Just-DNA-Seq

Release 0.1

VovixSW

Dec 26, 2022

CONTENTS

Just-DNA-Seq is a project to facilitate working with human genomes at all levels - from clinical cases to personal curiosity, education and longevity. We envision the future when genomics is becoming more *available*, *understandable* and *useful* for everybody, especially those interested in life extension and improving human condition.

For this purpose we use OakVar framework to integrate *annotators* which are tools or databases accumulating what we actually know about the genome - genes, their influence on health or drug response, polygenic risk scores and so on.

We created the **Oakvar-Longevity** module for OakVar. It annotates user genome with longevity associated gene variants using the longevitymap annotator, and provides the longevity PRS, variants, drugs and major risks (oncorisk) reports with the longevity-combined reporter.

Check out the *Getting Started* section for further information, including the assemblies section to understand whether you have the appropriate genome assembly.

NOTE: This project is under active development.

CHAPTER

ONE

CONTENTS

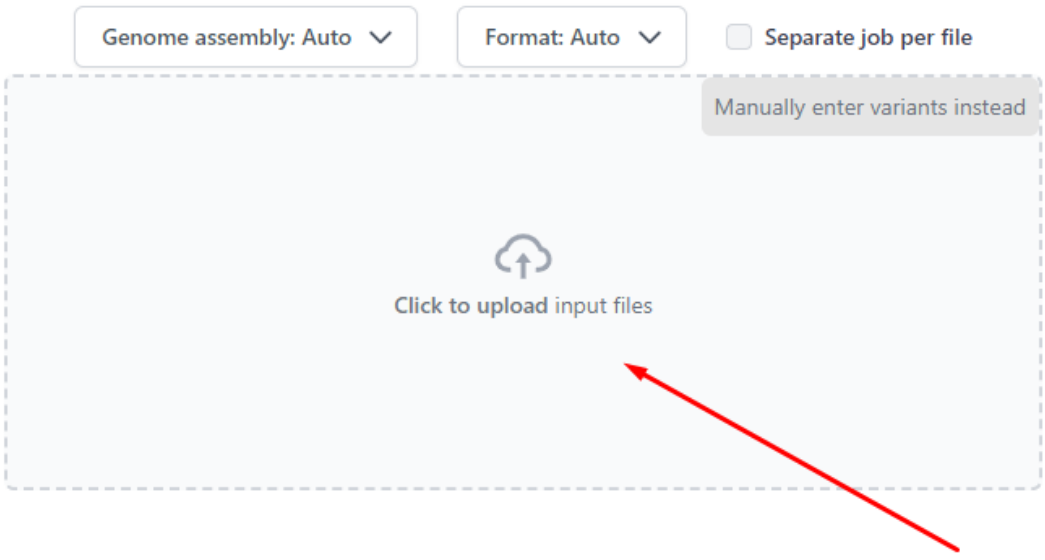
1.1 Getting Started

NOTE: *Both OpenCravat and OakVar can be used to annotate a human genome. At the beginning of the project, we used OpenCravat as a framework. However, as OakVar is based on OpenCravat and contains more advanced features customized specially for personal longevity genomics, we decided to base further development of the project on OakVar.*

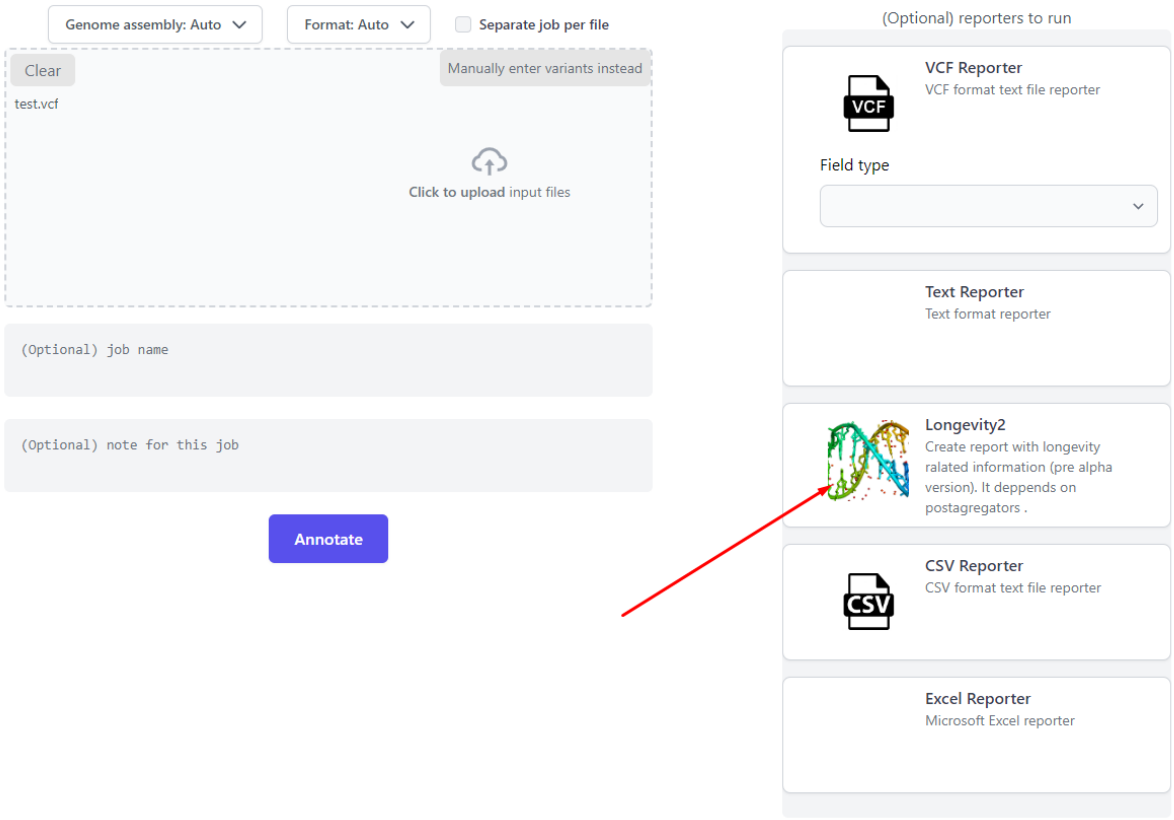
First, you need to open http://agingkills.eu:8081/index.html

1.1.1 Loading Genome Files

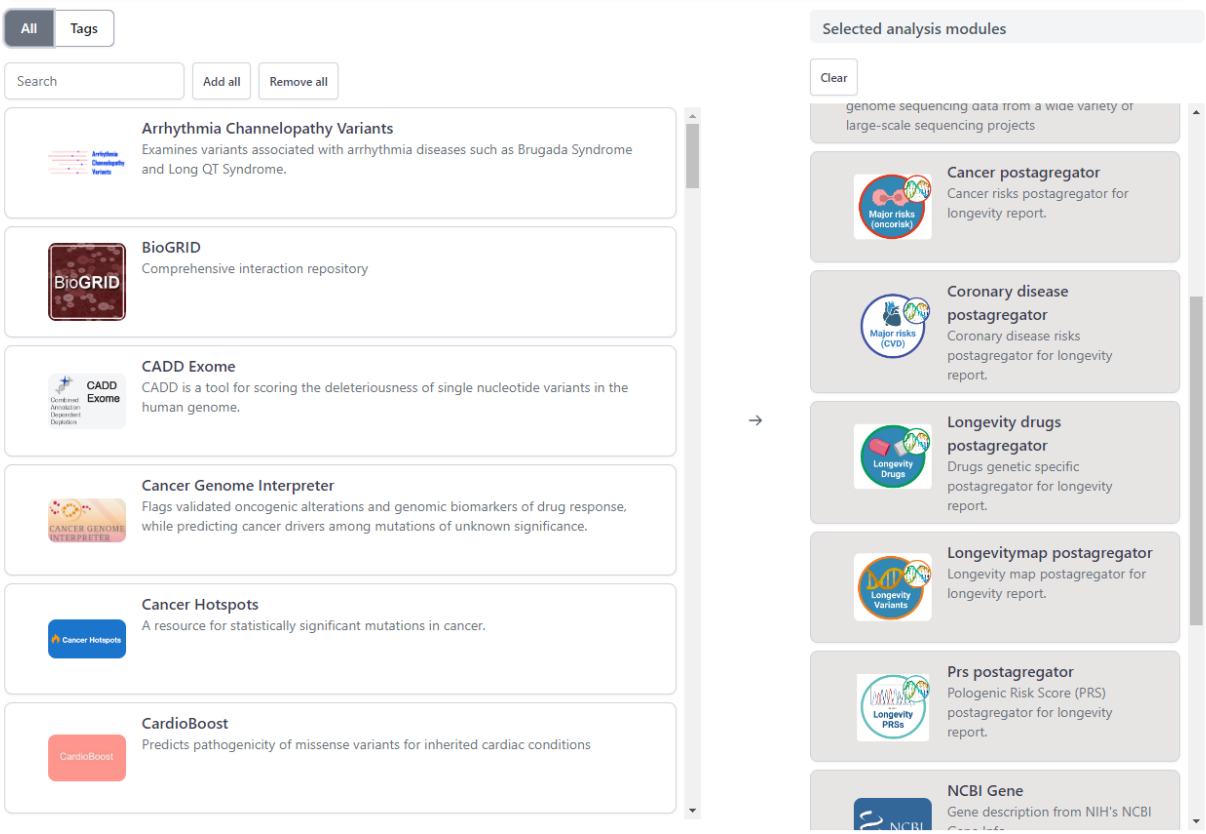
In the "Submit" section click on "Click to upload input files" to upload your genome file.



Then choose Longevity2 in the reporters window:

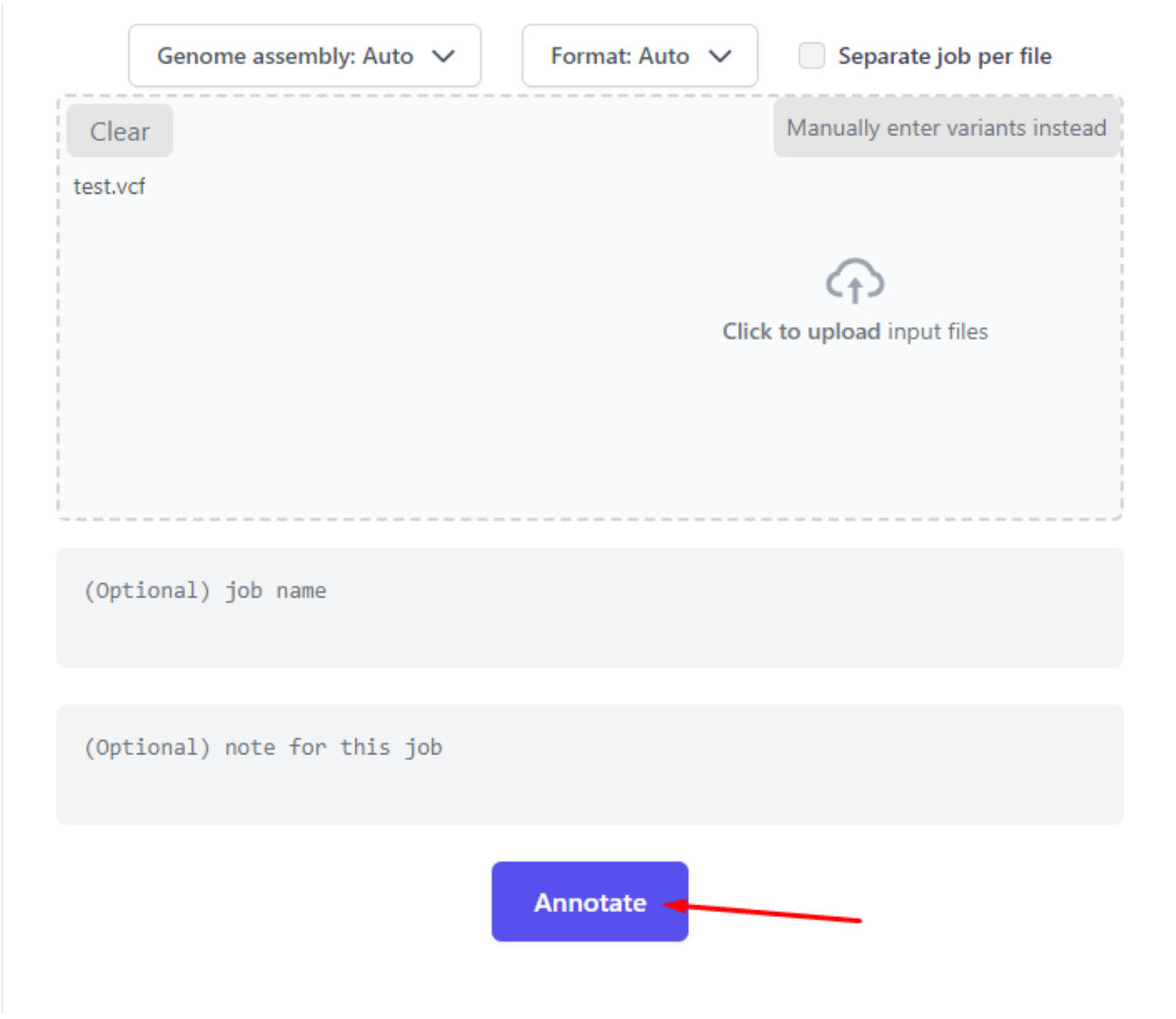


All the necessary modules will be automatically selected. If you are interested in the other modules you can select them in the modules section:

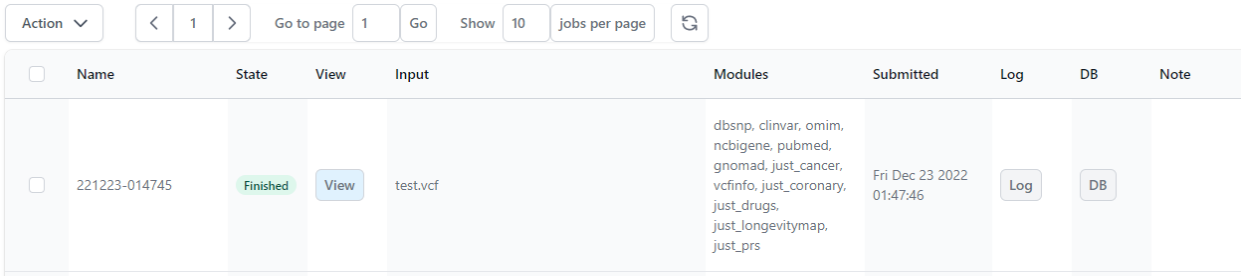


1.1.2 Annotating Your Genome

When you select all the annotators you need, click the **ANNOTATE** button below in the left area.



Annotating a large genome file may take some time. While loading, it will be displayed in the "jobs" section, displaying different stages of the processing in the **State** column, and when finished, the **View** button will appear:

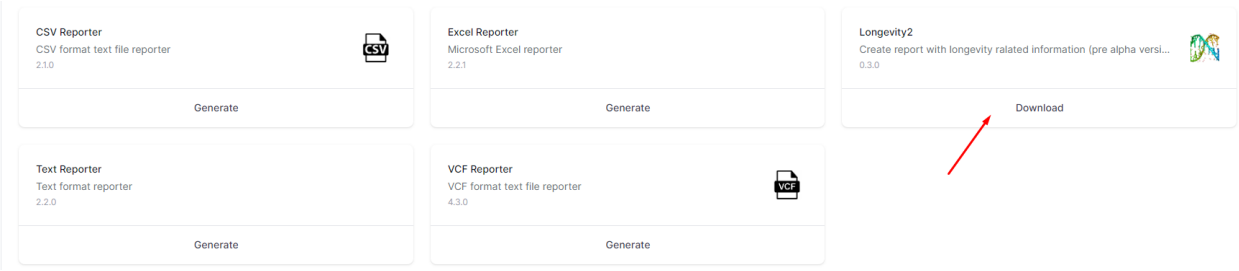


1.1.3 Opening Your Annotated Genome

Now click the **View** button, and the annotated genome will open in a new browser tab/window.

1.1.4 Getting Longevity2 report

In the new window open the "Reports" section and click the "Download" button in the Longevity2 reporter to download the longevity report:



1.2 Viewing Reports

Just-DNA-Seq makes a set of reports on a genome that contain data about PRS, longevity-related gene variants, known cancer risks, coronary artery disease risks, and drug responses.

You can learn how to get your longevity report and install Longevity Combined reporter in *Getting Started*.

Source code on GitHub (here you also can learn how to install it from GitHub): https://github.com/dna-seq/oakvar-longevity

1.2.1 Part 1: Polygenic risk scores (PRSs)

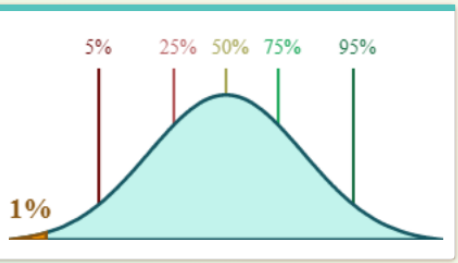
Longevity2 report contains Longevity PRS and Health Risk PRS.

PRS is represented as a percentile within a given population. For example, if you have the 95th percentile in Longevity PRS, it means your genetic chances to gain extreme longevity is higher than 95 out of every 100 people in a chosen population.

If you have the 70th percentile in Heath Risk PRS, which means your personal genetic risk of the disease development is higher than 70 out of every 100 people in a chosen population.

It does not take into account environmental factors and is based on genetics only.

Title	Longevity PRS (PRS5)
Sum	0.451574635478157
Count/total	1/332
Average	0.2257873177390785
Percentile	1%



1.2.2 Part 2: Longevity Significant Variations

This report contains gene variants that have a significant influence on longevity. It contains the following columns:

+ - clicking this green button opens detailed information on each entry (row), and the button becomes red with **-** sign. Clicking this **-** closed the details. Clicking **+** in the header opens the details for all rows and behaves in the same way (clicking **-** in the header closes all detail sections).

RSID - reference sequence ID of the variant.

Population - population(s) on which the research was conducted, e.g. Greek, Ashkenazi Jewish, etc., or multiple (for more details, open **+**).

Gene - gene the variant belongs to.

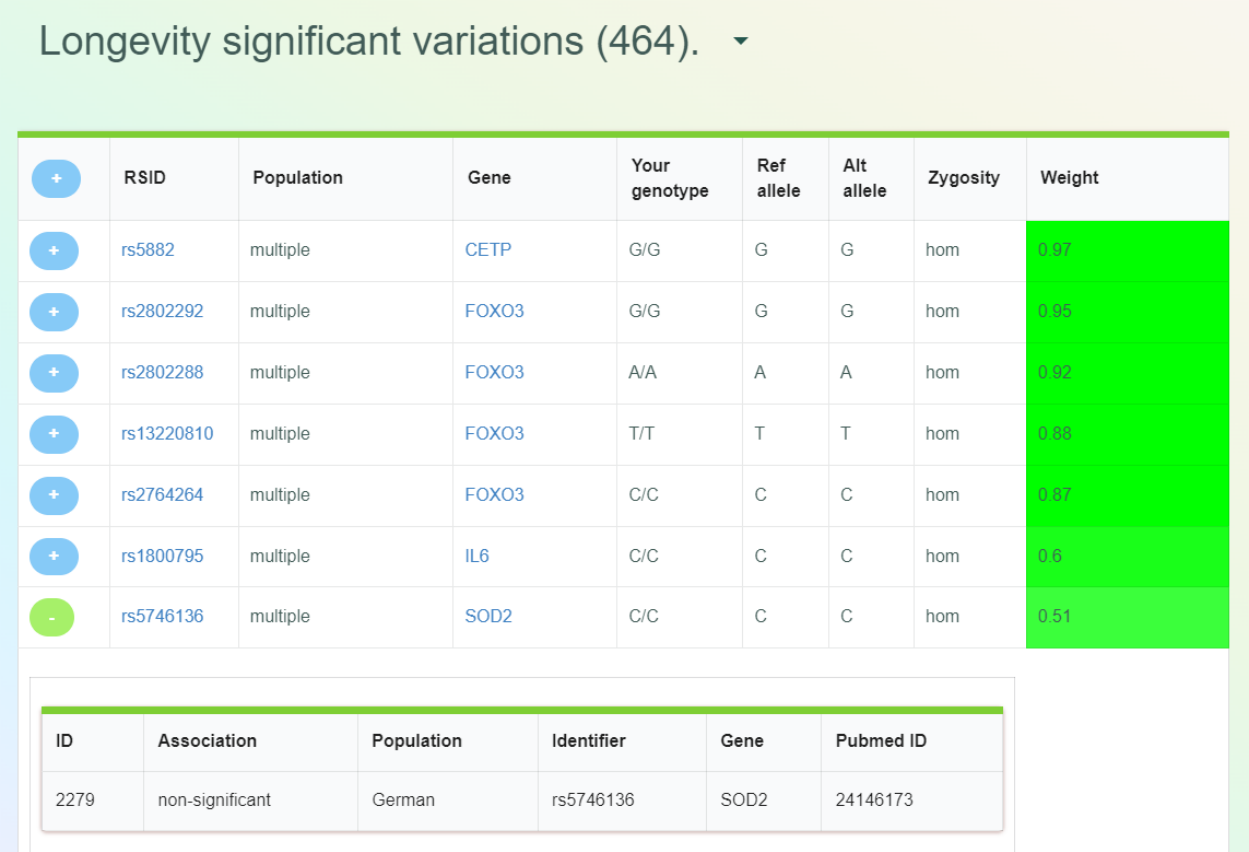
Your Genotype - which variants your genome contains. Note that in the case of homozygosity two letters should be the same, and for heterozygosity, they differ.

Ref allele - reference allele.

Alt allele - alternative allele.

Zygosity - hom (homozygosity) or het (heterozygosity).

Weight - the weight of this variant (the degree of significance).



Longevity significant variations (464).

+	RSID	Population	Gene	Your genotype	Ref allele	Alt allele	Zygosity	Weight
+	rs5882	multiple	CETP	G/G	G	G	hom	0.97
+	rs2802292	multiple	FOXO3	G/G	G	G	hom	0.95
+	rs2802288	multiple	FOXO3	A/A	A	A	hom	0.92
+	rs13220810	multiple	FOXO3	T/T	T	T	hom	0.88
+	rs2764264	multiple	FOXO3	C/C	C	C	hom	0.87
+	rs1800795	multiple	IL6	C/C	C	C	hom	0.6
-	rs5746136	multiple	SOD2	C/C	C	C	hom	0.51

ID	Association	Population	Identifier	Gene	Pubmed ID
2279	non-significant	German	rs5746136	SOD2	24146173

1.2.3 Part 3: Drugs Report

This report contains known issues of response to certain drugs associated with gene variants. It has the following columns:

- number of an entry (row).

Variant/Haplotypes - by rsID.

Drug(s) - name(s) of drug(s) response to which is affected by the variants.

Phenotype Category: Efficacy, Dosage, or Other.

Significance - yes or no.

Sentence - description of the case.

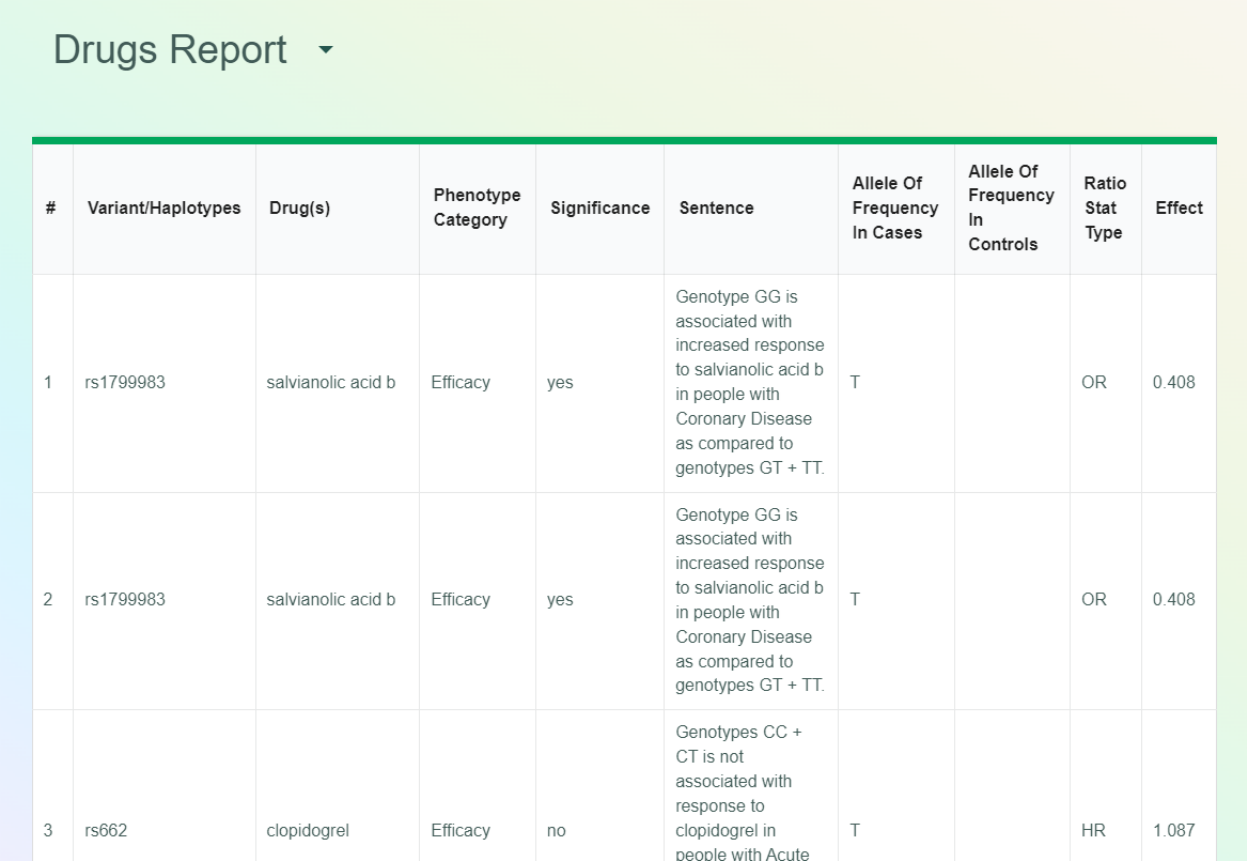
Allele Of Frequency In Cases - allele of the variant (one or more letters A, T, C or G) in cases involved.

Allele Of Frequency In Controls - allele of the variant in controls.

Ratio Stat Type

Effect

Note: In some browsers, the last one or two columns may be found beyond the visible area at 100% zoom level; in such cases try zooming out to 90%, 80%, and so on until everything is visible.



#	Variant/Haplotypes	Drug(s)	Phenotype Category	Significance	Sentence	Allele Of Frequency In Cases	Allele Of Frequency In Controls	Ratio Stat Type	Effect
1	rs1799983	salvianolic acid b	Efficacy	yes	Genotype GG is associated with increased response to salvianolic acid b in people with Coronary Disease as compared to genotypes GT + TT.	T		OR	0.408
2	rs1799983	salvianolic acid b	Efficacy	yes	Genotype GG is associated with increased response to salvianolic acid b in people with Coronary Disease as compared to genotypes GT + TT.	T		OR	0.408
3	rs662	clopidogrel	Efficacy	no	Genotypes CC + CT is not associated with response to clopidogrel in people with Acute	T		HR	1.087

1.2.4 Part 4: Cancer Report

This report shows variants with known cancer risks and contains the following columns:

+ - acting the same way in all reports (see above).

- number of an entry (row).

Chrom - chromosome the variant belongs to, e.g. **chr1**.

Position - position of the variant on the chromosome (number).

Gene - gene the variant belongs to, like in the previous report.

RSID - reference sequence ID of the variant.

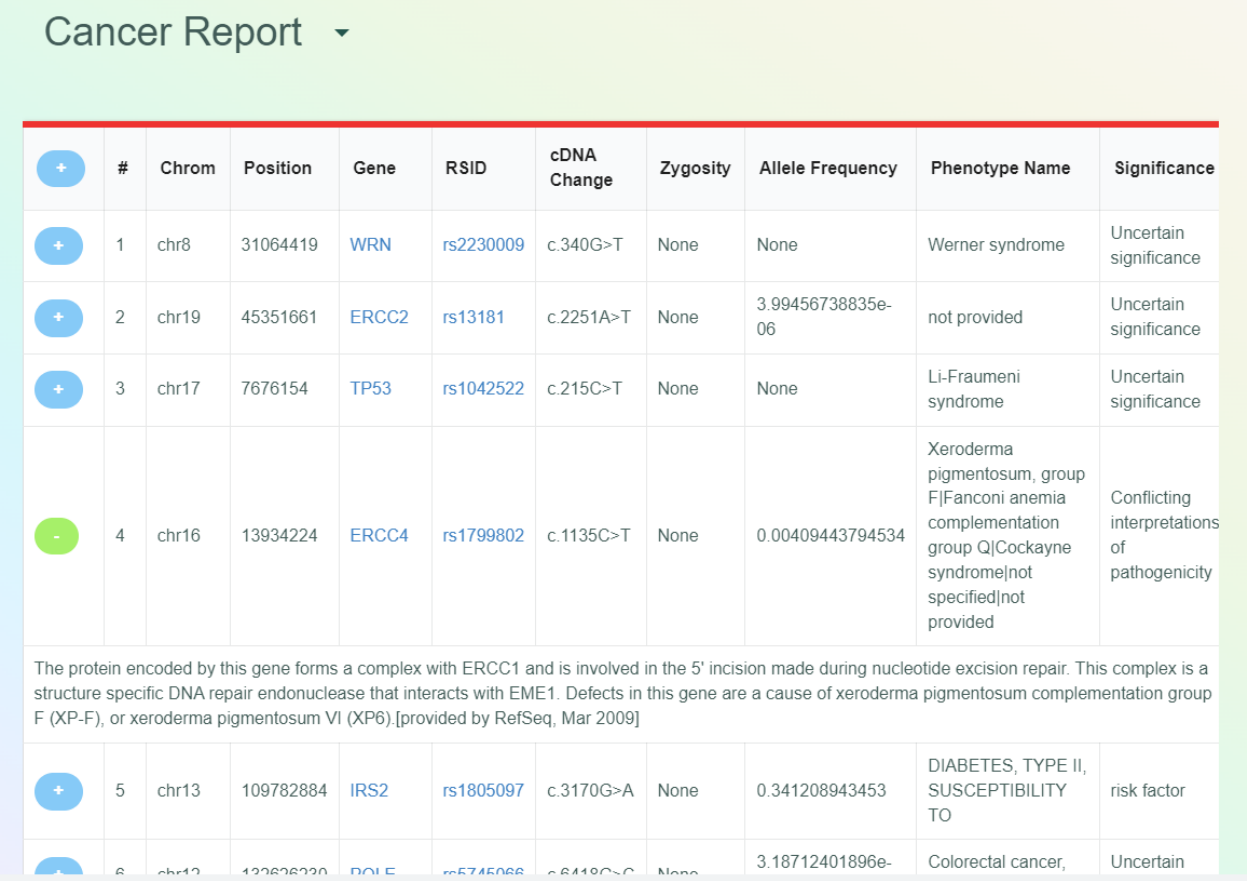
cDNA Change - change in the coding DNA by the variant.

Zygosity - hom or het (see above).

Allele Frequency - frequency of the allele.

Phenotype Name - description of condition(s) associated with the variant.

Significance - description of the significance of this variant.



+	#	Chrom	Position	Gene	RSID	cDNA Change	Zygosity	Allele Frequency	Phenotype Name	Significance
+	1	chr8	31064419	WRN	rs2230009	c.340G>T	None	None	Werner syndrome	Uncertain significance
+	2	chr19	45351661	ERCC2	rs13181	c.2251A>T	None	3.99456738835e-06	not provided	Uncertain significance
+	3	chr17	7676154	TP53	rs1042522	c.215C>T	None	None	Li-Fraumeni syndrome	Uncertain significance
-	4	chr16	13934224	ERCC4	rs1799802	c.1135C>T	None	0.00409443794534	Xeroderma pigmentosum, group F|Fanconi anemia complementation group Q|Cockayne syndrome|not specified|not provided	Conflicting interpretations of pathogenicity
The protein encoded by this gene forms a complex with ERCC1 and is involved in the 5' incision made during nucleotide excision repair. This complex is a structure specific DNA repair endonuclease that interacts with EME1. Defects in this gene are a cause of xeroderma pigmentosum complementation group F (XP-F), or xeroderma pigmentosum VI (XP6).[provided by RefSeq, Mar 2009]										
+	5	chr13	109782884	IRS2	rs1805097	c.3170G>A	None	0.341208943453	DIABETES, TYPE II, SUSCEPTIBILITY TO	risk factor
+	6	chr12	132626220	POLE	rs5745066	c.6418C>G	None	3.18712401896e-	Colorectal cancer,	Uncertain

1.2.5 Part 5: Coronary artery disease Report

The report shows variants with known coronary arthery disease risks and contains the following columns:

+ - acting the same way in all reports (see above).

- number of an entry (row).

RSID - reference sequence ID of the variant.

Gene - gene the variant belongs to, like in the previous report.

Risk Allele - the allele that contains the risk trait

Genotype - which variants your genome contains. Note that in the case of homozygosity two letters should be the same, and for heterozygosity, they differ.

Pubmed ID - the ID number to access corresponding articles on PubMed

Population - population(s) on which the research was conducted, e.g. Greek, Ashkenazi Jewish, etc., or multiple (for more details, open **+**).

P-Value - is a number describing how likely it is that your data would have occurred under the null hypothesis.

Weight - the weight of this variant (the degree of significance).

Coronary arthery disease

+	#	RSID	Gene	Risck Allele	Genotype	Pubmed ID	Population	P-Value	Weight
+	1	rs1333049	CDKN2B-AS1	C	C/G	PMID 17634449; PMID 18362232; PMID 20031606; PMID 19171343; PMID 18979498; PMID 24573017; PMID 23202125; PMID 22623978; PMID 17554300; PMID 33321069;	European; Multiethnic;	[PMID 17554300]: 1 x 10-13; [PMID 33321069]: 5 x 10-18;	0.9
-	2	rs6922269	MTHFD1L	A	A/G	PMID 17634449; PMID 17554300; PMID 19164808; PMID 22216278; PMID 17634449;	European	[PMID 17634449]: 3 x 10-8;	0.7
Conclusion: Polymorphisms in MTHFD1L, including rs6922269, have been implicated in risk for coronary artery disease (CAD). GA genotype is associated with 1.2x risk of coronary artery disease **Study design:** [PMID 17634449]: Discovery ancestry (country of recruitment): 4864 European (U.K.);Discovery sample description1,926 European ancestry cases, 2,938 European ancestry controls									
+	3	rs11206510	PCSK9	T	T/T	PMID 20864672; PMID 21378990; PMID 26343387;	Multiethnic; Asian	[PMID 26343387]: 2 x 10-8; [PMID 32469254]: 1 x 10-8;	0.0
+	4	rs55730499	LPA	T	T/T	PMID: 26343387	Multiethnic;	[PMID 26343387]: 5 x 10-39	-0.4
								[PMID: 29212778]:	

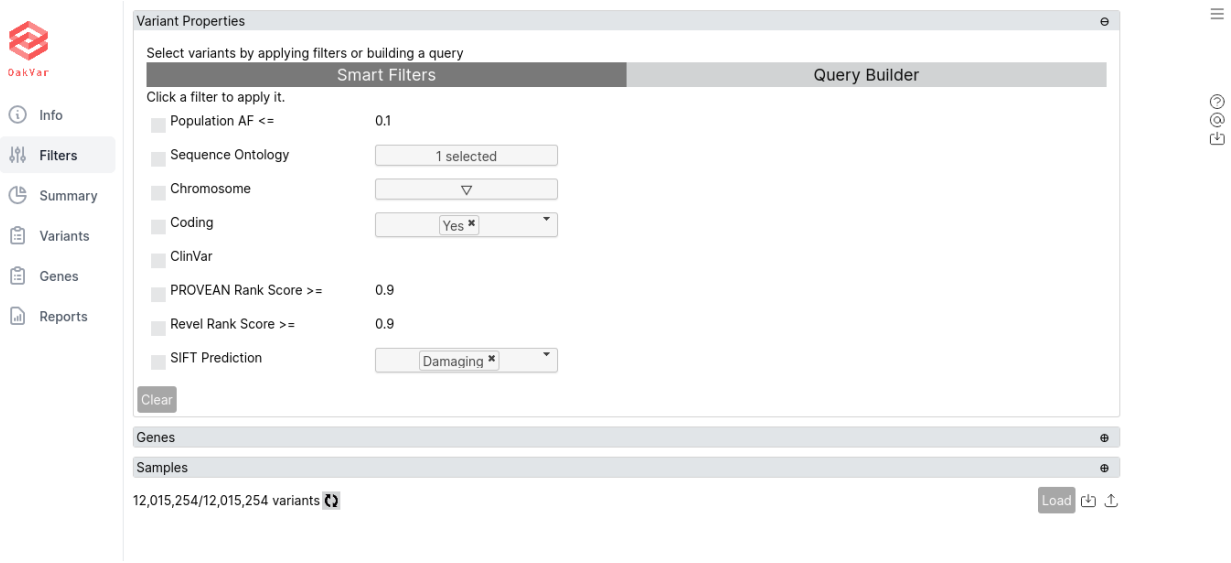
1.3 Filtering Variants

Filters in OakVar allow to select those variants which are relevant. As the number of variants in a genome usually is very large, you need to filter them first. OakVar cannot load more than 100,000 variants at once.

1.3.1 The Filter page

Select the **Filter** page in the Result Viewer. There are sections where you can filter the variants:

- *Variant Properties*, with **Smart Filters** and **Query Builder** tabs
- *Genes*, where you can type in any particular gene names
- *Samples*, which is used mainly for oncological purposes and is not used in Just-DNA-Seq.



1.3.2 Using Smart Filters

Here are various useful filters:

Population AF «==* allows to set the maximum allele frequency in population.

Sequence Ontology allows to choose one or more sequence ontologies.

Chromosome allows to choose one or more chromosomes *or their specific versions.* E.g. chr1, chr10 and so on.

Coding allows to include only coding or noncoding variants.

ClinGen allows to include only variants with data from ClinGen.

ClinVar allows to include only variants with data from ClinVar.

dbSNP Common ID allows to include only variants with dbSNP common IDs.

PROVEAN Rank Score >= includes variants with PROVEAN rank score not less than 0.9.

Revel Rank Score >= includes variants with Revel rank score not less than 0.9.

SIFT Prediction can be set as Damaging or Tolerated.

1.3.3 Using Query Builder

Here you can create a set of filter rules.

By default, an opening (left) parenthesis appears with **+** and **(** buttons in the lower left corner, and a greyed out **NOT** switch appears if you hover the mouse in the upper left corner, which allows to make the following rule negative by clicking on it. Clicking **NOT** once again deactivates it.

Click **+** to add a rule. A line of boxes will appear:

The first drop-down box is the source to which the rule will apply. For example: Variant Annotation, ClinVar, PharmGKB etc. The second drop-down box allow to select an item in the source to apply the rule. E.g. UID, Chrom, Position, Gene etc. The following "not" switch, greyed out (inactive) by default, allows to select if the following condition should apply or should not apply. For the latter, click the "not" word, and it will become black (active). To remove "not" from the condition, just click it again, and it will be greyed out. The next drop-down allows to select the condition from one of the following:

has data - if the item being searched contains any data

equals - opens one more box where you can enter what the item should be equal to

is empty - if the item being searched is empty

in range - opens two boxes where you enter the boundaries of the range where the item should be

<= - if the item is less or equal to the value in the following box

>= - if the item is greater or equal to the value in the following box

At the end of the line, a small "x" allows to delete the whole rule by clicking on it.

If you click **+** once again, another rule is added, and between them the **and** operator is displayed by default, meaning that to satisfy the filter, both rules should apply. You can change it to **or** by clicking on it, so that to satisfy the filter, one of rules being true may be enough. Clicking **or** once again turns it to **and** again.

You can add as many rules as you wish, and the operators **and / or** between them will follow the general priority logic of boolean operations, i.e. **and** has the priority over **or**, as in any program code.

To change the priority and build more complex logical rules, you can click **(** making a separate set of rules (in parentheses), which have higher priority, as in mathematical operations. Note than the **and / or** operator which appears before the parentheses depends of the previous operator selected, i.e. if it was **or**, the next one will also be **or**, and vice versa. You can always change the operators by clicking on them.

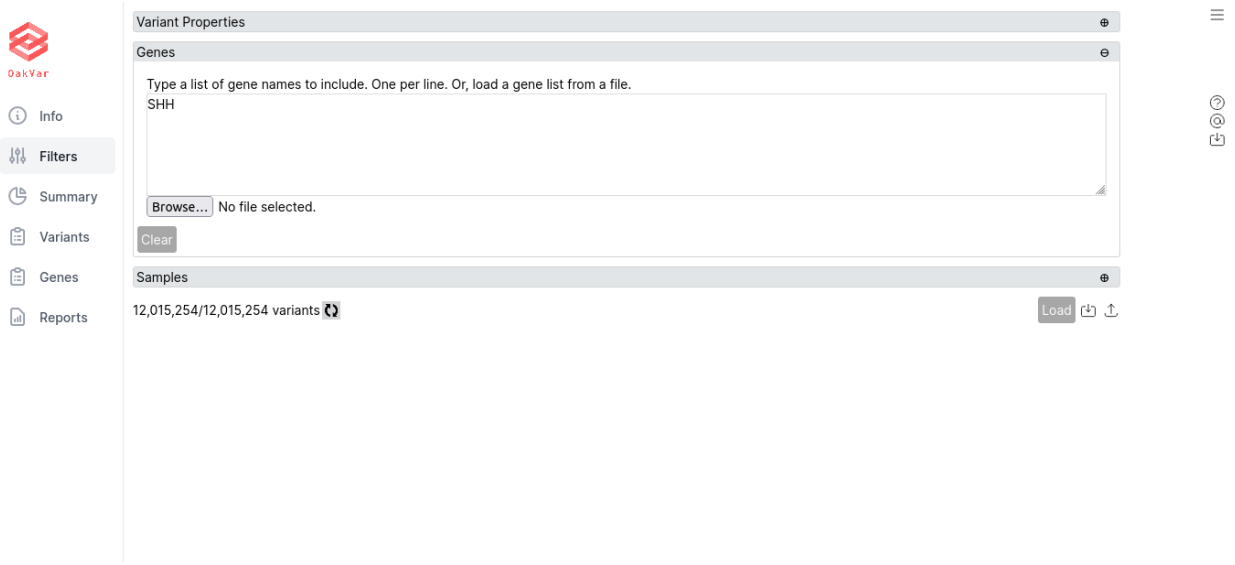
Within the parentheses, you can create any number of rules, and there are separate **+** and **(** buttons to add new rules and nested parentheses inside the parentheses. Also in the upper left corner a separate **NOT** switch appears if you hover the mouse over it.

You can also move any rule to another rule. To do this, drag an anchor **||** which appears from the left side of the rule if you hover the mouse there, and drop it on any rounded **+** anchors which appear between rules and/or parentheses (not on the **+** button that adds rules).

BUG NOTE: If dropping a rule just before or after itself, it redirects browser to an error page. In this case all the previously made filter settings may not be saved. Please avoid dropping a rule before or after itself until this bug is fixed in Oakvar.

1.3.4 Filtering by Genes

Switch to the **Genes** section and enter any particular gene names, one per line. Also you can load them from a file by clicking **Browse...**



1.3.5 Clearing Filters

Under any section you can click the **Clear** button to remove any filter settings from that section.

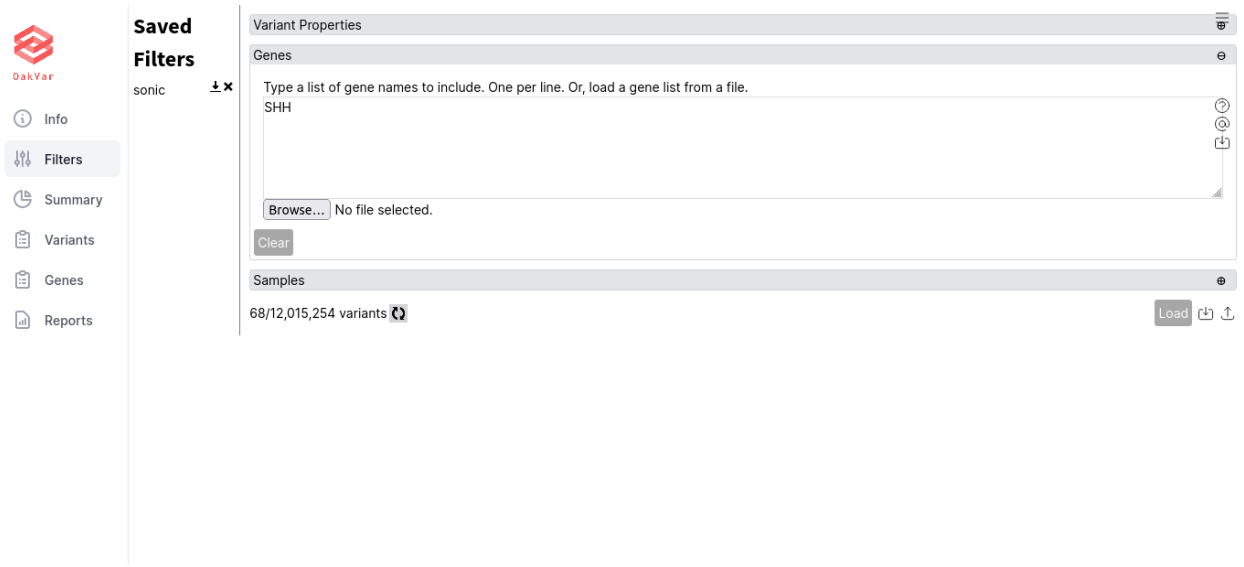
1.3.6 Saving and Importing Filters

You can save the filter (the whole set of rules) in OakVar for further loading, as well as exporting to a file, or import it from a file.

To save the filter, click the middle button ("inward arrow") in the lower right corner of the page, and enter the filter name.

NOTE: Filters are saved internally in OakVar, i.e. on the server if using a remote installation. To have a filter saved into a local file, export it after saving.

The saved filter appears in the left part of the page in the **Saved Filters** list:



To load a saved filter, just click its name. To export a saved filter into a file, click the icon with a down arrow next to its name. To delete a saved filter, click the **X** icon in its line.

To import a filter from a file, click the "up arrow" (rightmost) button in the lower right corner of the page, and browse for a file to import (e.g. pathogenic.json). Clicking **Open** in the browse window loads the filter. **NOTE:** the filter is not saved automatically, you need to save it using the "Save filter" (inward arrow) button if you want to keep it on the server for further working.

1.3.7 Loading Filtered Variants

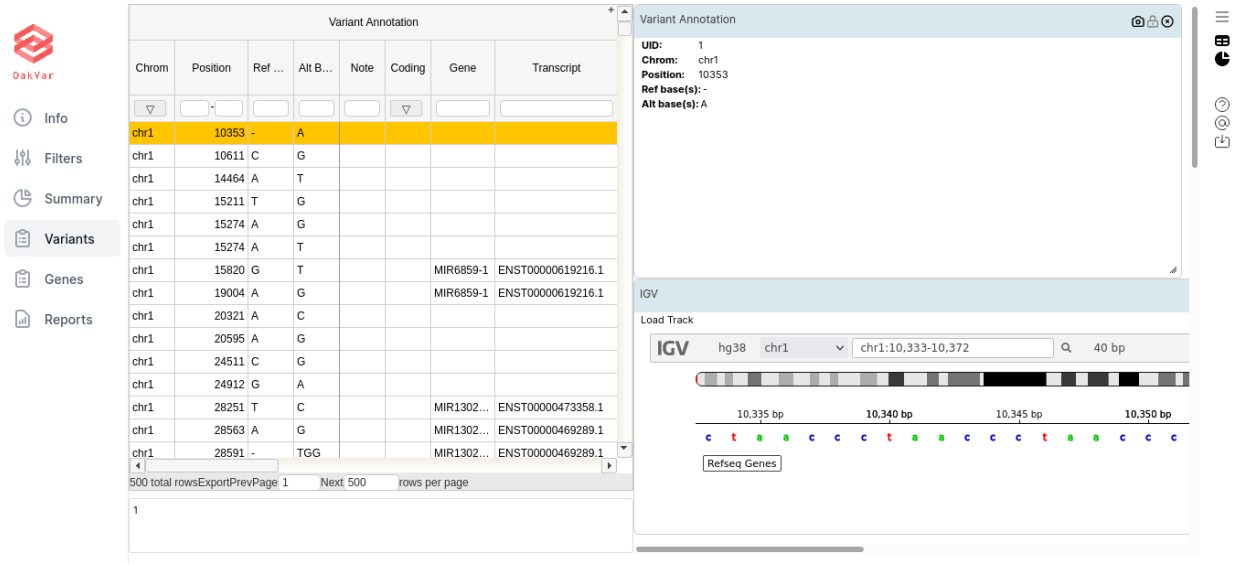
When building a filter, you can click the refresh button next to the number of variants (e.g. **68/12,015,254 variants**) in the lower left corner of the page to check how many results the filter provides. If the number is small enough, when the filter is ready, click **Load** in the lower right corner of the page. After loading the filter, the number of variants in the lower left corner (the first number before the slash, while the second one is the total number of variants and doesn't change) may be updated.

When the filtered variants are loaded, you can proceed to the **Variant** tab to analyse them (see the next section).

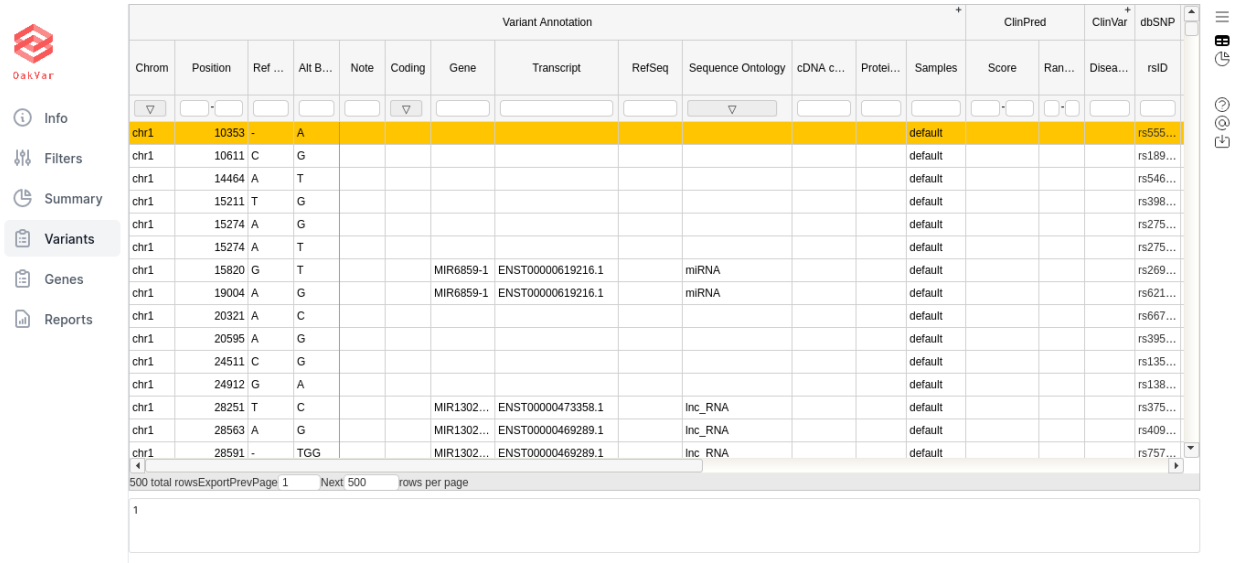
1.4 Working With Annotated Variants

After applying the necessary filters, select the **Variants** page.

By default a combined view is displayed, with both table and widgets:



By clicking the icons in the upper right corner, you can toggle on/off the table view (window-like icon) and the widgets view (piechart-like icon). For our purposes first of all we need the table view:



The table contains columns and column sets with general information about the filtered variants, as well as those connected to certain annotators. Some logically grouped column sets (by a particular annotator) can be extended or collapsed by clicking the **+/-** sign in the upper right corner of the column set (the topmost row). If you filtered by particular annotators, especially using "has data" condition, for other annotators it may show nothing for that particular variants, and they can be collapsed for convenience.

Each row of the table represents a variant that you can research.

The most important column groups for us are listed below, along with columns:

1.4.1 Variant Annotation

UID - the variant number in this (filtered) sequence

Chrom - chromosome where the variant is located. Chromosome names are 'chr1' to 'chr22', 'chrX', 'chrY' and 'chrM'.

Position - chromosomal position of the variant. The first position in each chromosome is position 1.

Ref Base - reference allele at this chromosomal position (one of A, C, G, T, and N).

Alt Base - alternative allele; called based on reads mapping to this chromosomal position.

Note - note for the variant, if available.

Coding - whether this gene variant is coding.

Gene - the gene this variant belongs to.

Transcript - GENCODE transcript.

RefSeq - the reference sequence.

Sequence Ontology - could be: missense variant, start lost, stop gained, or stop lost.

cDNA change - change of coding DNA.

Protein Change - change of protein being synthesized.

All Mappings - expression showing all the mappings.

Sample Count - the number of samples which contain the variant.

Samples - samples which contain the variant.

Tags - variant tags from the input file.

1.4.2 ClinVar

Clinical Significance - the level of clinical significance of the variant.

Disease Ref Nums - disease reference numbers.

Disease Names - names of diseases associated with the variant.

Review Status - the level of review supporting clinical significance.

ClinVar ID - ID in the ClinVar database.

Significance Detail - additional detail on clinical significance used when it is conflicting.

1.4.3 dbSNP

rsID - the database identifier ("rs" number) of this variant in dbSNP.

This column is empty if the observed variant is not described in dbSNP. Such variants can be extremely rare variants or technical artifacts.

1.4.4 LongevityMap Annotator

LongevityMap ID - ID(s) of the variant in LongevityMap.

Significance - could be: significant, non-significant, or conflicted.

Source Population - population the data were obtained on, e. g. Danish or American (Caucasian).

dbSNP id - ID of the variant in dbSNP.

Associated Genes - genes associated with the variant.

PubMed ID - ID of the variant in PubMed.

Info - additional information.

Description - detailed description of the research.

allele - allele associated with the variant.

state - state of the variant.

zygosity - zygosity of the variant vs OV.

weight - weight of the variant.

priority - priority of the variant.

1.4.5 VCF Info

Phred - Phred quality score.

VCF Filter - if the VCF filter is passed (PASS).

Zygosity - most likely zygosity of the variant in this chromosomal position, computed from the observed variant frequency (column 8). Can be "FP/HET" (<15%), "HET" (15-75%), "HET/HOM" (75-85%), or "HOM" (>85%).

Alternate reads - the number of reads showing the alternative allele.

Total reads - the total number of reads.

Variant AF - the variant allele frequency.

Haplotype block ID - ID of the haplotype block.

Haplotype strand ID - ID of the haplotype strand.

1.5 Developer guide

NOTE: *Both OpenCravat and OakVar can be used to annotate a human genome. At the beginning of the project, we used OpenCravat as a framework. However, as OakVar is based on OpenCravat and contains more advanced features customized specially for personal longevity genomics, we decided to base further development of the project on OakVar.*

1.5.1 Installing OakVar

Since our module is based on OakVar, you have to install OakVar first to run our module, if it is not already installed. OakVar docs: https://oakvar.readthedocs.io/en/latest/

Pre-requirements for Oakvar:

- installed conda/mamba environment management systems, or you can use their lighter versions: miniconda/micromamba
- installed python and pip

You can find documentation for mamba here: https://mamba.readthedocs.io/en/latest/index.html

And for conda here: https://docs.conda.io/en/latest/

The installation of OakVar and further work should proceed after the activation of an environment created by Conda/Mamba or Miniconda/Micromamba.

1.5.2 Installing Annotators

For the Longevity module to work, you need to install the following annotators:

1. clinvar
2. dbsnp
3. gnomad
4. ncbigene
5. omim
6. pubmed

You can install them by using terminal or Oakvar GUI.

Installation using terminal:

Use the following command:

`ov module install module_name`

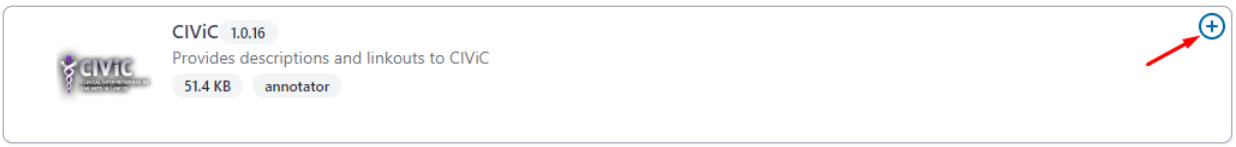
Installation using GUI:

To activate Oakvar GUI, use the following command:

`ov gui` or `ov gui --multiuser --debug` to open GUI in multiuser debug version.

After the execution, GUI will be opened in your browser.

Go to “Store” and find the annotators and install them:



1.5.3 Installing Postaggregators

For the Longevity module to work, you need to install the following postaggregators:

1. just_prs
2. just_longevitymap
3. just_drugs
4. just_cancer
5. just_coronary

You can install them by using terminal or Oakvar GUI.

Installation using terminal:

Use the following command:

`ov module install module_name`

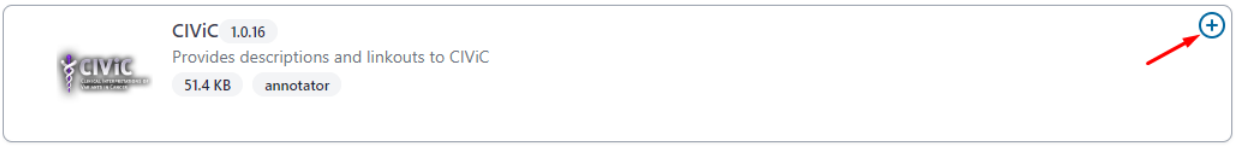
Installation using GUI:

To activate Oakvar GUI, use the following command:

`ov gui` or `ov gui --multiuser --debug` to open GUI in multiuser debug version.

After the execution, GUI will be opened in your browser.

Go to "Store" and find the annotators and install them:



1.5.4 Installing the Reporter

Installation using terminal:

Use the following command in terminal:

`ov module install longevity2reporter`

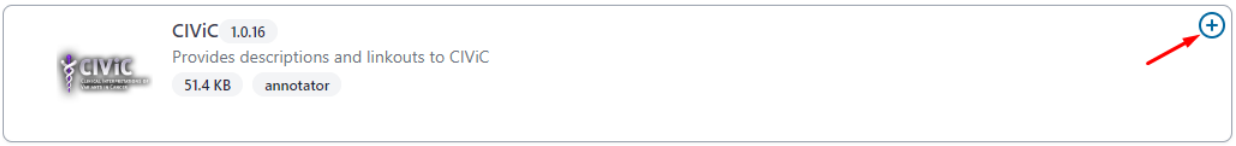
Installation using GUI:

To activate Oakvar GUI, use the following command:

`ov gui` or `ov gui --multiuser --debug` to open GUI in multiuser debug version.

After the execution, GUI will be opened in your browser.

Go to "Store" and find the annotators and install them:



All further work also can be done by using command-line interface, for more information check OakVar documentation: https://rkimoakbioinformatics.github.io/oakvar/cli/

Or by using GUI *Getting Started*

1.5.5 The description of some modules

All annotators can be divided into 2 groups:

1) Tools that predict pathogenicity (**bold**)
2) Tools that provide information like databases

Here are their internal (coded) module names:

- **cadd_exome** (1.6.1) - CADD is a tool for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome
- gnomad_gene (2.2.1) - gene-level population statistics from gnomAD
- pubmed (1.1.5) - articles related to a particular gene
- **clingen** (1.0.1) - NIH-funded resource that defines the clinical relevance of genes and variants
- **clinpred** (1.0.0) - prediction tool to identify disease-relevant nonsynonymous single nucleotide variants
- clinvar (2021.10.01) - ClinVar is an archive of reports of the relationships among human variations and phenotypes, as well as interpretations of clinically relevant variants (Uncertain significance, Likely pathogenic, Pathogenic etc.)
- mitomap (1.1.0) - a human mitochondrial genome database
- ncbigene (2019.08.02) - gene descriptions from NCBI (National Center for Biotechnology Information) Gene database
- omim (1.0.0) - a catalog of human genes and genetic disorders and traits
- **prec** (3.6.0) - provides a database identifying rare and likely deleterious loss-of-function (LoF) alleles
- **provean** (1.0.0) - a tool which predicts whether an amino acid substitution or indel has an impact on the biological function of a protein
- **revel** (2020.12.02) - ensemble method for predicting the pathogenicity of missense variants based on a combination of scores from 13 individual tools
- **sift** (1.2.0) - predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids
- GnomADD - aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects
- PharmGKB - an NIH-funded resource that provides information about how human genetic variation affects the response to medications
- dbSNP - the Single Nucleotide Polymorphism Database is a free public archive for genetic variation within and across different species developed and hosted by the National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI)