Extractor quick-guide

Flip Mulder
UMC Utrecht Genetics Department

Introduction:

Extractor is a tool which was intended to easily get information from the ExAC website, allowing you to query search terms in batch by specifying an input file containing a list of genes.

The content could then be quickly filtered using the graphical interface allowing you to quickly explore results without having to save the table and working in excel or having to do new searches each and every time.

The tool has been modified to also work with Gnomad but hasn't been tested as extensively yet, although it DOES seem to work just fine as well.

Just keep this in mind though…

Extractor is a work in progress and to be used at your own risk!

The program is still in development so bugs and issues are bound to exist, so please double check where needed.

Also if you have any issues, bugs, remarks or requests feel free to contact me at:

f.mulder@umcutrecht.nl

Where possible I will try to address them as quickly as possible. Also regarding certain features you might miss, workflows which don't make sense, etc… Please let me know and while I can not guarantee anything I will surely take it into account and will see if it's possible to adjust and improve things where necessary.

Using Extractor:

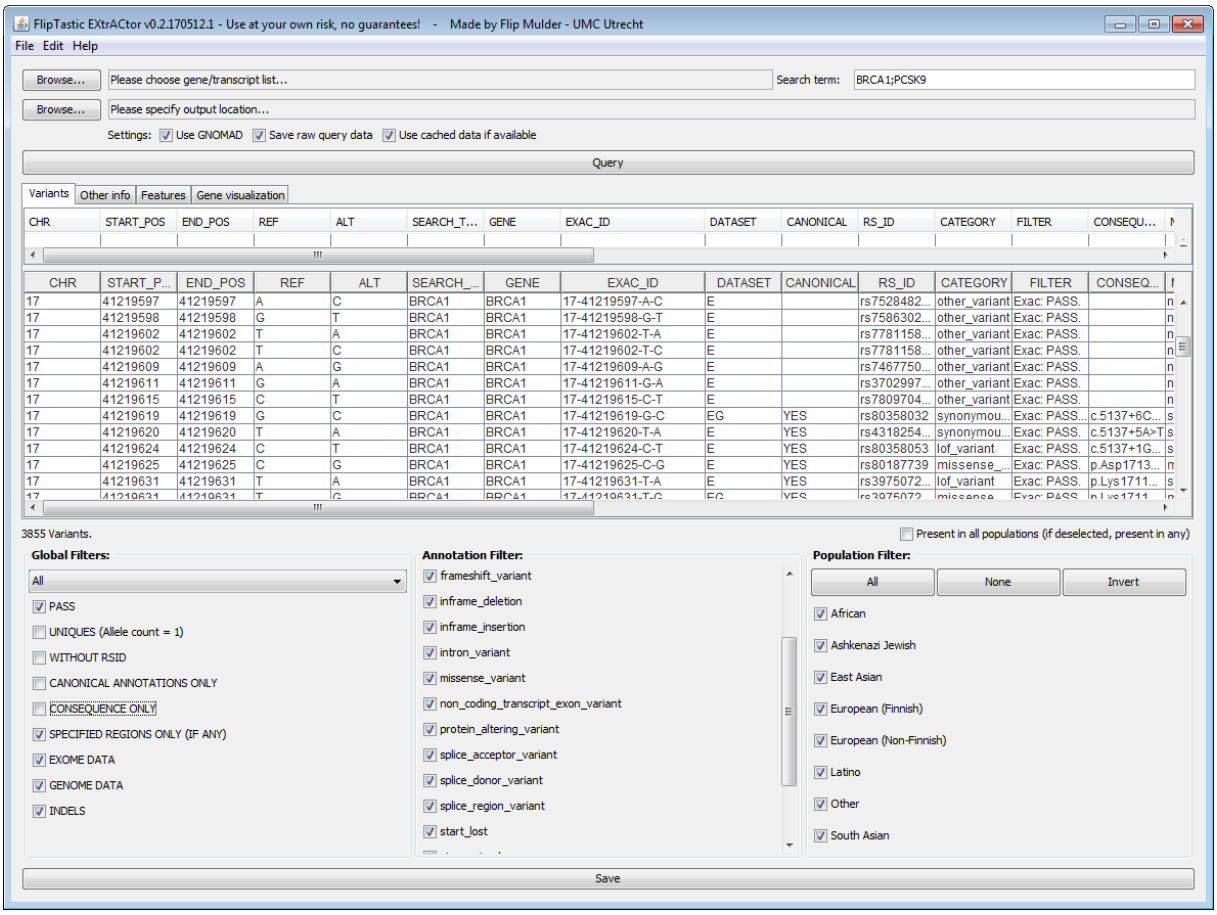
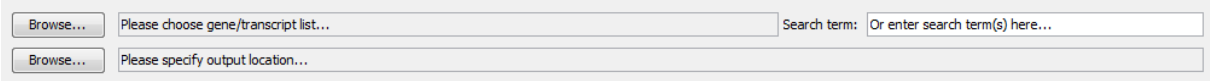


Figure 1: Extractor Mainscreen

In- and output

When starting Extractor the main screen will be shown. At the top there is a browse button allowing you to specify a tab-delimited text file containing a list of genes, transcripts, regions or genes/transcripts and locations.



At the right there is a search field where you can quickly enter a single search term or multiple search terms separated by a semi-colon (;).

An output file can be specified by pressing the corresponding browse button, but if none was specified you will be asked to specify it when you actually try to save the results.

Input format:

When an input file is used this should be a tab delimited text file consisting of one or two columns. The first column should contain the list of primary search terms, meaning gene name, transcript, or region.

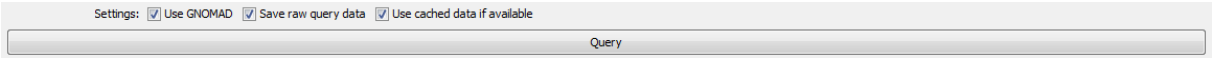
If desired a second column can be specified with sub-regions. This can be either in the format chr:start-end (for example 22:46615715-46615880) or certain protein change location (for example p.Gly23Ter).

In the latter case wildcards in the form of * can be used, or when the aminoacid after the position is left empty this is also treated as a wildcard.

Examples of uses are:

p.Gly23tTer:	Search for specific Gly to Ter change on position 23
p.Gly23* (Same as p.Gly23):	Search for Gly to any amino acid change on position 23
p.*23Ter:	Search for any amino acid to Ter change on position 23
p.*23*	Search for any changes on position 23

(Query) Settings

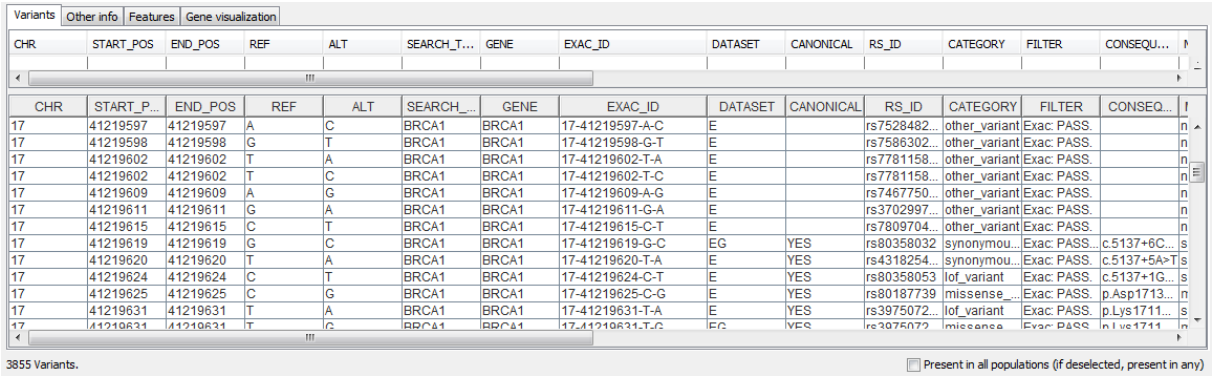


Above the query button, which will actually start getting the data, there are a few options. The tool works on both ExAC and Gnomad at the same time, but disabling 'Gnomad' will force it to ExAC only mode which was the original version. You can specify whether you wish to save the queried data (Save raw query data) so this can be used for future searches and whether you actually wish to use the cached data on the computer. Using cached data speeds up the process a lot as no new queries are necessary for each search term.

Results

Below the query-button there are a number of tabs which display the results from your queries.

Variants

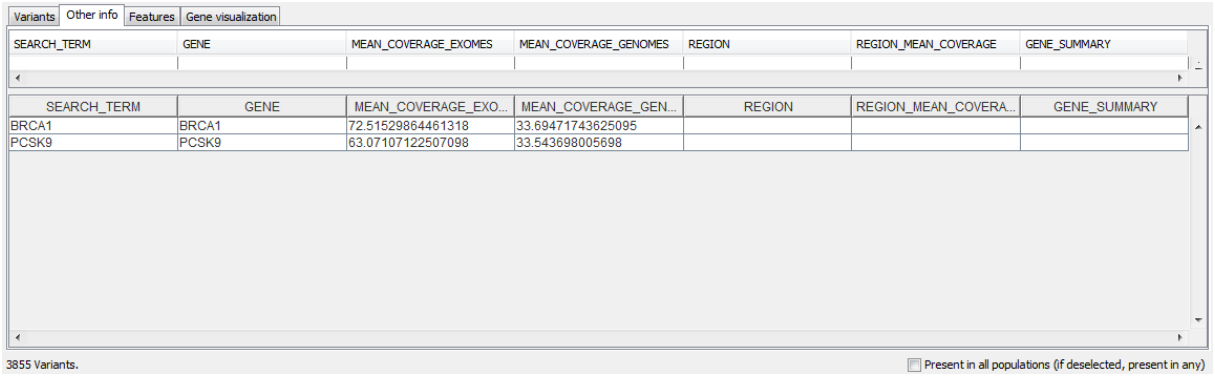


CHR	START_P...	END_POS	REF	ALT	SEARCH_...	GENE	EXAC_ID	DATASET	CANONICAL	RS_ID	CATEGORY	FILTER	CONSEQ...
17	41219597	41219597	A	C	BRCA1	BRCA1	17-41219597-A-C	E		rs7528482...	other_variant	Exac: PASS.	
17	41219598	41219598	G	T	BRCA1	BRCA1	17-41219598-G-T	E		rs7586302...	other_variant	Exac: PASS.	
17	41219602	41219602	T	A	BRCA1	BRCA1	17-41219602-T-A	E		rs7781158...	other_variant	Exac: PASS.	
17	41219602	41219602	T	C	BRCA1	BRCA1	17-41219602-T-C	E		rs7781158...	other_variant	Exac: PASS.	
17	41219609	41219609	A	G	BRCA1	BRCA1	17-41219609-A-G	E		rs7467750...	other_variant	Exac: PASS.	
17	41219611	41219611	G	A	BRCA1	BRCA1	17-41219611-G-A	E		rs3702997...	other_variant	Exac: PASS.	
17	41219615	41219615	C	T	BRCA1	BRCA1	17-41219615-C-T	E		rs7809704...	other_variant	Exac: PASS.	
17	41219619	41219619	G	C	BRCA1	BRCA1	17-41219619-G-C	EG	YES	rs80358032	synonymou...	Exac: PASS.	c.5137+6C...
17	41219620	41219620	T	A	BRCA1	BRCA1	17-41219620-T-A	E	YES	rs4318254...	synonymou...	Exac: PASS.	c.5137+5A>T
17	41219624	41219624	C	T	BRCA1	BRCA1	17-41219624-C-T	E	YES	rs80358053	lof_variant	Exac: PASS.	c.5137+1G...
17	41219625	41219625	C	G	BRCA1	BRCA1	17-41219625-C-G	E	YES	rs80187739	missense_...	Exac: PASS.	p.Asp1713...
17	41219631	41219631	T	A	BRCA1	BRCA1	17-41219631-T-A	E	YES	rs3975072...	lof_variant	Exac: PASS.	p.Lys1711...

The variants tab shows a table with all the variants matching your search terms and specified filter settings. This table is updated in real time so any changes you make to the filters are shown right away, including a count of the displayed results.

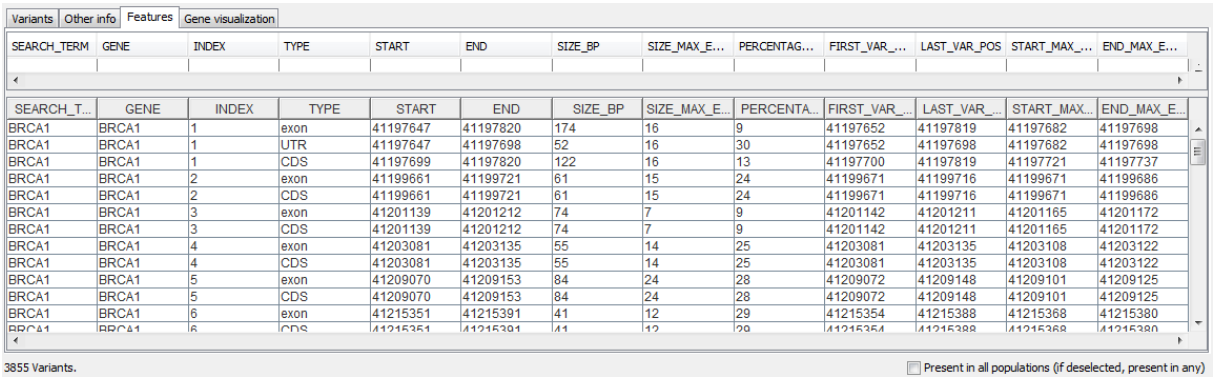
When saving the results this is also what is being saved, what you see is what you get.

Other info



Some other (general) information regarding your searches is displayed here, such as mean coverage info for the entire gene/transcripts and for the specified sub-regions if any. Also a gene summary field is included giving you some general information about the genes function, based on refseq gene summary data.

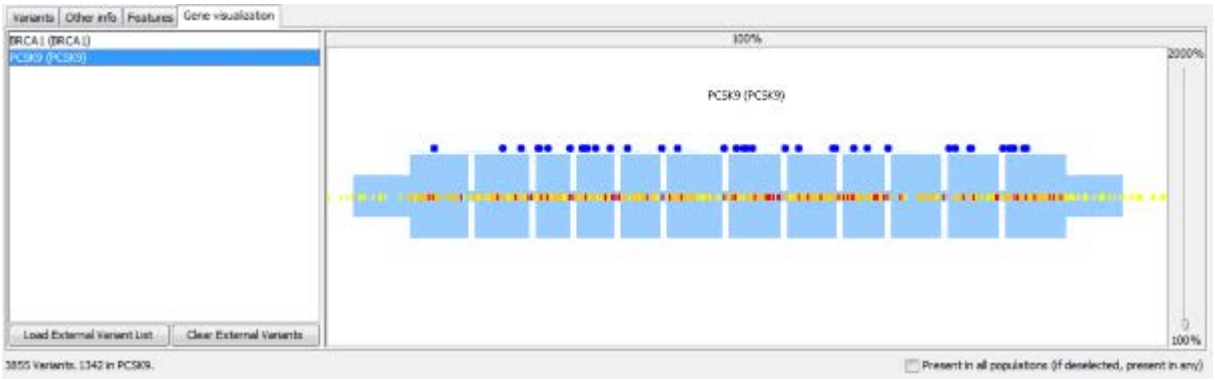
Features



The features tab contains information about the features of the (canonical) transcripts of the gene or transcript queried. This includes:

i. The type (UTR, exon, CDS)
ii. Start location of feature
iii. End location of feature
iv. Feature size
v. Max size of region without any variants
vi. Percentage of this max size of the feature which is without variants
vii. Position of first variant in the feature
viii. Position of the last variant in the feature
ix. Start of the largest empty region in the feature
x. End of the largest empty region in the feature

Gene visualization



The gene visualization tab displays a graphical representation of the selected gene or transcript. The variants are plotted in the gene so one can quickly see the distribution and locations of various variants. Again the variants displayed are based on selected filters and updated real time.

Hoovering over the exons and variants displays some information of the element, double-clicking a variant takes you to the selected variant in the variant-table from the variants tab.

The gene of interested can be selected from the gene list on the left and one can also zoom in and out (either by ctrl^ dragging with the mouse, ctrl^ scroll wheeling or using the zoom bar on the right side of the plot.

By right clicking in the gene visualization panel a small menu is shown allowing the user to save the currently drawn image to .png format or to display the actual UTR sizes. By default UTR's are drawn as a fixed size in order to show the exons as clearly as possible.

Below the gene lists the user can find two buttons, Load External Variant List and Clear External Variants. The first allows the user to specify a list of variants, in the (tab delimited) format:

chr start_pos end_pos <other_columns>

The variants are then plotted in the gene visualization tab so one can see where variants of interest are located in the gene and compare that with localizations of other variants. They are represented as blue dots above the gene. When hoovering over these variants the specified <other_columns>, which can simply be any number of columns following end_pos, are used to display information about the variant. Simply in the form of one extra row per column.

The Clear External Variants button can be used to clear the currently loaded variants.

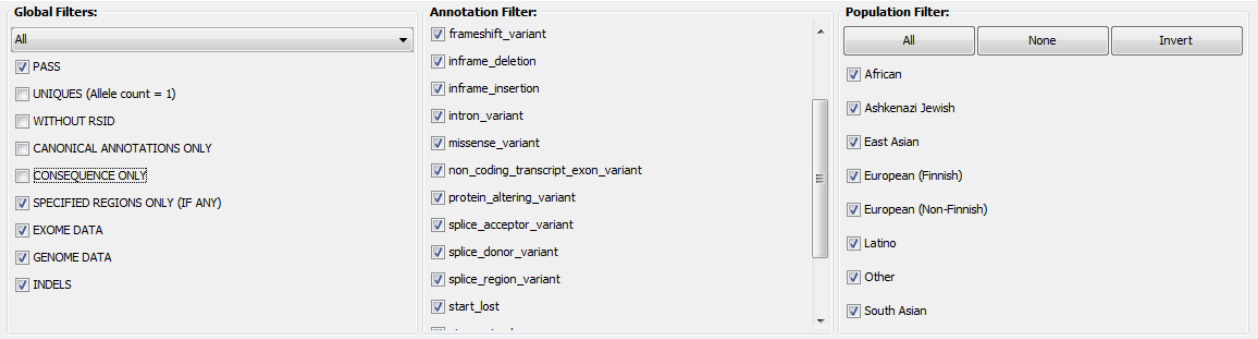
Quick Filter

The tables from the result tabs contain an extra row at the top, this is the so-called quick filter. By entering a value in this row the user can quickly filter on data as well, showing only the rows matching that value. Multiple columns can be combined and by right-clicking the

row a popup menu is shown allowing you to adjust some specific settings or clear the entries.

This quick filter row can be used in combination with the Filter panels at the bottom. More on the quick filter in the Filters section below.

Filters



The filters allow the user to quickly explore and make selection in the data. All changes made here will be applied instantly to the results from above.

As can be seen there are three sections:

Global Filters

These are basically the same filters as available on the ExAC/Gnomad website. One can specify to display only LOF, Missense+LoF or All variants. Then one can choose to only show those having a PASS entry, only Unique varians, Only Canonical annotations, Only variants with an annotated consequence and only those in specified regions (By default the entire gene is queried so disabling this will show the all the variants returned in the gene or transcript corresponding the region of interest). Whether to display results from exome and/or genome data (as indicated by a G, E, g or e in the dataset column in the results table).

Annotation Filter

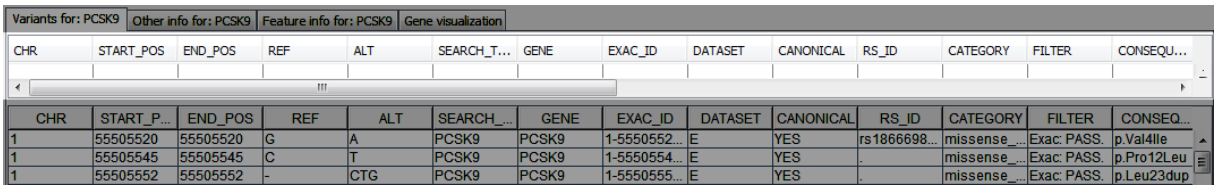
This allows the user to easily select specific variants, the entries are dynamic in the way that they are based on the data retrieved. Using the All, None and Invert button one can easily select/deselect larger parts.

Population Filter

This allows the user to only show variants specifically occurring in certain populations (or not occurring in them). Frequency and allele counts based on populations in the main variant table are adjusted based on these settings.
The user can also choose to use any or all mode for the population filter by (un)checking the appropriate box above it. And also here the All, None and Invert buttons are available.

Quick Filter

The quick filter allows the user to quickly search for value by entering them in the quick filter row which is displayed above the table data itself.



When the column data type is a numeric value (such as allele counts, frequencies) the filter also accepts comparisons such as larger then and equal or smaller then. The possible options are: >, <, >=, <=, !=. By right clicking the quick filter row a menu is displayed where other options can be specified (case sensitivity, whole words or substring, clear all settings). Each setting is applicable for the corresponding column only.

Saving the data

Once the user has selected the variants of interest by choosing the appropriate filtering steps the currently displayed variants can be saved to a file by pressing the 'Save' button at the bottom.

If an output file was specified in step a) above this will be used for saving the data, otherwise the application will ask the user to specify the output file in this step.

Saving (filter) settings

All most recently used changes will be stored by the program and restored at next launch.