



VcfView v0.5.0.5 beta

Documentation

Created at



Cambridge University

sam.haldenby at easih.ac.uk

General Overview

VcfView allows researchers to view and manipulate data generated from human exome sequencing projects. It takes Variant Call Format (VCF / .vcf) files as input and displays the contained data in tabulated format. The researcher can filter variants based on any data field present in the file, and the software also provides tools for comparison of multiple exome project data, which can be especially useful for pedigree analysis projects.

Motivation

The Variant Call Format is highly extensible and allows the storage of a wide range of information on variants discovered in exome sequencing projects. However, the trade-off for this level of data storage is readability. While the main portion of a VCF file is tabulated and therefore amenable to text parsing, any additional data stored on a variant is represented by a large single text field (the INFO field), which can be more tricky to parse, especially for researchers without much bioinformatics/scripting knowledge.

VcfView was created to allow researchers to very easily view and manipulate their results through a point-and-click interface, obviating the need for scripting and basic bioinformatics assistance.

Installation

VcfView is written in Java (Version 6.0) and will therefore run on any Windows, Linux or Mac machine with Java installed. For all platforms, extract the contents of the downloaded archive file to the directory of your choice and follow the instructions below.

Windows: Double-click the file named VcfView.<VERSION>.jar.

Mac: Double-click the file named VcfView.<VERSION>.jar. Any machine running OS X version 10.5 or later should have Java 6.0 installed by default. Earlier versions may require manual installation.

Linux: Open a console and type `java -jar VcfView<VERSION>.jar`

Contact

This is a beta release and there are therefore likely to be a number of bugs present in the software. We would be very grateful if you could report bugs to us, including the steps taken to cause the bug. Also, if you have any questions, problems or feature suggestions please contact **sam.haldenby at easih.ac.uk** and we will do our best to address them.

Interface

The screenshot shows the VcfView main interface. It features a menu bar at the top with 'File', 'Analysis', and 'Extras'. Below the menu bar is a large table displaying variant data. The table has columns for Sample, #CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, AC, AF, AN, DB, DP, DS, and Dels. The table is currently displaying 94621 variants. On the left side, there is a filter panel with sections for Sample, #CHROM, POS, ID, QUAL, and FILTER. The bottom of the interface shows a status bar with 'Displaying Entries: 94621 / 94621 (100.00%)' and a progress bar.

Sample	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	AC	AF	AN	DB	DP	DS	Dels
A060001...1	1	91176	rs10399...	C	A	52.26	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2		3		0.00
A060001...1	1	91214	.	T	A	163.16	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2		9		0.00
A060001...1	1	91256	.	A	G	80.27	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2		9		0.00
A060001...1	1	91268	rs4245762	G	A	106.60	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	9		0.00
A060001...1	1	136652	.	A	G	37.26	HARD T...	AC=1;A...	GT:AD:D...1	0.50	2	2		21		0.00
A060001...1	1	136954	.	A	G	36.27	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2		33		0.00
A060001...1	1	136995	.	C	G	35.27	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2		32		0.00
A060001...1	1	137825	.	G	A	82.27	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2		12		0.00
A060001...1	1	234308	rs7195389	A	G	230.31	HARD T...	AC=1;A...	GT:AD:D...1	0.50	2	2	true	59		0.00
A060001...1	1	234378	rs62053...	A	G	121.51	HARD T...	AC=1;A...	GT:AD:D...1	0.50	2	2	true	67		0.00
A060001...1	1	234481	rs8179403	T	A	116.71	HARD T...	AC=1;A...	GT:AD:D...1	0.50	2	2	true	29		0.00
A060001...1	1	663097	rs61769...	G	C	111.60	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	17		0.00
A060001...1	1	679604	rs61865...	C	A	133.63	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	31		0.00
A060001...1	1	691544	.	T	A	2134.70	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2		90		0.01
A060001...1	1	701946	rs61769...	A	G	102.60	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	4		0.00
A060001...1	1	761732	rs55962...	C	T	70.25	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	6		0.00
A060001...1	1	762273	rs3115849	G	A	47.26	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	3		0.00
A060001...1	1	762589	rs71507...	G	C	54.26	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	3		0.00
A060001...1	1	762592	rs71507...	C	G	58.26	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	3		0.00
A060001...1	1	762601	rs71507...	T	C	58.26	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	3		0.00
A060001...1	1	762632	rs61768...	T	A	94.27	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	4		0.00
A060001...1	1	783304	rs2980295	T	C	69.27	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	3		0.00
A060001...1	1	787262	rs56108...	C	G	91.27	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	5		0.00
A060001...1	1	787399	rs2905055	G	T	568.80	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	15		0.00
A060001...1	1	788840	rs71507...	T	C	153.71	HARD T...	AC=1;A...	GT:AD:D...1	0.50	2	2	true	49		0.00
A060001...1	1	788844	rs71507...	C	G	150.61	HARD T...	AC=1;A...	GT:AD:D...1	0.50	2	2	true	50		0.00
A060001...1	1	789141	.	G	C	178.31	HARD T...	AC=1;A...	GT:AD:D...1	0.50	2	2	true	52		0.00
A060001...1	1	789256	rs3131939	T	C	1489.80	HARD T...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	45		0.00
A060001...1	1	804115	rs9725068	G	A	41.26	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	10		0.00
A060001...1	1	808631	rs11240...	G	A	125.61	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	4		0.00
A060001...1	1	808922	rs6504027	G	A	69.27	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	5		0.00
A060001...1	1	808928	rs11240...	C	T	108.60	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	6		0.00
A060001...1	1	809639	.	G	T	34.51	shallow20	AC=1;A...	GT:AD:D...1	0.50	2	2	true	7		0.00
A060001...1	1	851777	rs13303...	A	G	35.27	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	2		0.00
A060001...1	1	871434	rs4072383	G	T	55.26	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	2		0.00
A060001...1	1	876499	rs4372192	A	G	51.26	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	2		0.00
A060001...1	1	877715	rs6605066	C	G	98.27	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	4		0.00
A060001...1	1	879170	.	C	A	43.51	StdDf...	AC=1;A...	GT:AD:D...1	0.50	2	2	true	3		0.00
A060001...1	1	879676	rs6605067	G	A	259.46	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	7		0.00
A060001...1	1	879687	rs2839	T	C	227.73	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	6		0.00
A060001...1	1	881627	rs2272757	G	A	348.91	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	9		0.00
A060001...1	1	883625	rs4970378	A	G	189.16	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	6		0.00
A060001...1	1	887560	rs3748595	A	C	130.61	StdDf...	AC=2;A...	GT:AD:D...2	1.00	2	2	true	4		0.00
A060001...1	1	887801	rs3828047	A	G	186.16	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	5		0.00
A060001...1	1	888639	rs3748596	T	C	204.73	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	6		0.00
A060001...1	1	888650	rs3748597	T	C	163.16	shallow20	AC=2;A...	GT:AD:D...2	1.00	2	2	true	5		0.00

The screen-shot above shows the main interface for VcfView. It is composed of four main sections:

- (1) Menu bar: Files are loaded and saved, and analyses can be carried out from options in these menus.
- (2) Table view: One a VCF file has been loaded, information extracted for variants is displayed here in table format.
- (3) Filter panel: Prior to loading a VCF file, the user is given the option to apply filters to the information fields found in the file. Any selected filters are displayed in the panel.
- (4) Information bar: Contains miscellaneous information, i.e. the number of variants currently loaded vs the number of variants displayed after filtering, a bar which tracks the progress of tasks being carried out (e.g. file loading, saving, etc.), and a meter which tracks how much memory is available to the program.

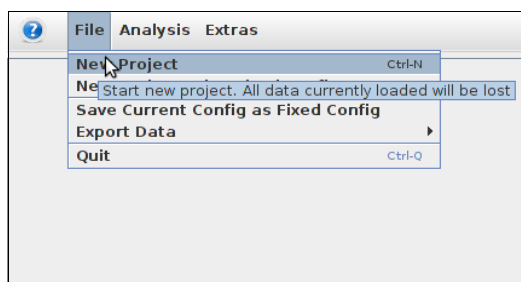
Usage

Viewing and Filtering Datasets

There are 4 basic steps to follow for viewing and filtering exome data:

Select file(s) to load

Select File → New Project, and select one or more VCF files to import.



Select filter types


Once files have been selected, VcfView will analyse the data fields present in each file and present the user with the option of applying filter-types to any of these fields.

The filter-type selection screen (top-right) contains two columns: the field identifier column and the associated filter-type column.

The user can select fields they wish to view by selected the check-box next to the field name. Once selected, the adjacent drop-down box will be enabled allowing the user to choose a filter-type for that particular field (The various filter-types are described in the following section).

Once a selection has been made, the user can either select 'Accept', or choose to save the filter-type configuration for future use (File → Save Config) (bottom-right). Previously saved configurations can be loaded from the File menu, and a default configuration can be set which will be loaded automatically each time a new project is started.

Field	Filter Type
<input checked="" type="checkbox"/> #CHROM	multi-choice
<input checked="" type="checkbox"/> POS	range
<input checked="" type="checkbox"/> ID	match
<input checked="" type="checkbox"/> REF	none
<input checked="" type="checkbox"/> ALT	none
<input checked="" type="checkbox"/> QUAL	range
<input checked="" type="checkbox"/> FILTER	multi-choice
<input checked="" type="checkbox"/> INFO	none
<input checked="" type="checkbox"/> DP	none
<input type="checkbox"/> Total Depth	none
<input type="checkbox"/> HM3	none
<input type="checkbox"/> AA	none
<input checked="" type="checkbox"/> AC	none
<input checked="" type="checkbox"/> AN	none
<input type="checkbox"/> GP	none
<input type="checkbox"/> BN	none
<input type="checkbox"/> NR	none
<input type="checkbox"/> AR	none
<input type="checkbox"/> OR	none
<input type="checkbox"/> MP	none



View and filter data

Once loaded, the selected data are shown in table format (right).

The user can rearrange the column order by dragging the column-headers and moving them to the desired position. Additionally, entries can be sorted by double-clicking the column-header of the field the user wishes to sort by. Double-clicking again will reverse sort.

Selected filter-types are shown at the left of the screen. The type of filter that the user should select can be determined by considering the type of field that the filter is to be applied to. The filter-types are:

[illegible]

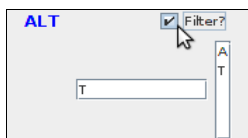
Match

☒ Filter?

☒ Partial Match?

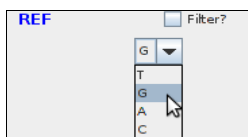
The user types their query into the lower-left text box. If the filter is activated by selecting the upper-right check box, only variants that match the query text in the filtered field will be displayed. There is also the option to allow partial matching to the query as well, by selecting the lower-right check-box. e.g. partially-matching 'rs3' in the ID field will display all variants which *contain* the letters 'rs3', rather than ones that exactly match 'rs3'

Multi-Match



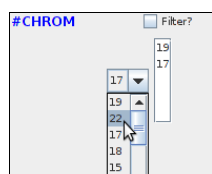
Similar to the Match filter-type, except multiple queries can be applied. The user types the desired query into the text-box and presses return. The query text will then appear in the list on the right and will be used as a filter if activated by selecting the upper-right check-box. To remove a query from the list, either double-click, middle-click or right-click the query on the list.

Choice



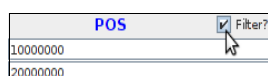
A suitable option if the field contains a fairly low number of possible entries. The user selects the filter query from the lower-right drop-down list. If the filter is then activated, only variants containing the chosen option in that field will be displayed. This filter-type is not suitable for fields containing a large number of possible entries, as it becomes very difficult to find the desired query from a very long drop-down list. For this reason, VcfView will warn the user if this filter-type is applied to a field with a large number of possible entries (e.g. ID).

Multi-Choice



Similar to the Choice filter-type, except that multiple selections can be made. The user selects their query choice from the drop-down list and is then entered into the list on the right. Once the filter is activated, any queries on the list will be used for filtering (e.g., in this case only variants present on chromosomes 17 and 19 would be displayed). Queries can be removed from the list by double-clicking, middle-clicking or right-clicking the query on the list. As with the Choice filter-type, the software will warn the user if this filter-type is applied to a field with a large number of possible entries.

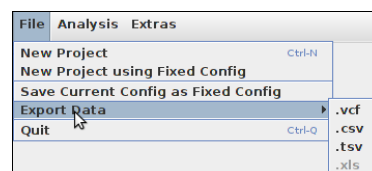
Range



Only suitable for use on fields containing numerical data, e.g. chromosomal position, quality scores, etc. The user enters a lower-bounds value in the top box, and an upper-bound value in the bottom box. When activated, only entries that have data in this field that lies between the upper- and lower-bound values will be displayed (e.g. in this case, only variants between chromosomal positions 10m and 20m will be shown).

Saving data

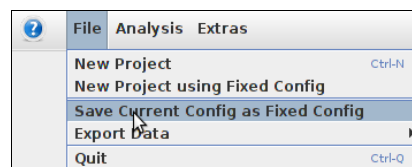
Once the user is happy with their data, they can export the variants displayed to file (right) by selecting File → Export Data → .vcf/.csv/.tsv. The data can be exported back to VCF format or to tabulated format for simple viewing and editing. If multiple VCF files were



loaded and the data are saved in tabulated format (either tab-separated or comma-separated), all samples will be saved to the same table. However, if the user chooses to output data as VCF, one file will be generated for each sample loaded.

Saving configuration

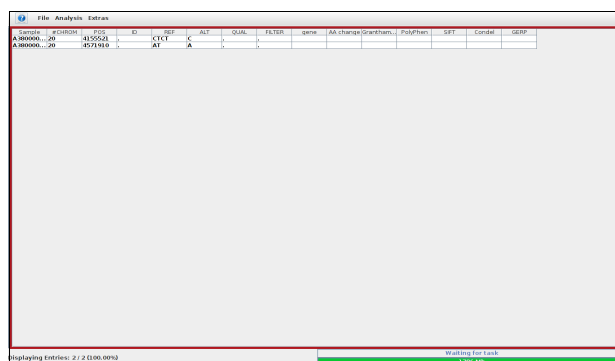
If the user wishes to reuse the selected filter parameters at a future point, they are able to save them to file (right) by selecting File → Save Current Config as Fixed Config.



The configuration/set-up is classed as a 'Fixed' one because when it is used again, the parameters can not be changed. This can be useful for filtering multiple data-sets using the same parameters for each one. By having the filter-parameters fixed, the risk of accidentally changing one of the parameters is eliminated (unless the user manually changes the contents of the fixed configuration file).

Loading fixed-configuration data

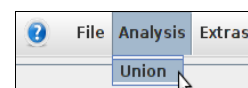
To use a previously saved fixed-configuration file in a new project, instead of selecting File → New Project, the user instead selects File-> New Project using Fixed Config. They will then be prompted to select a fixed configuration file (.fc) and then one or more VCF files. To remind the user that they are in fixed configuration mode, the border of the table is painted red (right).

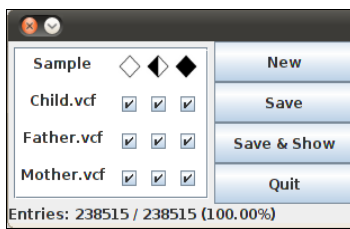


Comparing data-sets

Multiple data-sets can be compared, and filtered using the Union tool. The Union tool is named so because VcfView scans the input VCF files and generates a list of all positions in all data-sets where a variant is found, effectively creating a union of all variants. From this list, filters can be applied. e.g., Show variants that are present in sample A but absent in sample B, or variants that are heterozygous in sample A but homozygous in sample B.

To carry out a union analysis, the user selects Analysis → Union (right). They will then be prompted to select at least two VCF files for comparison. Once loaded, the user will be presented with the following screen:





The left panel contains four columns. The first column shows the sample file name, and the following three columns are check-boxes used to apply filtering criteria: The check-boxes are for selecting non-variants, heterozygous variants and homozygous variants, respectively.

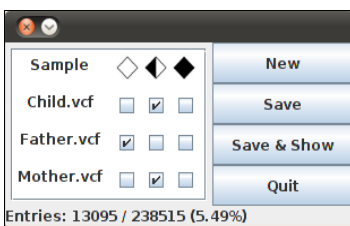
The right panel contains options to restart the analysis with new files ('New'), to export the filtered VCF files ('Save'), to export the filtered VCF files and display the results in the main table view ('Save & Show'), or to quit the union analysis ('Quit').

The bottom bar shows the number of variant positions currently selected vs the total number of positions where variants are present among all samples.

The usage of the union analysis is best illustrated with an example:

A researcher has exome data for a mother, father and child. They know that the mother and child both have an inherited condition, and that the father does not have this condition. They are also aware that the condition is very rare, leading them to the assumptions that the pathogenic variant is dominant, that the mother and child are both likely to be heterozygous for the pathogenic variant and that the father does not have the variant.

From this, the researcher can use the union analysis to narrow down a list of candidate variants by selecting the check-boxes as follows:



The result of this selection is that only variants which are (1) absent in the father, (2) heterozygous in the child and (3) heterozygous in the mother, will be saved. In this example, this corresponds to 5.49% of the dataset, narrowing down the search by ~20-fold.

The user can then save and show the data in table format for further filtering. When data is saved, the user is asked to provide a file prefix and a new VCF file will be generated for each sample that was used in the analysis, named <PREFIX>_<ORIGINAL-NAME>.vcf.

Important: It should be noted that the genotype of each variant is determined by analysing the genotype (GT) field present in the VCF file. This field is present in the last column, immediately following the FORMAT column. Currently, VcfView will only accept files where the GT information is the **first** variable present in this field. This will be modified in a future release.

Also, VcfView only checks the chromosome and position of variants to determine whether they are present in each sample. It does not account for any other information such as the reference and alternative bases, quality scores, or whether the variant passed filtering. This should be considered when carrying out this analysis, and will be addressed in a future release.