.



Filter by NS Functional Predictions Track

All variants will be annotated against:
dbNSFPFunctionalPredictionsandScores2.9-GHI_2015-02-03_GRCh_37_Homo_sapiens.tsf

Note
Only non-synonymous missense coding variants predicted. Synonymous and non-coding are ignored.

☑ Filter

If SIFTori is smaller than 0.05 (rankscore>0.55) the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)".

And inactivate variants SIFT predicted as...

□ damaging ☑ tolerated

By default, the variants will be filtered on SIFT, PolyPhen2 HVAR, MutationTaster, MutationAssessor and FATHMM. In addition, other predictions or conservation scores provided in the latest dbNSFP source can be used for filtering.

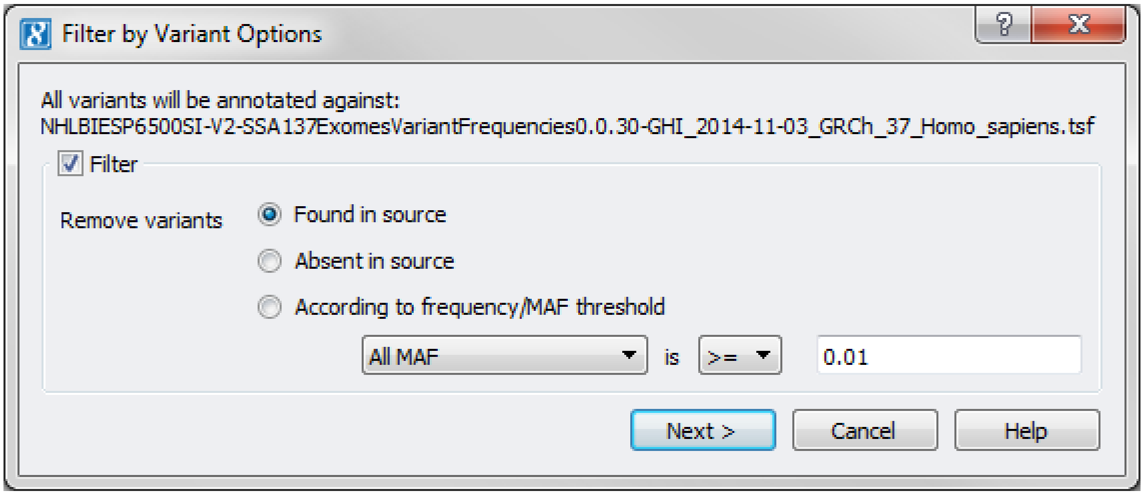
Information on each score or prediction is included on each tab of the filtering dialog.

The default values reflect a filtering scheme that inactivates variants that are predicted or known to have little or no affect in causing disease. These variants are deemed uninteresting and therefore filtered out of the data.

Note: This tool has an implicit filtering requirement that the alleles in the spreadsheet must match the alleles listed in the source.

Filter on Variant Frequency Catalog

This tool allows the user to filter a spreadsheet against an allele frequency field found in an annotation source. For example, you can use this tool to filter out variants with an minor (or alternate) allele frequency greater than a desired threshold. This tool allows the user to simply annotate the variants or annotate the variants and apply the filtering mechanism.



Filter by Variant Options

All variants will be annotated against:
NHLBIESP6500SI-V2-SSA137ExomesVariantFrequencies0.0.30-GHI_2014-11-03_GRCh_37_Homo_sapiens.tsf

☑ Filter

Remove variants
● Found in source
○ Absent in source
○ According to frequency/MAF threshold

All MAF is >= 0.01

If the filtering option is selected, the user may choose to filter variants by presence or absence in the source or by an appropriate allele frequency field (if more than one are detected), a comparison operator (>= or >) and a threshold value must be selected for this option.

If the minor or alternate allele frequency is greater than 0.5 for a particular variant in the selected variant source, then we assume that the major/minor or ref/alt alleles were not correctly specified. The alleles and allele frequencies are flipped so that the filtering is performed using the true minor or alternate allele and its corresponding frequency.

If the alleles present in the spreadsheet contains only the major or reference allele, then it will always be filtered out regardless of the minor/alternate allele frequency as there are no rare alleles present for the variant in the spreadsheet.

On the other hand, if there are more than 3 observed alleles for a variant in the annotation source it is impossible to determine which is the minor allele and as a result a conservative approach is taken and the variant is not filtered out.

Note: Some annotation sources have more than one numeric field which will result in several options in the combo box. For example, the 1kG Phase3 – Variant Frequencies source (only available for the GRCh_37 human build) has a separate field for each sub-population's alternate allele frequency.

Filter on Variant Scores

This tool allows the user to filter a spreadsheet against a prediction score or other real valued field found in an annotation source. For example, you can use this tool to filter out variants with a p-value of less than 0.05 or a prediction score of greater than 0.5. This tool allows the user to simply annotate the variants or annotate the variants and apply the filtering mechanism.