Instructions for Use

Refers to VarSome Clinical v11.9

Table of Contents

Explanation of symbols

	Catalog Number
	Manufacturer
	Consult Instruction for Use
	The device complies with European Directive 98/79/EEC
	When followed by "**CAUTION**", this symbol means "Caution! Failure to observe could result in forfeited time or effort, or the need to abort use of the device."

Introduction

This document provides general information and procedure to use VarSome Clinical, available by using the following url:

https://clinical.varsome.com/

⚠ **CAUTION** This Instruction For Use (IFU) contains important information for the safe use of this product. Please read the entire IFU before using our product.

Manufacturer:

Saphetor SA

EPFL Innovation Park – C 1015 Lausanne
+41 21 552 09 77

www.saphetor.com

Authorized Representative

Saphetor SA Greek Branch

Leocharous 3 - Athens 105 60
Greece

Result Validation

We can provide detailed validation results for known individuals, whose cells/DNA can be ordered by post:

https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM12878.

The samples should be run through normal laboratory procedures. Then send us the FASTQ files and we will validate the variants against the Genome in a Bottle (GIAB) dataset giving detailed specificity and sensitivity.

Regulatory Information

According to the Annex III of the European Directive 98/79/EEC, VarSome Clinical is considered as an In-Vitro Medical device.

The use of VarSome Clinical is subject to the terms and conditions provided on our website. Please consult them.

Device Description

VarSome Clinical is an online software application that can process the results of Next Generation Sequencing (NGS), generating genetic variation information based on standard guidelines and databases with clinical evidence.

VarSome Clinical is composed of two main modules, with module 2 being subdivided in two submodules:

1) **Module 1: Variant Calling**

   VarSome Clinical utilizes variant calling pipelines for analysis of Next Generation Sequencing data (Illumina or MGI technologies), from whole genome assays, targeted DNA capture-kit assays (e.g. exome, clinical exome) and amplicon kits for germline and somatic DNA.

   Variant calling is a computational method used in genetics to detect differences (variants) in DNA sequences and involves comparing an individual's DNA to a reference genome.

2) **Module 2.1: Annotation**

   VarSome Clinical will annotate all variants identified by bringing in known information about them from external databases. This will include, but is not limited to the gene(s) the variants fall in, the known functions of those genes, any publications mentioning the genes or the variant in their title or abstract, any reported clinical information about the variant (e.g. from ClinVar).

3) **Module 2.2: Variant Classification**

   VarSome Clinical will then use the information from the Annotation step above to identify which germline or somatic variant classification rules are applicable to a variant and so reach a variant pathogenicity verdict. VarSome Clinical uses SVC v3 and AMP rules, which are guidelines considered standard in the field for determining pathogenicity for germline or somatic variants, respectively, issued by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (Richards et al., 2015, Riggs et al., 2020, and Li et al., 2017).

VarSome Clinical provides the user with tools that allow filtering the results according to user-defined criteria and can also propose algorithmic filters for more sophisticated filtering such as identifying de-novo variants in family trio analyses or finding cases of compound heterozygosity.

Principle of operations

VarSome Clinical is accessible using internet browsers including Chrome, Safari or Edge and it is installed either on Saphetor's servers, or in Google Cloud or on the client's premises. The system consists of a number of software tools and databases installed on the servers which are

accessed by the users via a Graphical User Interface (GUI) provided by the VarSome Clinical Web Portal. The software tools are either developed in-house or obtained from third parties.

The scheme reported in Figure 1 is a summary of the analysis pipelines used in VarSome Clinical. VarSome Clinical is intended to be integrated within a routine NGS genetic testing workflow, which begins when a patient initially visits a specialized doctor, and the doctor suspects a genetic disease based on the clinical examination of the patient, or the patient suffers from cancer. The genetic testing workflow is carried out according to the following steps, which include the primary, secondary and tertiary analysis of NGS data (only the secondary and tertiary analysis are performed by VarSome Clinical, therefore the primary analysis is out of scope of this Performance Evaluation):

--- Previous to VarSome Clinical and out of scope of the current Performance Evaluation---

Step 1 – Once the patient is prescribed with a genetic test for the identification of DNA variants in one or multiple candidate genes, the laboratory takes a biological sample from the patient (blood is the most common).

Step 2 – DNA extraction is performed.

Step 3 – The extracted DNA is fragmented and prepared in a sequencing library.

Step 4 - During the **primary analysis**, millions of DNA fragments are sequenced using the NGS machines. The output of this step will be one or more text files ("fastq" files) with the sequence of DNA bases (A, C, T or G) as detected in the patient's sample along with the quality score for each sequenced DNA base. NGS sequencing results in millions of such small sequences ("reads"), many of which will correspond to the same areas of the sample's genome.

Of note, steps 1 to 4 are independent from VarSome Clinical. However, to guarantee a good quality of sequencing data for a clinical setting, steps 1 to 4 shall be performed with CE-marked medical devices.

--- VarSome Clinical ---

Step 5 - VarSome Clinical platform use starts. A specific interface which is available for uploading FASTQ or VCF files (see Section 2.8.3. for the description of the input files) is selected. The user simply needs to select the files to be uploaded. Once the files are uploaded, the user can select the files to be analyzed and launch the test.

Step 6 - During the **secondary analysis**, VarSome Clinical performs an analysis of the DNA reads. This analysis step involves aligning the reads to the reference genome and calling variants, in the case of starting from FASTQ files. Alternatively, the physician may choose to use a different service to perform the variant calling, in which case they would upload a VCF file containing the identified variants to VarSome Clinical and start using VarSome Clinical at Step 7.

Step 7 - During the **tertiary analysis**, VarSome Clinical annotates and classifies the identified or provided variants, that can easily be in the order of tens of thousands. First, it annotates the variants based on their possible effects on genes (e.g. truncation of protein, amino acid change, etc) and information gathered from external databases, such as relevant publications, evidence on their effect of a specific phenotype, clinical trials, etc. VarSome Clinical uses information from over 140 databases to provide the most comprehensive set of annotations. These data from external databases are first gathered and stored in Saphetor's server to create an integrated database (Molecular DB). You may find the list of databases we have integrated in VarSome on the following link https://varsome.com/datasources/.

Secondly, based on the annotated information on the variants, VarSome Clinical follows a set of rules and guidelines to classify the variants with respect to their pathogenicity (i.e., their likelihood of affecting the physiological state of an individual, which can translate to disease in some cases). For germline variants, the SVC v3 guidelines are followed, and variants are classified as either "Benign", "Likely Benign", "Likely Pathogenic", "Pathogenic" or "VUS" (Variant of Unknown Significance). For somatic variants, the AMP guidelines are followed, and variants are classified as Tier 1, Tier 2, Tier 3 and Tier 4 according to their implication in cancer (from clear to little evidence, respectively). Of note, CNVs are classified based on specific CNV guidelines (Riggs *et al.,* 2020).

Step 8 (optional) - Once the variant calling and/or the variant annotation/classification has finished, an email is sent to the user, who can then access the results through the user interface. Once the sample analysis is finished, the user can access the detected genetic variants and the quality reports associated with the sample and the analysis. At this point, the user can manually adjust each of the classification criteria based on prior information (such as a variant's de novo status, patient's and their family's health background and phenotypes, etc.) or their own domain knowledge to reach a final interpretation. To filter out the irrelevant variants, the user needs to design the filters that will be applied to the variant table. The system provides a built-in editor which can be used to generate a report .

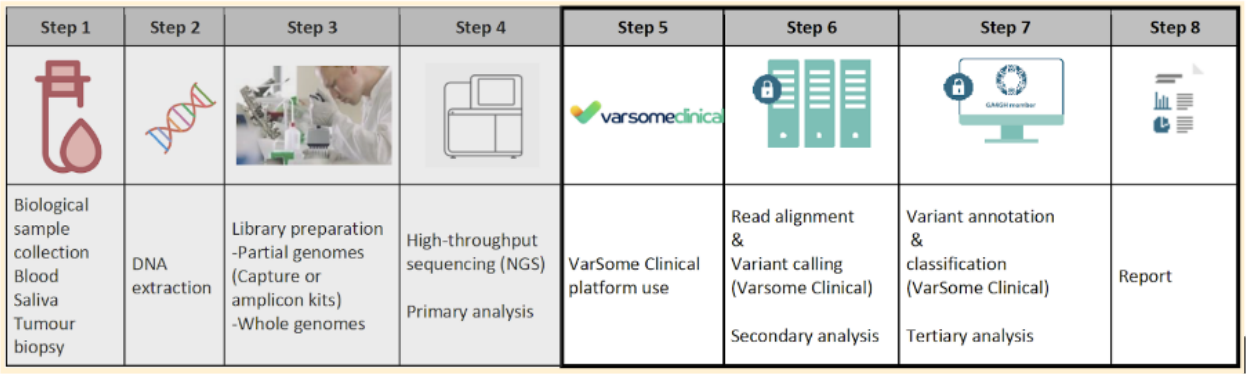


Figure 1. Genetic test laboratory workflow. Steps 5, 6, 7 and 8 are supported by VarSome Clinical, which include the secondary and tertiary analyses, and report generation (optional).

Intended Use

The VarSome Clinical platform aids in making an informed decision and diagnosis, and in the selection of appropriate treatments for cancer, Mendelian and other genetic disorders by processing next generation DNA sequencing data.

VarSome Clinical utilizes variant calling pipelines for analysis of Next Generation Sequencing data (Illumina technologies), from whole genome libraries, DNA capture-kit libraries (exome, clinical exome) and amplicon kits for germline and somatic DNA. The Clinical platform identifies and reports SNVs, CNVs, indels and substitutions, as well as variant annotation to facilitate the variant curation and classification. Multi-sample analyses are also available, to perform joint-genotyping to generate a multi-sample VCF file with the merged genotypes for two or more samples. VarSome Clinical can also propose algorithmic filters for inheritance hypothesis testing.

Indication for Use

VarSome Clinical offers clinicians and researchers a solution to assist in making an informed decision and diagnosis on the selection of the best therapies for the treatment of patients of all ages for the following conditions:

- Patients suspected to be affected by heterogeneous mendelian diseases or undiagnosed (rare) diseases.
- Patients with common diseases that are in fact a group of rare diseases with eclectic phenotypic features, being caused by mutations in specific genes, such as developmental delay or intellectual disability, degenerative neurological diseases, immunodeficiency diseases, cardiomyopathies, etc.
- Patients with hereditary cancer syndromes.

Warning, caution and precautions

⚠ **CAUTION** When using NGS to investigate the cause of a disease, clinicians and patients must be aware that the test may cover many genes, most of them unrelated to the patient disease. This would lead to discovery of secondary or incidental findings on genes causing other diseases, e.g. cancer predisposition status (Kuhlen and Borkhardt 2015). Recently, policy statements from two major regulatory agencies (AMP and ACMG) have provided a list of recommendations (Green et al. 2013; Allyse and Michie 2013; ACMG Board of Directors 2015; Hegde et al. 2015). However, each country has its own directives about incidental finding reporting. Therefore, it is extremely important that the patient, or their legal representative, is informed of the incurred risks and that the patient's decision is recorded in an “informed consent” document (Anderson et al. 2016). It is the responsibility of the treating clinician to collect the informed consent form before ordering any NGS test, and ensure that the patient understands the potential outcomes of the test.

Users should be warned that the result reported shall not be used alone to modify the patient care.
Because certain chemotherapeutic agents can cause DNA damage, patients undergoing chemotherapy treatment may have variants created by the chemotherapy process itself. Such variants can be included in the results of VarSome Clinical.

There is a risk of misclassification of variants recently discovered and not documented enough in the scientific literature.
The DNA sequencer used to generate these data shall be CE-marked to guarantee the minimum standards of data quality.

The following warnings are sub-divided according to the VarSome Clinical modules: Variant calling and Variant annotation/classification.

Variant-Calling

- Large-scale genomic rearrangement analysis (structural variants, including CNVs, whole-chromosome deletions etc.) may not be guaranteed because such genomic events cannot be reliably detected from short read next generation sequencing data. See, for example, Tan R. *et al., Hum Mutat*. 2014 Jul;35(**7**):899-907. doi: 10.1002/humu.22537 and references therein). This is a limitation of NGS sequencing itself, caused by the short (~150-200bp) length of the reads, and not of the VarSome Clinical device.
- Performance was not validated specifically for hard to detect variants (those falling in regions of segmental duplication or other hard to sequence and/or process regions of the genome) because the system does not provide special tools for such variants.
- The performance of VarSome Clinical cannot be guaranteed for samples with very low concentrations of tumor cells.

Variant Annotation & Classification
VarSome Clinical implements three automated classifiers according to the ACMG & AMP

guidelines.
The classification results obtained with VarSome Clinical aim to reflect common expert consensus. The user shall be aware that the results generated after changing the classification settings are not part of the in vitro medical device scope and are therefore not validated.
The classifiers implement an automated evaluation of the **generic** guidelines for ACMG, AMP and CNVs. These automated classifications cannot substitute for the knowledge, experience and additional case information available to a human curator and are limited by the availability of machine-legible data.
Only the "generic" ACMG classification rules have been implemented, specifically, **disease-specific guidelines have not been implemented**. These typically require further adjustments to the various classifiers, rules, thresholds, exception lists etc. However, all the ClinGen expert panel reviewed variants are available in VarSome, visible in the ClinGen tabs (these entries are also replicated by ClinVar).
Please see the online documentation for the full detail of how the classifiers have been implemented in VarSome:

- **ACMG - germline:** https://varsome.com/about/resources/germline-implementation/
- **AMP - somatic:** https://varsome.com/about/resources/somatic-implementation/
- **CNVs:** https://varsome.com/about/resources/sv-implementation/

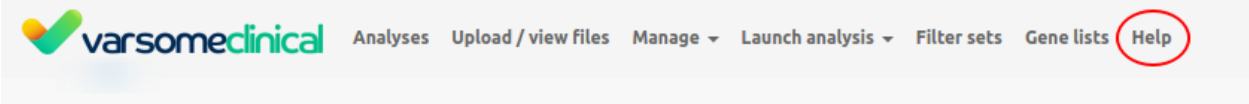
⚠️ Please note that a visible discrepancy between the two platforms may be encountered since VarSome is updated prior to VarSome Clinical. There are two reasons why a variant may have a different pathogenicity annotation on VarSome Clinical and VarSome: updates in the databases utilized to infer the pathogenicity and improvements in VarSome's Germline Variant Classification..

Introduction to VarSome Clinical

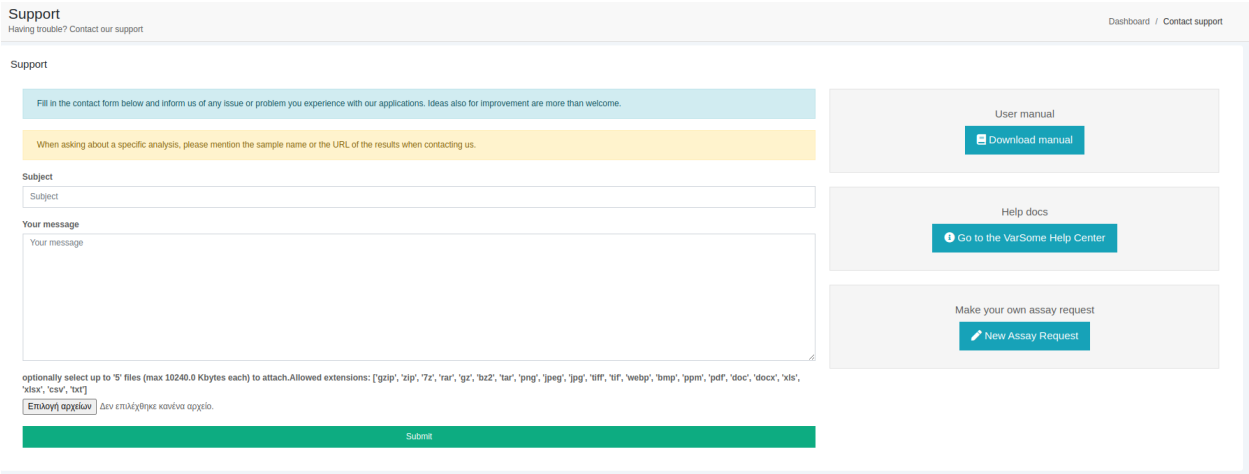
1. Getting help from support and citation

Get help via VarSome Clinical

In case any help is needed either to understand how VarSome Clinical works or if you are facing a problem, it is possible to contact our support directly from VarSome Clinical from the Help menu option.



A new window will be displayed with the subject and main text field where you can describe your problem or request help. It also allows you to attach a document. Our support team will reply to you via email.



Other methods

It is possible to reach out the support team directly by email (support@varsome.com) or by filling out the help center form (https://docs.varsome.com/en/kb-tickets/new).
Please provide detailed information on the question or the problem you are facing as well as screenshots that may help us identify and reproduce the issue.

How to cite

If you use VarSome or VarSome Clinical for your work, please cite our paper in your publications and other communications:

VarSome: the human genomic variant search engine, Christos Kopanos, Vasilis Tsiolkas,

Alexandros Kouris, Charles E Chapple, Monica Albarca Aguilera, Richard Meyer, Andreas Massouras, Bioinformatics, Volume 35, Issue 11, 1 June 2019, Pages 1978–1980, https://doi.org/10.1093/bioinformatics/bty897

You can find VarSome citations on the following link: https://varsome.com/about/general/varsome-citations/.

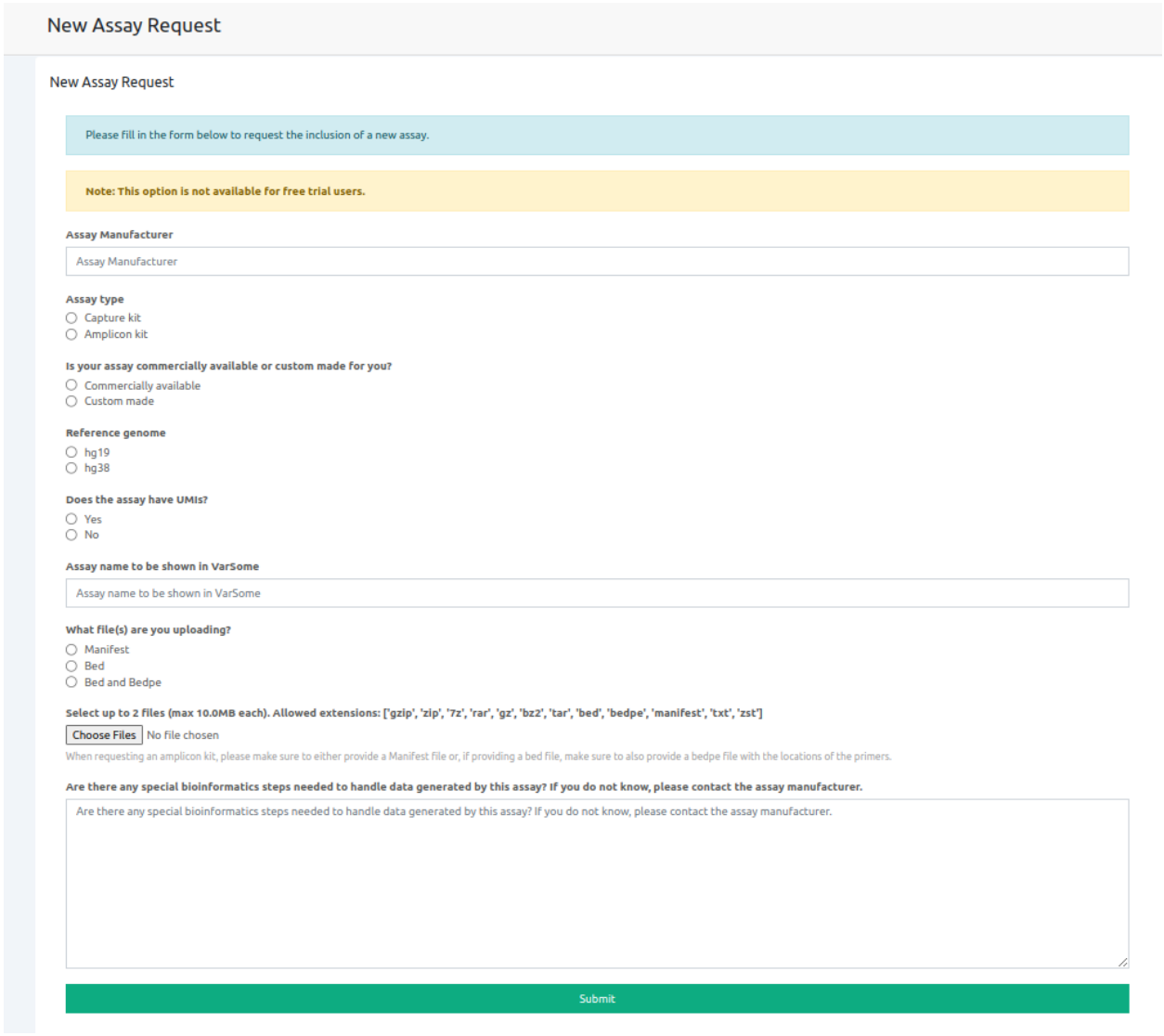
2. Adding an assay to VarSome Clinical

VarSome Clinical customers

If you are a VarSome Clinical customer and your assay is not yet available in VarSome Clinical, we can add your assay if it is an assay we can support.

If you need us to add a new assay, the first question is whether this is a standard assay or if it requires special handling. Examples of special handling include but are not limited to assays with Unique Molecular Identifiers (UMIs) (of those, VarSome Clinical currently supports only Paragon CleanPlex UMI Lung Cancer Panel), assays that are designed to produce mixed RNA and DNA sequencing data, assays targeting specific classes of variants such as certain large deletions in genes like CFTR and various other special cases. If you do not know if an assay requires any special treatment, please talk to your assay providers and ask them if any special bioinformatics approaches are required to handle the data the assay will be used to generate.

You can now request to add your assay of preference by filling the corresponding form which is available in the following links depending on the installation you are working from:
https://ch.clinical.varsome.com/new-assay-request/
https://eu.clinical.varsome.com/new-assay-request/
https://us.clinical.varsome.com/new-assay-request/.



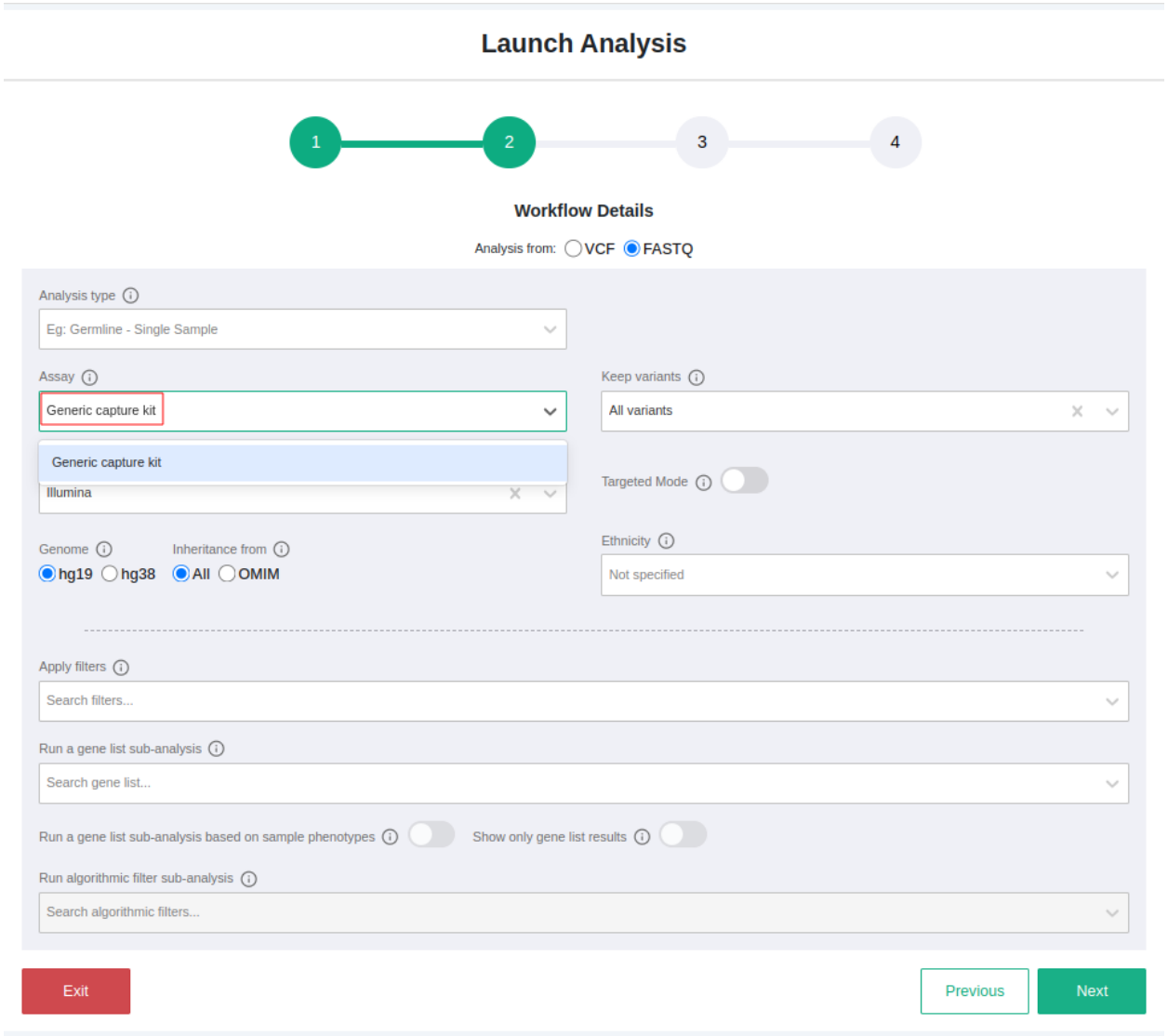
On this form you will need to fill the requested fields:

- The assay's manufacturer name
- Whether it is a capture or amplicon-based assay
- Whether it is a commercially available or custom made assay
- The version of reference genome (hg19 or hg38)
- If the assay includes Unique Molecular Identifiers (UMIs)
- The name of the assay to be shown on Varsome Clinical
- The type of files you are uploading (manifest, .bed or .bed and .bedpe)
- Choose the corresponding files
- Specify if any special bioinformatics approaches are required to handle the data the assay will be used to generate.

Please keep in mind that for assays targeting relatively small regions of the genome (i.e. assays that are neither whole-exome nor whole genome) variant calling will be performed in [targeted mode], only looking at the regions mentioned in the assay.

⚠️ If you ask us to add a new assay targeting mitochondria on hg19/GRCh37 reference genome, please use the rCRS “MT” coordinates in the accompanying BED file.

If your assay is not listed, you can choose the "Generic capture kit", for the case of capture-based libraries.



As long as you choose the right chemistry type as described above, the results will be the same as if we had added your assay. The only differences will be that we will not give you statistics about the percentage of reads that fall on target and we will not run the variant calling in targeted mode (since we do not know the targets) so you may see variants that fall outside your regions of interest. However, the results will be correct and reliable.

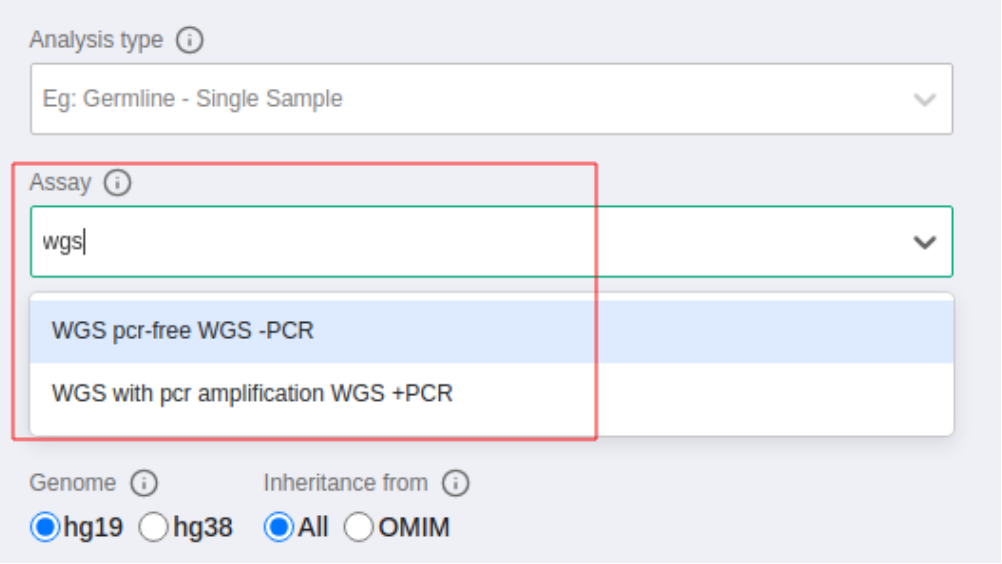
The correct choice of the chemistry type is important because it affects:

- The way we handle possible PCR duplicates and
- The default thresholds set for minimum coverage needed to call a variant. For samples sequenced using an amplicon kit, since such kits tend to have very high coverage, the minimum threshold coverage is higher than for the capture kits.

⚠️ Please, note that the samples analyzed using one of the "Generic" kits can not be used for CNV sub-analyses.

The assay's information will be used to calculate alignment statistics and coverage of the targeted regions. The statistics are used to generate the QC report.

If you want to run a whole genome analysis you need to select the 'WGS' option when launching the analysis, as shown on the pic below. This option will guarantee the sample will be available for subsequent CNV analysis, if you wish to perform it.



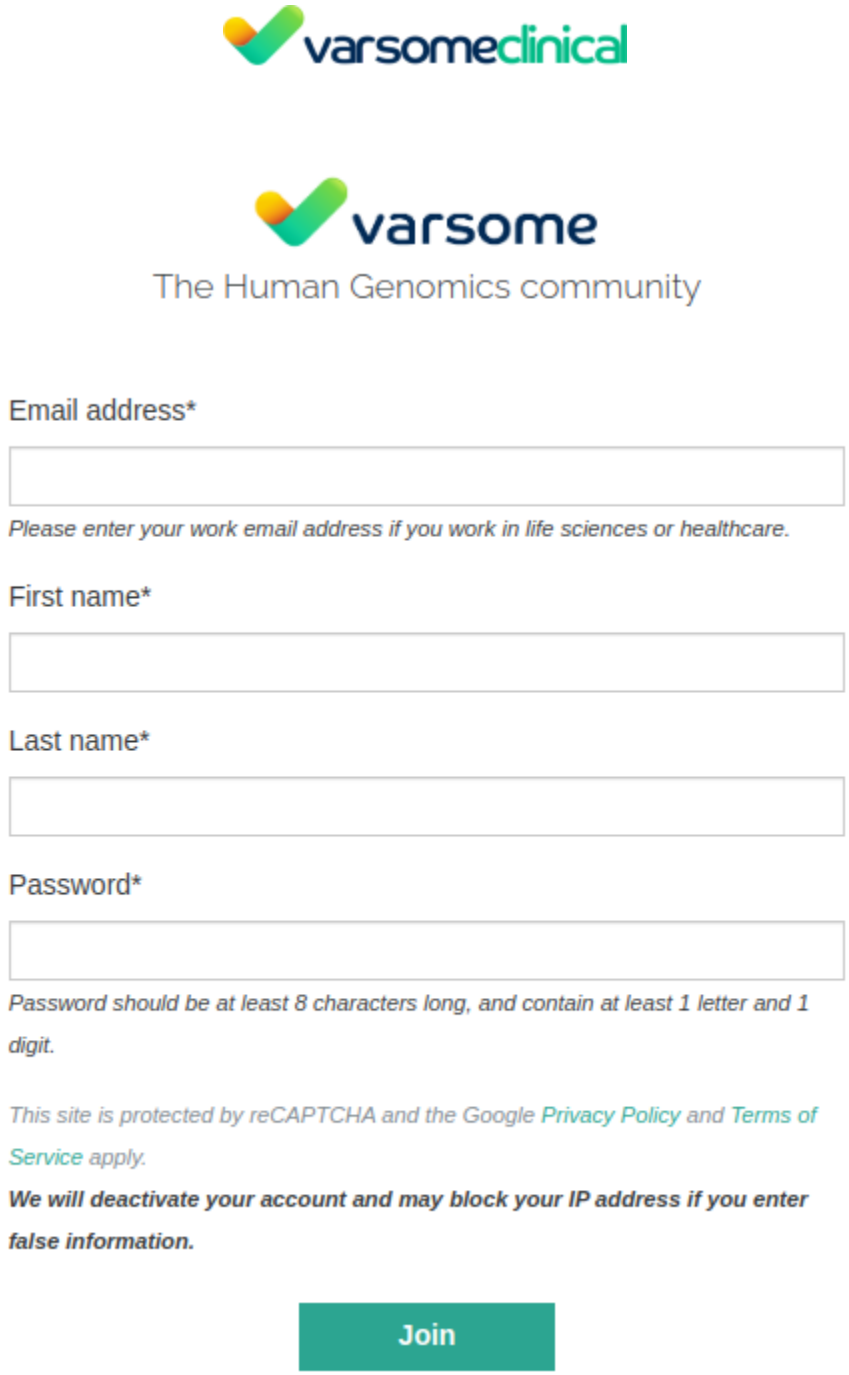
⚠️ Please note that as VarSome Clinical can process any kind of NGS data, the list of Assay is very long. To eliminate the need to scroll through the long list of Assay whenever you are launching an analysis, you or your supervisor can simply mark certain assays as favorite. Remember to check your settings for favorite assays.

3. Installation

VarSome Clinical is a web based software that connects to our servers to run all the analyses. It can be used from different browsers such as Google Chrome, Firefox, Microsoft Edge and Safari.
In order to have access to VarSome Clinical a subscription is required. If you want to get a trial or acquire a subscription please contact sales@varsome.com.
VarSome Clinical is available from 3 different servers, 2 from Google Cloud, one located in Germany (EU) and the other in the United States (US). The last one is on a server that is located in Switzerland (CH).

Registration/ Join VarSome

In order to be able to have access to VarSome Clinical either with a subscription or a trial you have to register to https://varsome.com/ first.
On the top right click on Join and a new window will be displayed where you can provide your contact details, Email address, First name and Last name.



An email will be sent confirming the registration. Now you can request your VarSome trial / subscription to our business team.

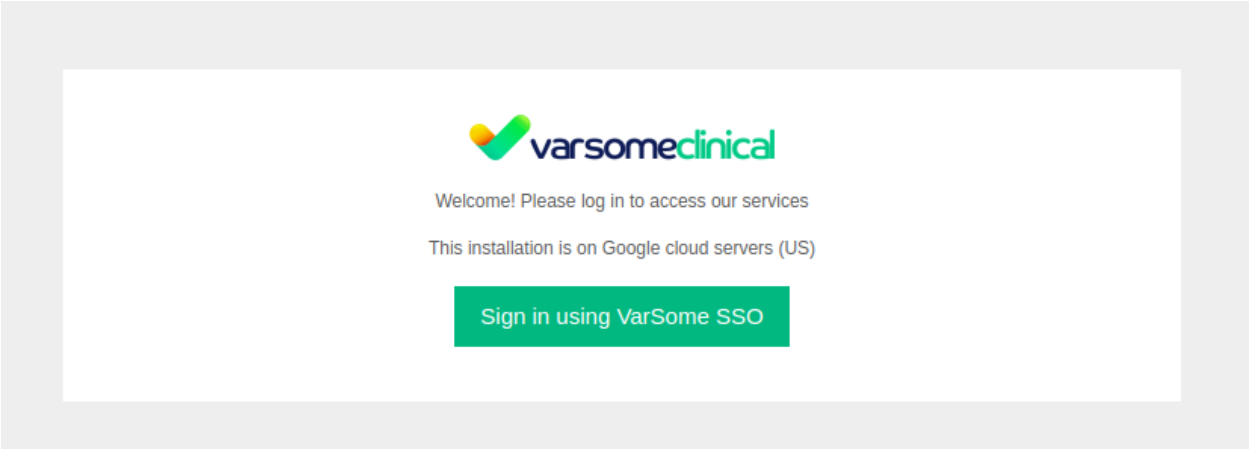
Login to VarSome Clinical with 2 Factor Authentication

Once someone from our account team confirms that your trial / subscription has been activated then it is possible to log in to VarSome Clinical. The activation of your VarSome Clinical could be either on the CH, EU or US server.

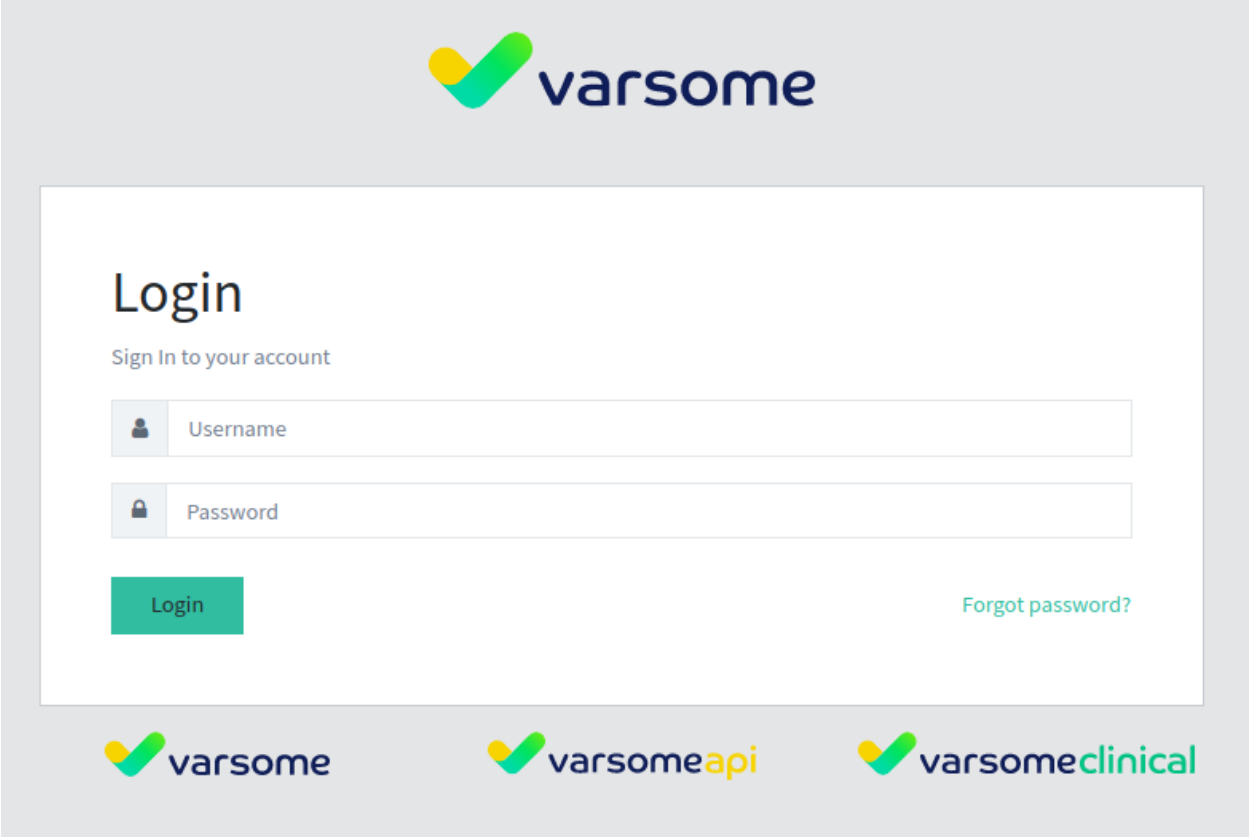
You will only be able to log in to one of the servers:

https://ch.clinical.varsome.com/
https://eu.clinical.varsome.com/
https://us.clinical.varsome.com/

Once you open the VarSome Clinical web page you will see the following sign in information.



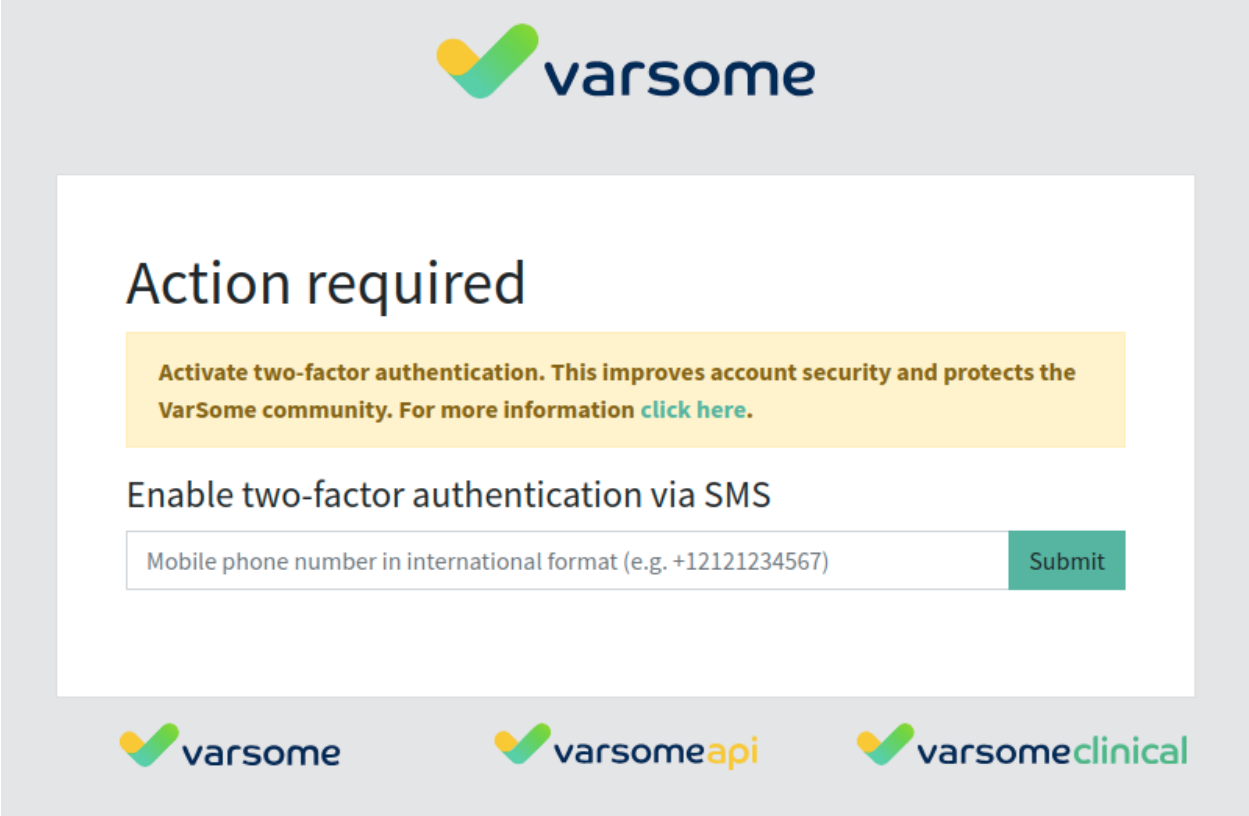
By clicking on ***Sign in using VarSome SSO*** you will be redirected to our SSO (Single sign-on) page and you can now provide your user name (email address) and password which you used to register to varsome.com.



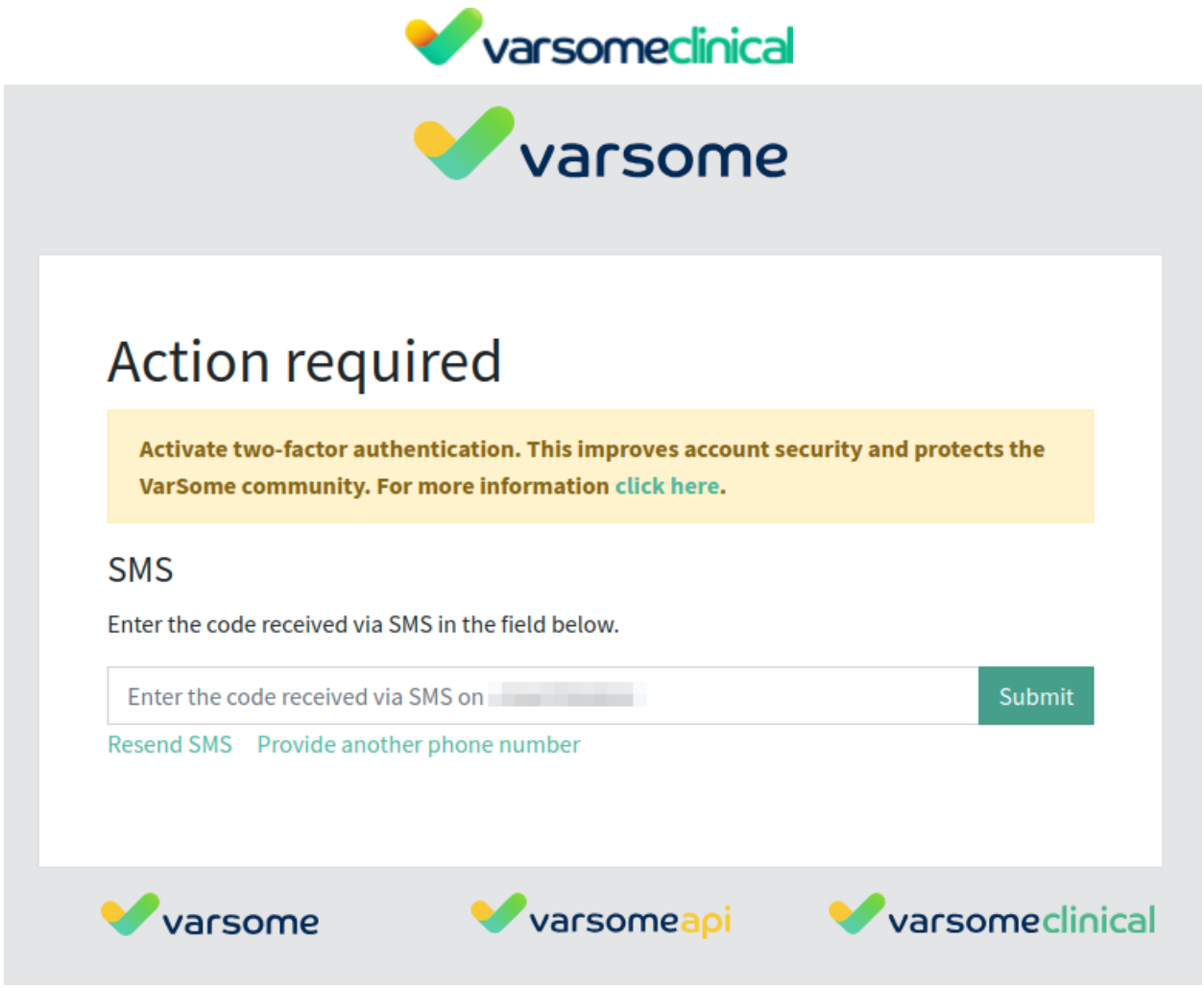
The next step to complete the log in, is to set up the two-factor authentication, which is mandatory for new users, by providing a mobile phone number.

⚠️ Please note that we do not accept numbers generated through a service that provides temporary telephone numbers. If such a number is used, the following error message will be displayed:

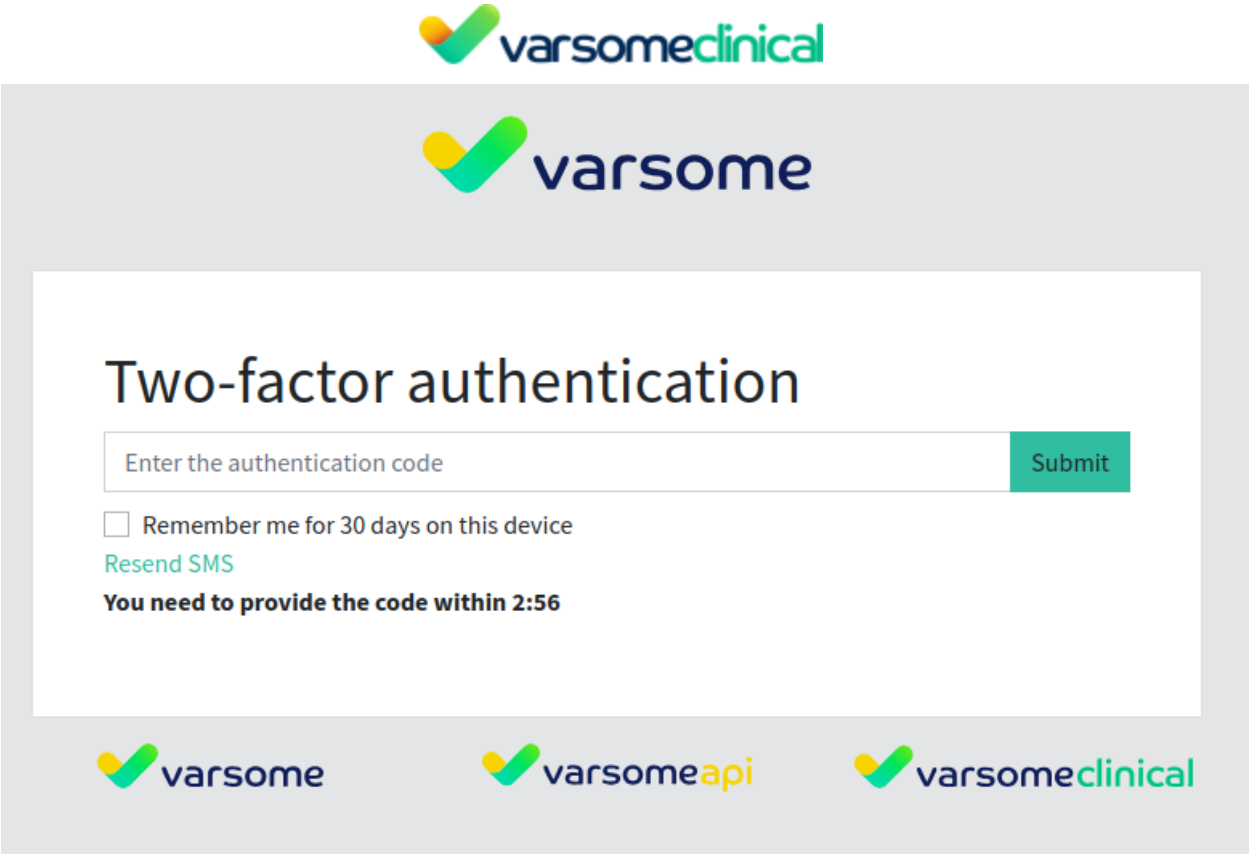
“This mobile phone number does not pass our system's security checks. Please use another number.”



After clicking Submit, a text message with a code should have been sent to the provided mobile number. Once you receive the code this has to be given in the next step in order to complete the log in.

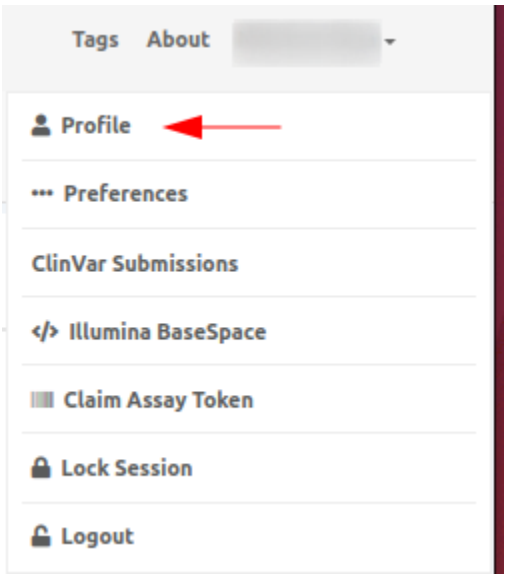


Everytime you log in to VarSome Clinical the Two-Factor Authentication method will be required. However, it is possible to remember the browser for 30 days and during that time you will not be asked for authentication.

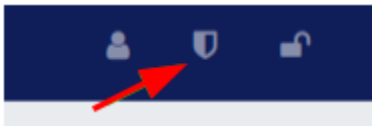


Other 2FA methods

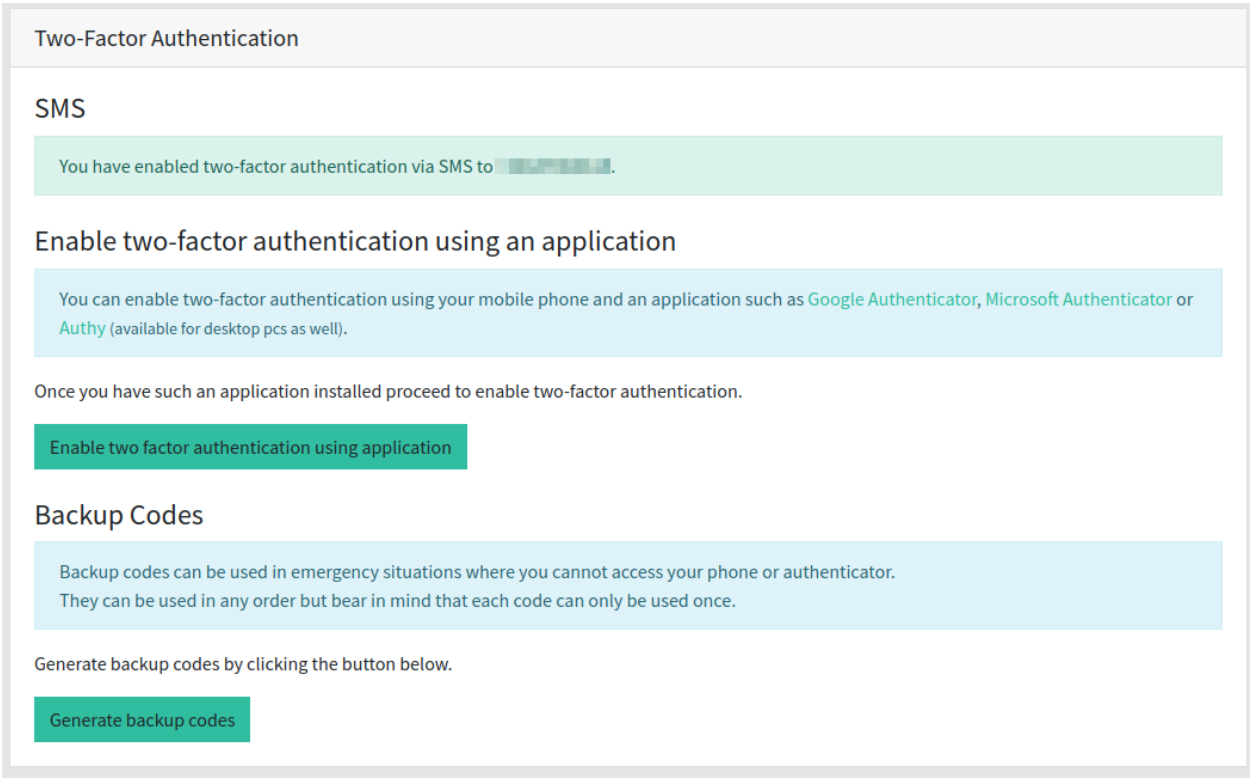
It is possible to enable 2 additional authentication methods, by using an application or saving backup codes. To enable the other methods there is a need to log in to sso.varsome.com, your VarSome account manager page. To land on the same page from inside VarSome Clinical, click on your name at the top right and then on Profile. This will redirect you to your VarSome account manager.



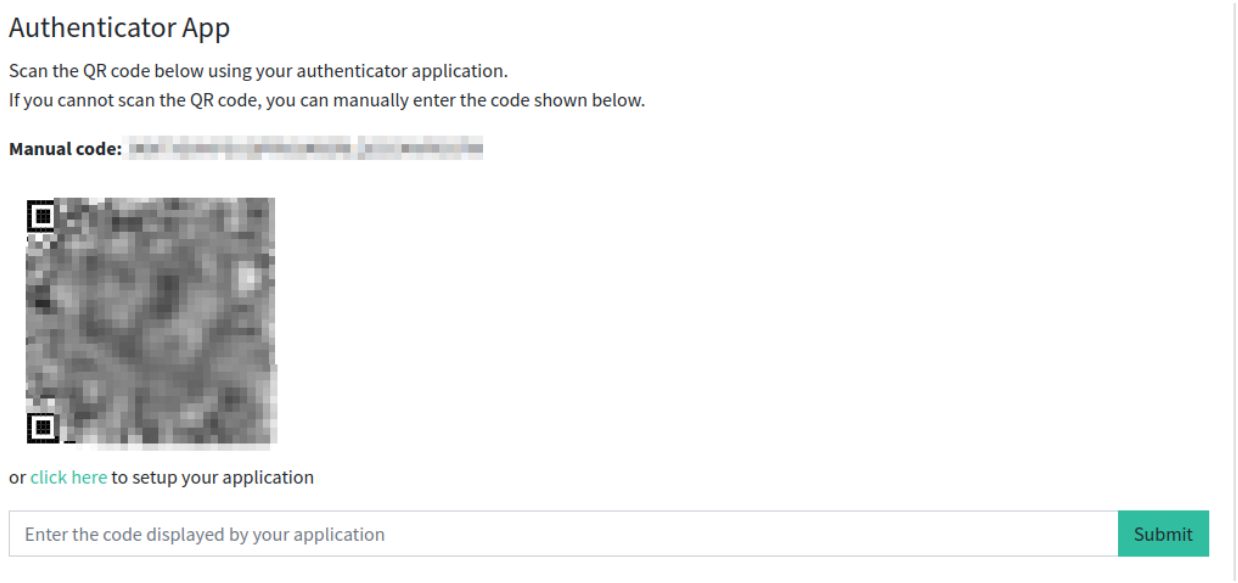
Once logged in to the VarSome account manager, on the top right of the page there is a shield icon.



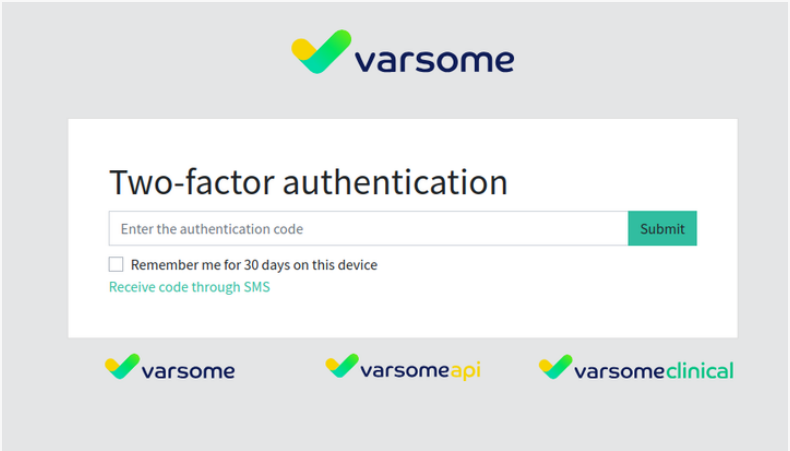
When clicking on it the user will be able to configure their privacy setting, such as change password, enable other authentication methods or see the API token.



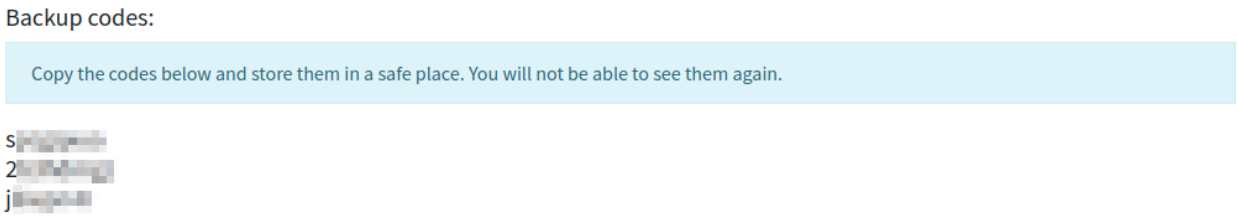
After enabling the application method you can scan the QR from e.g. Google Authenticator and verify it by providing the code displayed.



⚠️Please note that when using the Authenticator App, no SMS will be sent automatically. Instead, you should enter the code provided by the Authenticator App when prompted by VarSome Clinical. You may still request an SMS code to be sent by clicking on the "Receive code through SMS" link:



Alternatively, users can generate backup codes that will be shown once and can be copied and pasted to a secure place. Only 3 codes will be displayed and once you use those you can generate new ones.

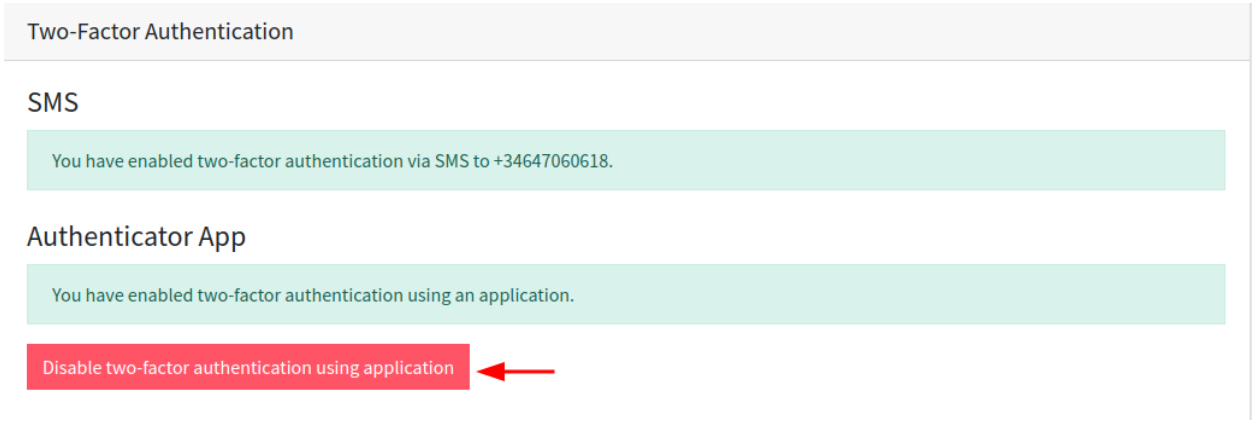


The backup codes should be used as the last option in case the user loses their phone or does not have access to the authenticator App.

Disable 2FA Authenticator APP

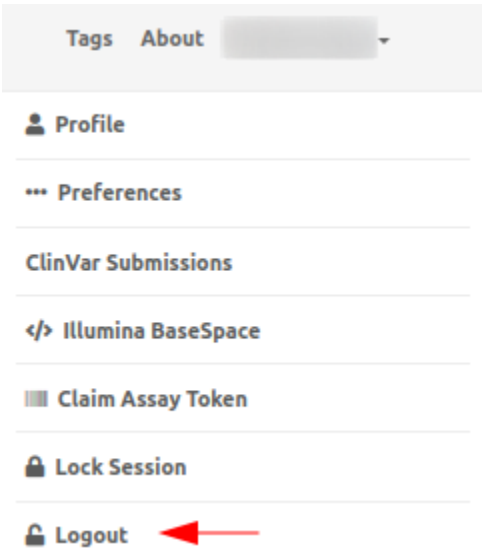
The Authenticator App can be disabled from the VarSome account manager by clicking on the "Disable two-factor authentication using application" button.

⚠️ Please note that it is not possible to disable 2FA entirely, so the the SMS option cannot be disabled.



Log out

Once you finish working with VarSome Clinical you can log out from the platform by clicking on your name on the top right corner and then Logout.



For security reasons VarSome Clinical logs you out after 15 minutes of inactivity.

Management

1. Group supervisor

In VarSome Clinical it is possible to set up a group supervisor for your account. The group supervisor is a VarSome Clinical user with certain privileges that other users from the same group will not have.
This user will have the ability to create, change and track:

- Workflows: create workflows with fixed parameters which can be used by you or other colleagues to launch analyses using pre-defined parameters.
- Launch analysis without a workflow: the supervisor can limit this option so that users can only launch analyses defined by him or her.
- Assay preferences: select an assay of a list of assays so that those are shown in the first place when someone is selecting an assay for launching an analysis.
- Storage preferences: as a group supervisor, you can modify the storage preferences of your group and decide how much time you would like to keep FASTQ or BAM files for example. Find more information here.
- Analysis preferences: you can enable the sensitive mode for CNV calling and decide if VarSome Picks should run automatically.
- Audit trail access: this keeps a record of different actions performed by the users of your group.

If you do not have a group supervisor yet or would like to change the group supervisor role to a different user, please contact our support team to change this.

2. Storage Management

Storage fees are calculated at the end of each month, based on the volume of data that has been uploaded into VarSome Clinical in order to run analyses. They reflect the storage costs of Saphetor data centers or the storage costs charged by 3rd parties (like GCP - Google Cloud Platform) to Saphetor.

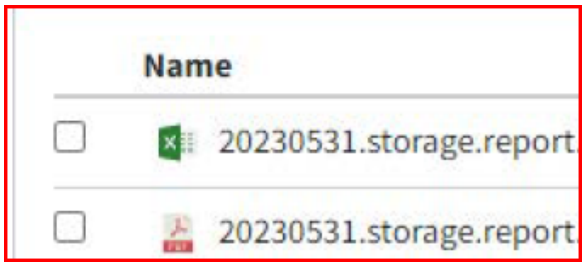
If a new sample has been uploaded during a given month, the storage costs for that month will be prorated based on the day of the upload.

Saphetor is not responsible for monitoring or managing the storage cost of any customer. This should be done by each end customer, possibly in coordination with the local distributor.

Monitoring the storage space

In order to check the volume of data that will be charged, each Account Administrator can visit sso.varsome.com (using his/her institutional user email address) and click on the Billing menu:

This window contains the monthly storage reports, in pdf and xls format, as well as the most recent billing entries, invoices history and account statements. The account administrator can then either filter on specific files or choose from the list.



Managing the storage space

There are 5 different options to possibly limit the volume, and hence the related costs, of samples data storage:

1. Keep all the data
   - The Storage Fee applies to all the data available in the client's account. This is the default option and needs no user action to be applied.

2. Remove FASTQ; keep the BAM and results.
   - Space occupied is lower than option 1 by approximately 25%.

3. Remove BAM; keep the FASTQ and results.
   - Space occupied is lower than option 1 by approximately 50%.

4. Remove all the raw data (FASTQ and/or BAM); keep the results only.
   - Space occupied is approximately 75% lower than option 1, as the annotation data doesn't take up much space.

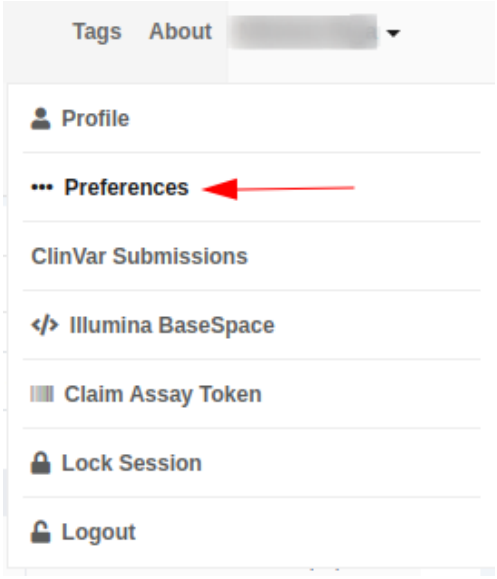
5. Archive the samples
   - No Storage Fee applies.
   - Samples results/annotations not available for browsing in VarSome Clinical.
   - The archived sample is still cross-referenced with other samples and sample links are still shown in the tab with cross-referenced samples.
   - Custom Variant Classifications set up initially for the archived sample will stay available when browsing other active analysis. The phenotypes assigned to the archived sample are still available in the sample information.
   - All sample data are deleted and not available anymore. Only the VCF file is stored for possible sample re-annotation when unarchiving the sample.

- The client may unarchive the sample. When doing so, the re-annotation fees will apply. For more details, please contact sales@varsome.com .
- When un-archiving the sample, VarSome Clinical will annotate the sample (VCF file) against the latest annotation data, which may cause annotation and classification differences between the original sample and the unarchived sample.

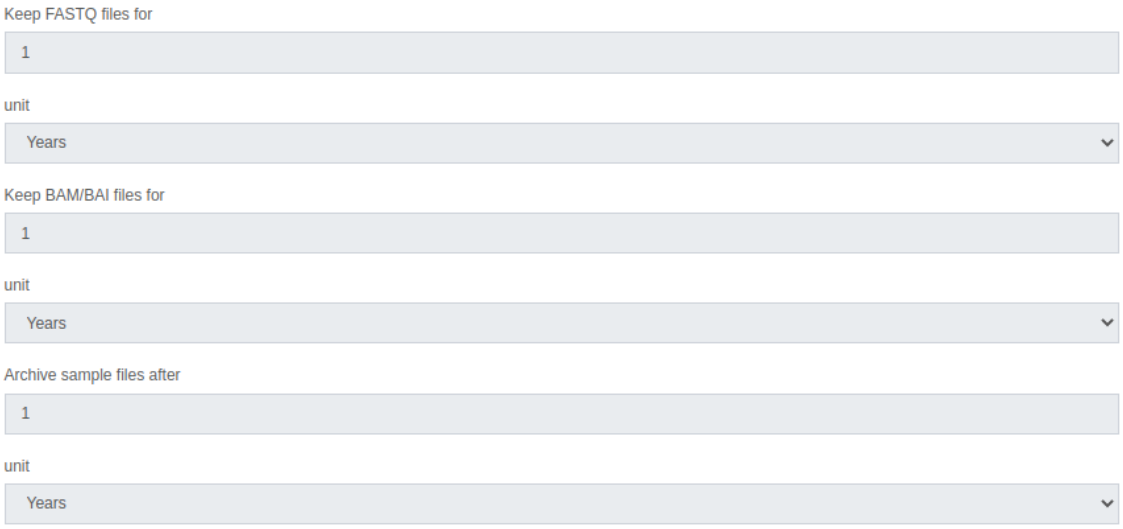
2. Setting up VarSome Clinical storage preferences

The storage preferences within VarSome Clinical can be modified in order to define for how long the user wants to store the FASTQ and BAM files and when they want their analyses to be archived.

In the VarSome Clinical platform, hover over your username on the right top of the window, then select "Preferences":



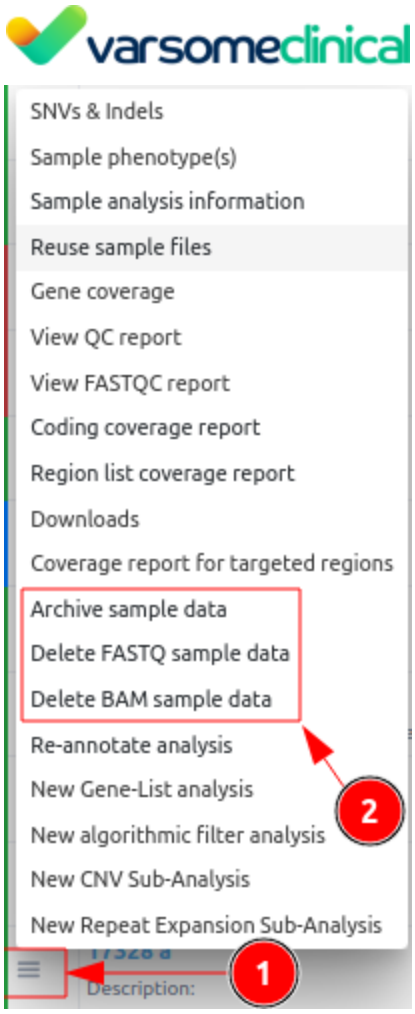
Only the group supervisor will have permission to modify these preferences. If you do not have a group supervisor yet, please contact support@varsome.com and let them know who should be your group supervisor. Once the group supervisor has accessed to the Preferences menu, they will be able to set up the storage preferences for each type of file:



3. Deleting / archiving data

Within VarSome Clinical it is also possible to immediately delete some FASTQ and BAM files, or also archive some analyses, without waiting for the time period defined previously.

For each single sample analysis that was launched from a FASTQ file, the user can click on the three horizontal bars on the right side of the sample name. They will then find the following options:



4. VarSome Clinical Token management

SSO platform and Account Administrator

Saphetor has developed an SSO platform on VarSome (sso.varsome.com) that can be used by the Account Administrator of the distributor to facilitate the usage of VarSome.
To use this application, one unique user from the distributor must be defined as the Account Administrator. This user is set up by the Saphetor Billing Department.

Token Definition

A token is a form of prepayment, an activation code for VarSome Clinical, which upon claiming, permits the client to analyze a certain number of samples belonging to a certain NGS assay.

Some tokens are for somatic and some for germline analyses.

Tokens support only FASTQ files.

How to get a Token

Tokens are issued by our partners, mainly distributors and assay manufacturers. Get in touch to learn more.

Issue a token

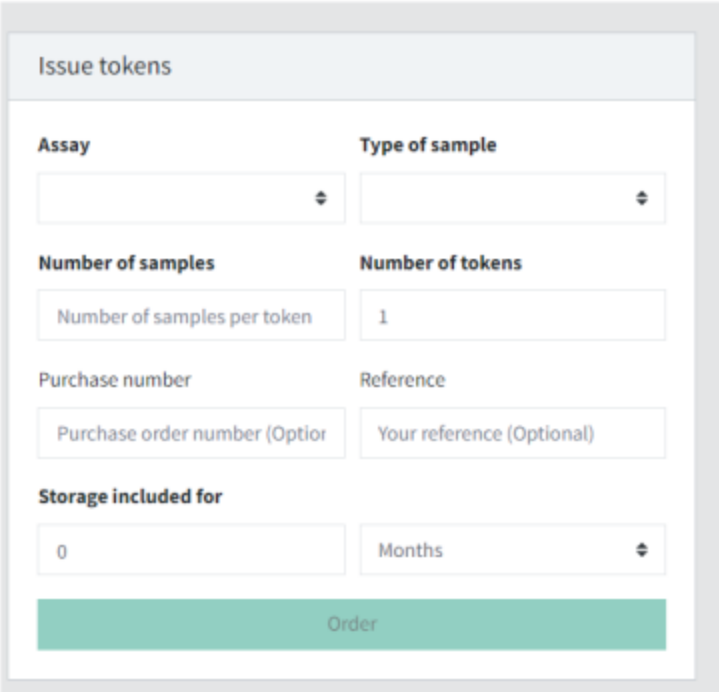
First, the token specifications for the user group must be set up by the VarSome Support Team (e.g. implementation of the token and setting up the group's account to access specific assays on the Clinical back-end). Also, ⚠️ please note that for this to work, the Account Administrator of the distributor ordering the token with specific assays should have access to VarSome Clinical, because assays are fetched from the Clinical platform.

Once all of the above actions are completed, the Account Administrator of the distributor needs to follow the steps below to issue a token:

i. The Account Administrator must go to sso.varsome.com and click on the Billing menu:

ii. Choose “Partners” button

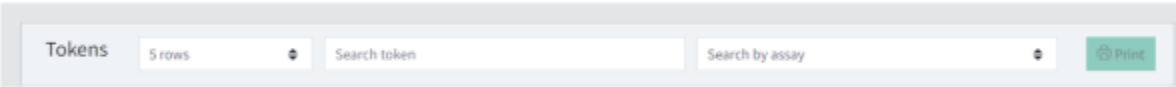
iii. Click on the screen “issue tokens”, type the assay and the type of sample (germline or somatic), the storage included in months, the number of samples and the number of tokens. Then, press order:



Print a Token and see all past token orders

On the next screen, your token will appear, as shown below, as well as all the tokens you have ordered so far. In the future, you will be able to print your tokens by pressing the button “print”.

On this screen, you will also be able to see all your token orders listed and filtered by assay.

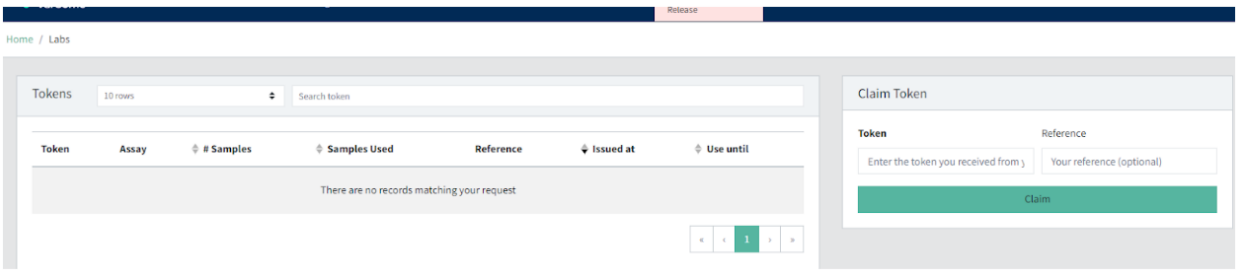


Claim a Token

Claiming a token means activating it. A token must be claimed by the end customer in order for an analysis to be run.

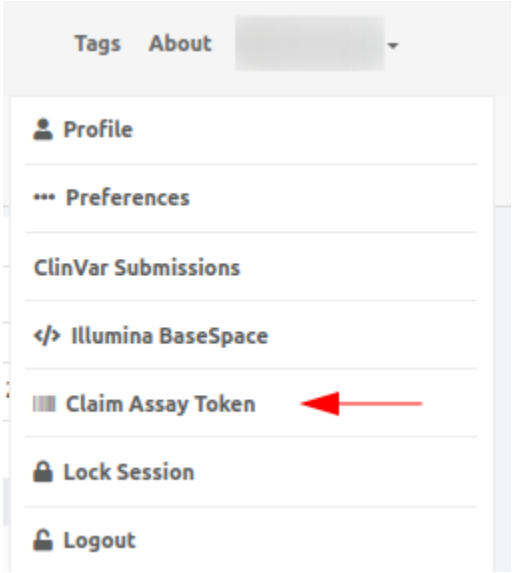
1. Claim a Token in SSO

Any user from the end customer needs to log into sso.varsome.com, as it drives more accurate results than using the VarSome Clinical platform, and click on Labs. They will see all the tokens that have been purchased but not yet claimed and can claim them by entering the token code on the Claim token window (on the right of the page) and clicking on the “Claim” button:

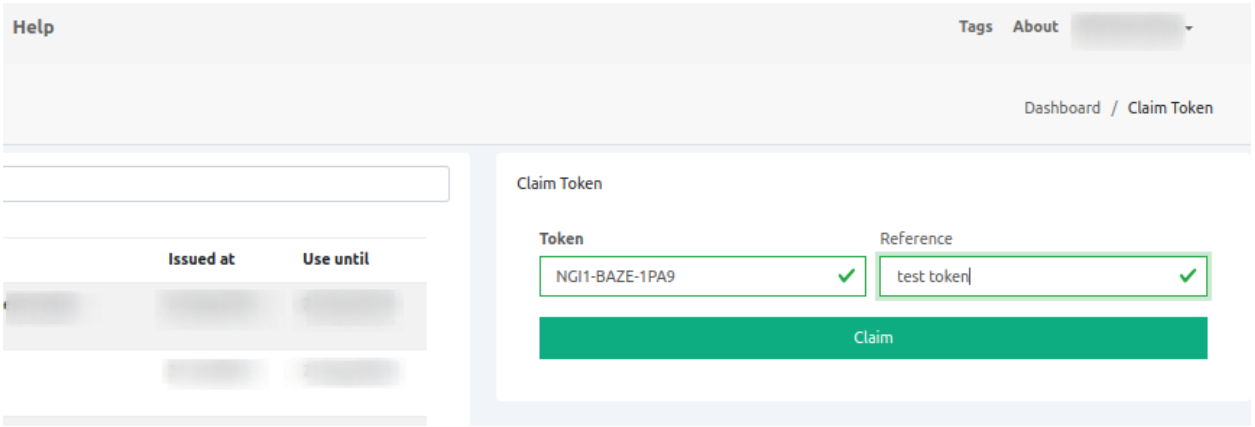


2. Claim a Token in VarSome Clinical

The user needs to log into VarSome Clinical and click on their username on the top right, and then choose the Claim Assay Token option from the drop-down menu.



Once the token has been received, the user will need to claim it to activate VarSome Clinical for the corresponding analysis.



Using the token to run an analysis in VarSome Clinical

- Single Sample Analyses

Once a token is claimed, no further action is needed in order to run analyses on VarSome Clinical. The platform, however, will validate whether a user attempting to run an analysis possesses non-expired tokens for the specific assay and sample type (germline fastq, somatic fastq) to be run.

Once an analysis is initiated, the system will deduct the number of analyses initiated by the user from the number of analyses available in the token. For example if a token allows a user to run ten (10) germline analyses from fastq and the user runs six (6) analyses, four (4) more analyses will be available for future use of the token.

- Multi Sample Analyses

For multi sample analyses the same validation occurs as with single sample analyses.
Once a multi-sample analysis runs, the number of available analyses for that token will be reduced by the number of samples included in the multi-sample analysis. For example, if a token allows a user to run ten (10) germline analyses from fastq and the user initiates a multi sample analysis from fastq files composed of three (3) samples, seven (7) more analyses will be available for future use of the token.

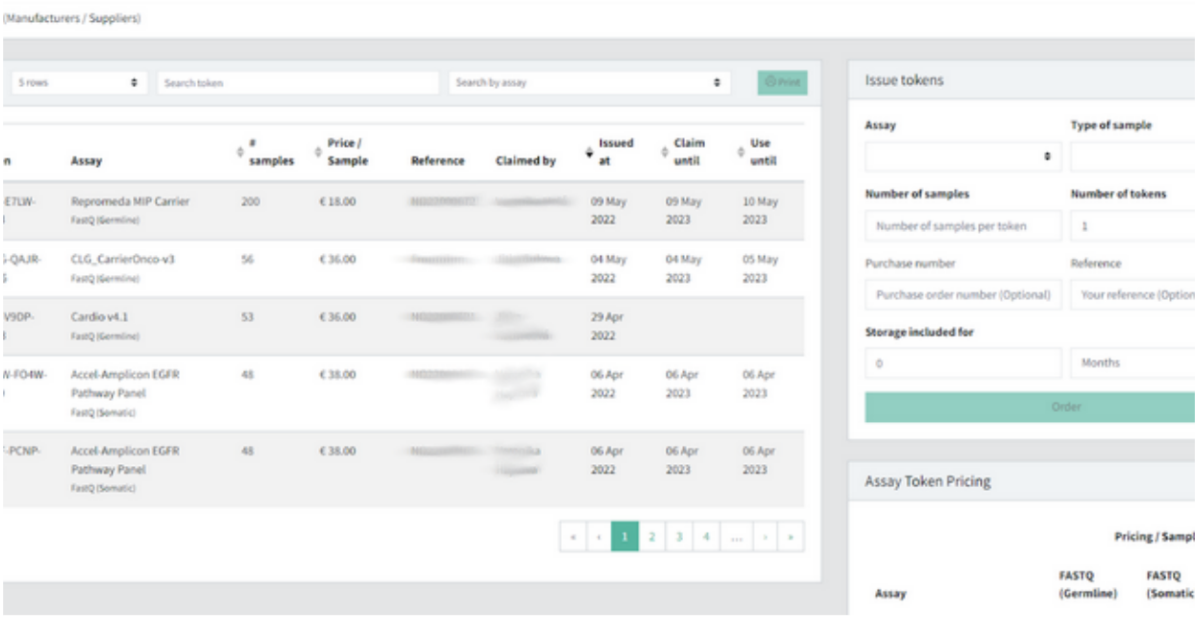
Token Usage Tracking

Here you can find all tokens you have claimed, along with the number of analysis remaining for each of them and the number of analyses performed.

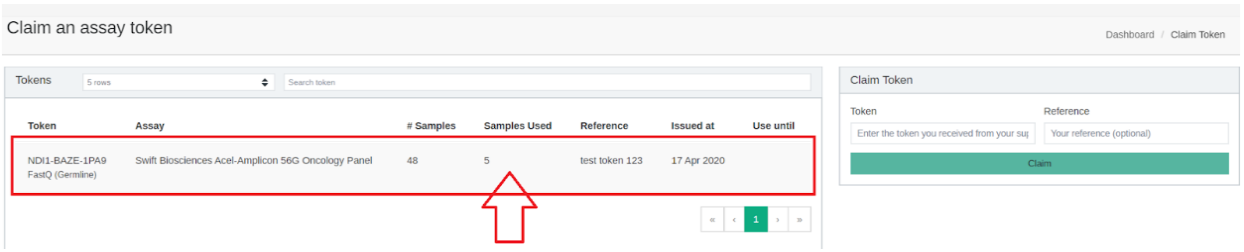
i. Token Tracking in SSO

The account administrator (from the distributor) or any end user must log into sso.varsome.com, as it drives more accurate results than using the VarSome Clinical platform, and click on the Partners tab.

All the tokens that have been bought will show on this page together with their actual status in terms of the number of analyses used and remaining, the price of the token per sample, the date of issue, the claiming date and the expiration date. There is also a price per sample list for all assays.



ii. Token Tracking in VarSome Clinical



Token expiry dates (claim and sample analyses)

The tokens have two (2) different expiry dates: one for the claim and one for the samples' analyses.

1. Token Claiming expiry date: twelve (12) months after the date of issue. This means that the end customer must claim the token within twelve (12) months of the token's issue date. If a token has not been claimed within twelve (12) months after it was purchased, it will be lost.
2. Analyses Expiry date: twelve (12) months after the claiming date. This means that the token is valid for twelve (12) months after it has been claimed. If some analyses linked to a token have not been used within twelve (12) months after the token was claimed, they will be lost. A billing entry is generated upon token issuing, regardless of whether the token will be claimed by the client or not.

Token-supported pipelines

Tokens only support analyses starting from FASTQ files. They can be issued for either somatic or germline analyses. By default, somatic tokens include AMP classification, which triggers extra charges. If a customer doesn't want to pay for this AMP classification, they must contact billing@saphetor.com to remove this default feature from the customer's account.

Tokens cannot be issued as combined pipelines (somatic and germline).

When a token is used, it cannot be converted to another assay or another version of the same assay than the one it was already issued for. For example, if a customer has ordered a germline token, has claimed it and started to use it, it could not be converted into a somatic one, in case they realized that it was ordered by error. The only way that this could be done is before the token is used, ask the Saphetor team to delete it and issue a new token instead. In such a case, the fact that it was claimed doesn't prevent the Saphetor team from deleting it.

Combined token analyses

Tumor normal analysis (paired analyses)

Tumor-normal analysis definition
The user has two different samples obtained from the same patient: a tumor sample and a "normal" (healthy) sample. The pipeline will report somatic variants observed only in the tumor sample.
For a tumor normal analysis, 2 tokens of the same assay must be used (a somatic token and a germline token) in conjunction, in order to obtain the desired results. The customer must buy both tokens (germline/somatic) of the same assay.

VarSome Premium for token-based customers

VarSome Premium is normally not available for commercial labs. If a token-based customer was exceptionally allowed to subscribe to both VarSome Premium and VarSome Clinical, they will not get any free monthly analyses on VarSome Clinical (up to the monthly value of the Premium subscription). The reason for this is that in our customer billing platform, we set the monthly credit limit to zero to avoid analyses that would go beyond the token amount and hence create extra charges. On the other hand, a VarSome Clinical account as a stand alone solution is set up with no monthly balance limitation and this is why it can be perfectly combined with the VarSome Premium product which allows limited free access to VarSome Clinical up to the monthly value of the Premium user.

⚠️ Please note that we developed the tokens solution mostly to answer distributors' specific needs to bundle assays and analyses. These tokens have proved to be quite complex to handle, especially when they are used outside of the most common analyses, so please contact Saphetor for any non-standard request from a token-based account.

Instructions

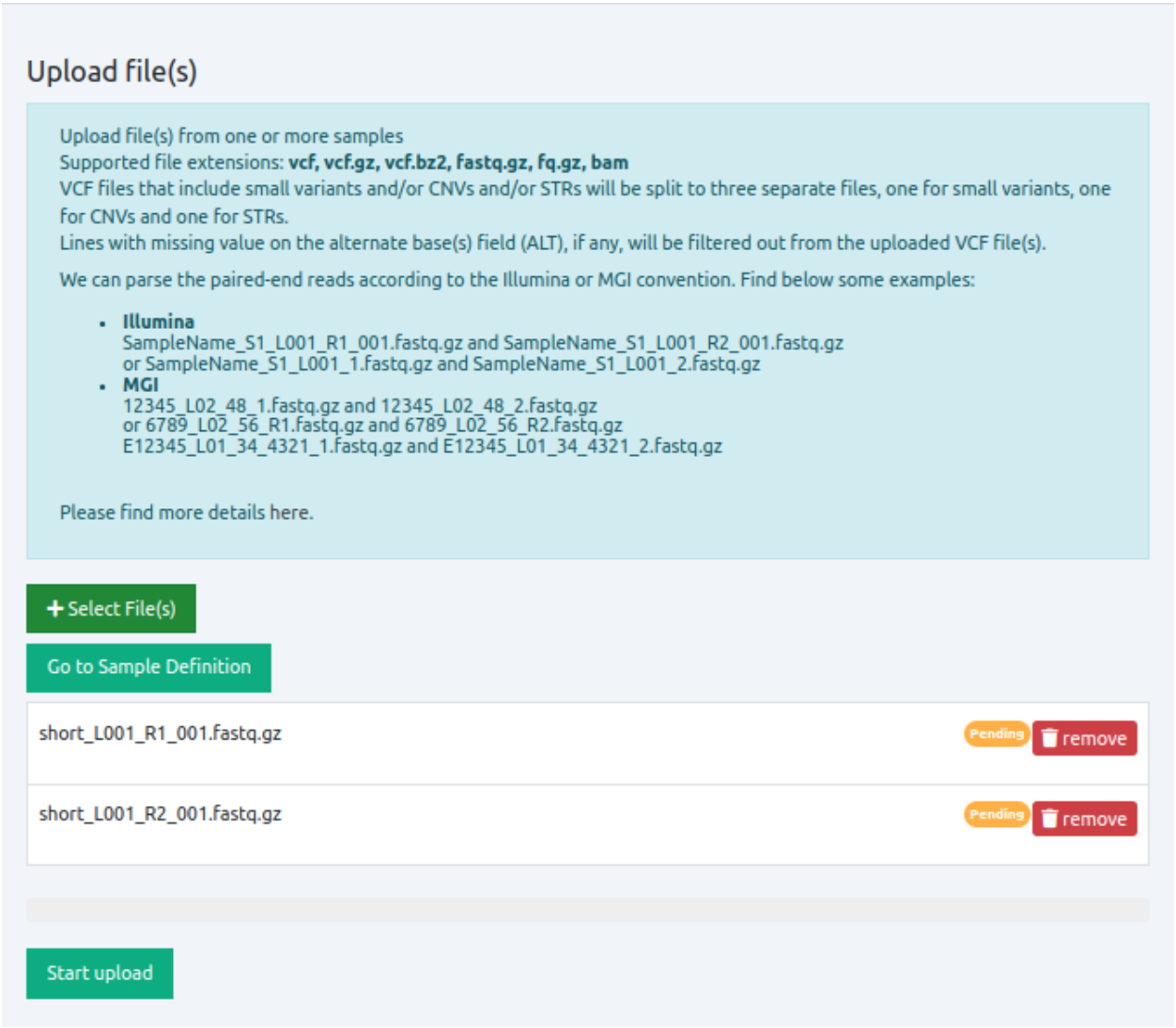
The sequence of actions to analyze a sample in VarSome Clinical are the following:

- Upload files to VarSome Clinical
- Create or define the samples
- Create a workflow (optional)
- Launch a new analysis: select a workflow/use blank and connect it to the samples to analyze

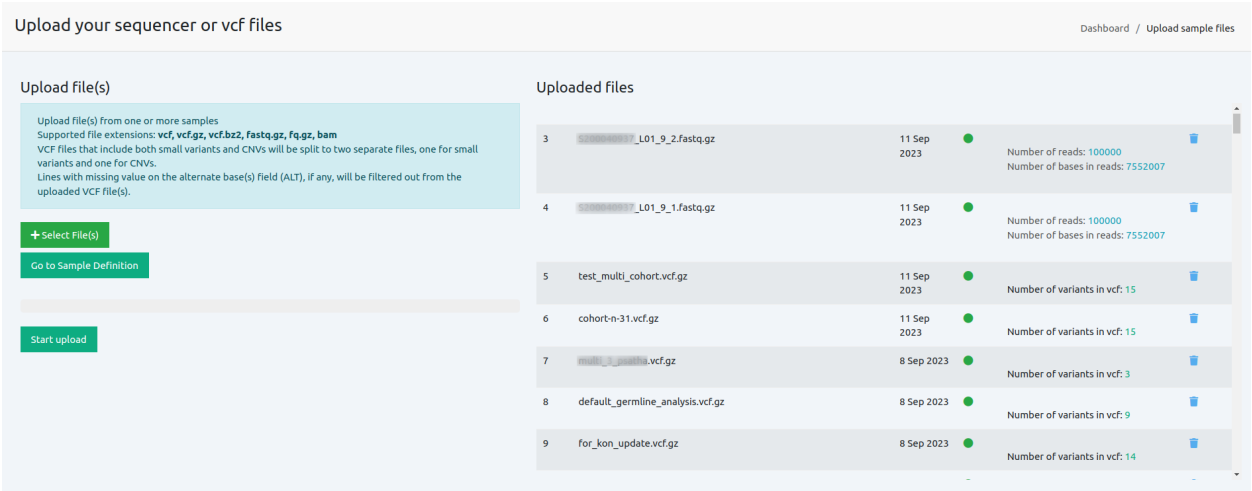
In the following sections you will find the detailed information on how to perform each of the steps.

1. Upload files to VarSome Clinical

To upload files to VarSome Clinical, click on ***Upload/View files***. Then, click on "Select File(s)" and, in the file browser window that will appear, select all the files you want to upload. Files do not need to be from the same sample or in the same format. Once all files have been selected, the file names are displayed under the green "Select File(s)" button. To upload the files, click on "**Start Upload**".



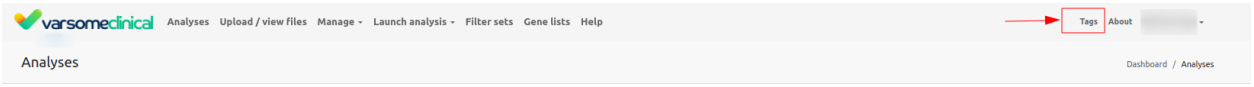
Files are uploaded and checked, and the number of reads and bases in reads is calculated and displayed for each file. Files can be deleted before or after upload. Files with status set to dark green can be used for a subsequent analysis.



Click on “Go to Sample Definition” to define your samples from the uploaded files.

⚠️ Please note that files that have been uploaded and not used for more than 30 days will be automatically deleted from VarSome Clinical.

You can organize your analyses using Analysis tags. It is possible to create new tags or edit existing ones by clicking on the "Tags" option next to the User name.



To create a new tag, enter a name in the "Tag label", choose a color and click on save.

To edit an existing tag, click on the edit icon displayed on each tag, and to delete a tag click on the trash icon:

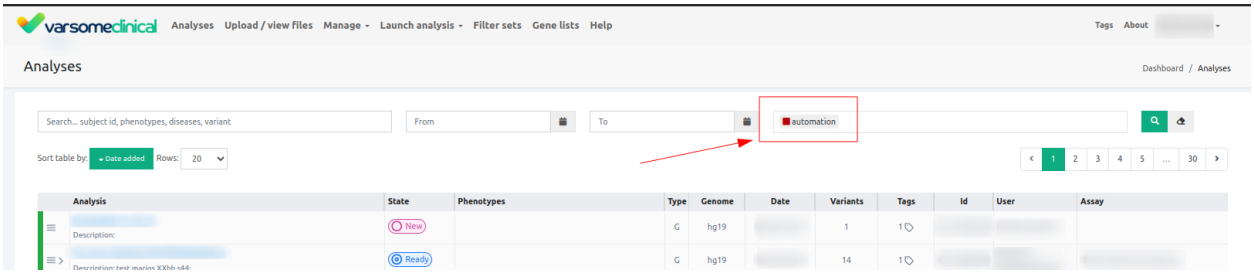
Edit icon: 

Trash icon: 

To add a tag to an analysis, click on the tag icon shown in the Analyses Table and select the tag you want to add to the sample:

Tag icon: 

To filter samples using tags, click on the "Tags" box on the top of the page and select the tag that you want to use to filter your samples. Now, only analyses containing the selected tag are displayed:



1.1 Accepted input files

The accepted input files to run analyses on VarSome Clinical are either:

- FASTQ files only from Illumina or MGI sequencers
- VCF files which conform to the VCF standard, regardless of sequencing platform. Users may also optionally upload an alignment BAM file for the VCF sample which can be used to visualize the coverage of the variants provided in the VCF file.

1.1.1 Accepted file names for FASTQ

We expect files that conform to Illumina's or MGI naming convention.
When providing paired-end FASTQ files, we require that reads are properly coordinated between them. Paired-end reads provided in a single FASTQ file are not accepted.

For Illumina pair-end files, we will consider pairs to be files with the exact same name except for the number of the read, for example SampleName_S1_L001_R1_001.fastq.gz and SampleName_S1_L001_R2_001.fastq.gz. We accept files in which the read number is specified alone (for example SN1234_S1_L001_1.fastq.gz and SN1234_S1_L001_2.fastq.gz) or with an “R” before the number (for example SN5678_S1_L001_R2.fastq.gz and SN5678_S1_L001_R1.fastq.gz). For further instructions in terms of naming conventions, please refer to Illumina.

For MGI pair-end files, we will parse the files as follows: [flow cell ID]_[lane ID]_[barcode ID]_(optional_id)_[read 1/2].fastq.gz and we accept the number of the read to be specified alone (for example, 12345_L02_48_1.fastq.gz and 12345_L02_48_2.fastq.gz) or with an “R” before the number (for example, 6789_L02_56_R1.fastq.gz and 6789_L02_56_R2.fastq.gz)

In cases where there are more than two paired-end files per sample, all the paired reads should be provided: R1 with R2, R3 with R4, R5 with R6 and so on.

1.1.2 Requirements for submitted VCF files

VarSome Clinical accepts VCF files for SNPs/INDEL and CNV annotation. You can upload VCFs containing only SNPs/INDELs or CNVs, but you can also upload VCFs containing both types of variants. If you upload a VCF containing both types of variants, it will be divided into two files: one file to annotate SNPs/small INDELs (**filtered.vcf.gz*) and one file to annotate CNVs (**cnv.vcf.gz*).

1.1.2.1 Required format for SNPs/INDELs annotation

VCFs containing SNPs and small INDELs can be used to launch a somatic or germline analysis: (Launch analysis > New analysis > Germline/Somatic Illumina analysis from VCF).

The VCFs uploaded to analyze SNPs/small INDELs variants must have the following requirements:

1. Are compliant with the VCF standard.
2. Include *specific* SNVs and INDELs. In order to annotate a variant, we need to know exactly what that variant is, so we cannot handle cases where the variant's sequence isn't specified. For example, we cannot annotate "NON_REF" variants:

```
#CHROM  POS       ID        REF   ALT
chr1    10052     .         C     <NON_REF>
```

Or variants with an "N" in the ALT field:

```
#CHROM   POS       ID       REF   ALT
chr22    30998425  .        C     CTTTTTNT
```

3. Include a valid genotype (GT) field for each variant entry.
4. The files should contain the variants found in a real human sample. We expect a maximum of around 4 or 5 million variants in a sample.

1.1.2.2 Required format for CNVs annotation

VCFs containing CNVs (deletions and duplications) can be used to launch a CNV subanalysis from VCF. The VCFs uploaded to annotate CNV variants must have the following requirements:

1. Are compliant with the VCF standard.
2. Include duplications and/or deletions where the type of copy number variant is shown in the ALT field:

```
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1
chr12 133040735 . C <DUP> . PASS SVTYPE=DUP;SVLEN=140;END=133040875 GT:CN 0/1:1.50 chr12
133049934 . G <DEL> . PASS SVTYPE=DEL;SVLEN=78;END=133050012 GT:CN 0/1:0.50
```

3. According to the VCF Specification, the CNV category should not be used when a more specific category can be applied. Therefore, the following VCF format is not accepted:

```
##ALT=<ID=CNV,Description="Copy Number Variant">

chrX   133559227 . G <CNV> . . SVTYPE=CNV;SVLEN=140;END=133559366   GT:CN 0/1:1.50
```

4. Include a valid genotype (GT) field for each variant entry.
5. Do not include other types of SV variants such as large chromosomal rearrangements (e.g. inversions, translocations) or gene fusions. We currently do not support these types of SV variants.

There are various ways to ensure that the format of your VCF files is correct. For more details please see below:

❗ **Tip: checking the format of a VCF file**

Ensuring that your VCF file is structured correctly and ready to be uploaded to VarSome Clinical is a recommended practice that could facilitate your analyses and save valuable time.

An easy way to check that your VCF file is valid is to try to run a bcftools command on it. Bcftools, a set of utilities that manipulate VCF files, is very sensitive to malformed VCFs, so it will fail if the file doesn't conform to the standard.

After installing Bcftools according to the instructions, the following command can be executed, where file.vcf represents your input VCF file:

```
bcftools norm -m -any -NO v file.vcf
```

This command will attempt to perform certain actions: check that REF alleles match the reference, split multiallelic sites into multiple rows, or recover multi allelics from multiple rows. If the fields in your file are complete, the command will be executed smoothly. However, if it comes across a non-compliant field like the following,

```
chr1    16366632        .       CC      GC,GT   193.02  PASS    AB=0.5;
```

the command will fail. In the row above, the field allelic balance (AB) is incomplete, as this is a multiallelic site with two alleles in a single row and two numbers are expected. This information will be provided with an error message:

```
Error: wrong number of fields in INFO/AB at chr1:16366632, expected 2, found 1
```

Other alternatives to VCF validation:

- https://github.com/EBIvariation/vcf-validator
- http://vcftools.sourceforge.net/perl_module.html#vcf-validator

which can be used to locate other types of errors (e.g. a malformed or missing header).

Another quick test is to just see if a standard program like bcftools recognizes the file and doesn't complain.

ℹ️VarSome Clinical API
The platform comes with full API, allowing you to automate each step of the data analysis process, including the data upload. Documentation for VarSome Clinical API

1.1.2.3 Required format for Repeat Expansion annotation

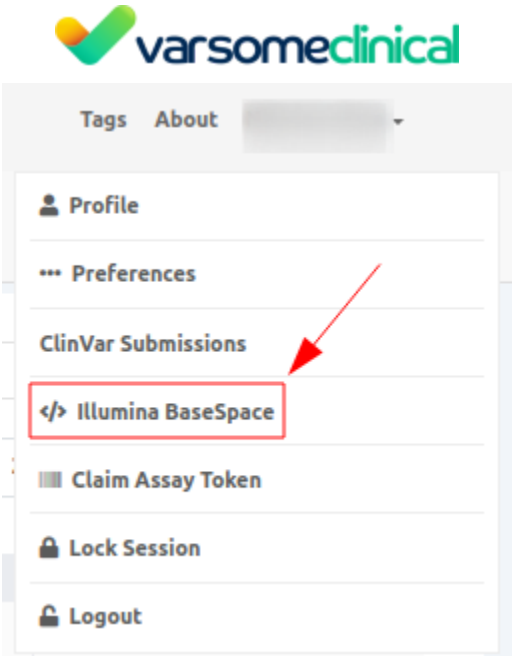
VCFs containing short tandem repeat (STRs) can be used to launch a Repeat Expansion sub-analysis from VCF. The VCFs uploaded to annotate STRs must meet the following requirements:

1. Are compliant with the VCF standard.
2. The number of repeats is shown in the ALT field as < STRn > where *n* is the number of repeats.
3. The INFO field contains the repeat unit in the following format: `DisplayRU=CCG`
4. The FORMAT field contains the number of repeats spanned by each allele.

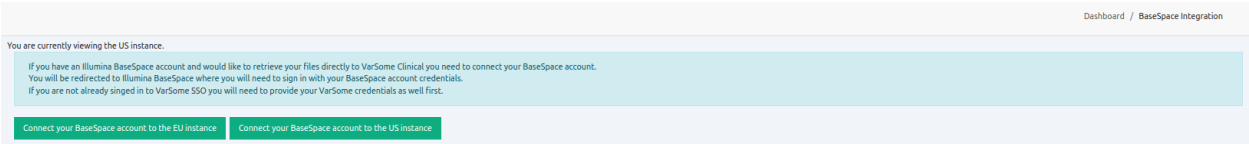
```
##FORMAT=<ID=CN,Number=A,Type=Integer,Description="Number_of_repeat_units_spanned_by_the_allele">
```

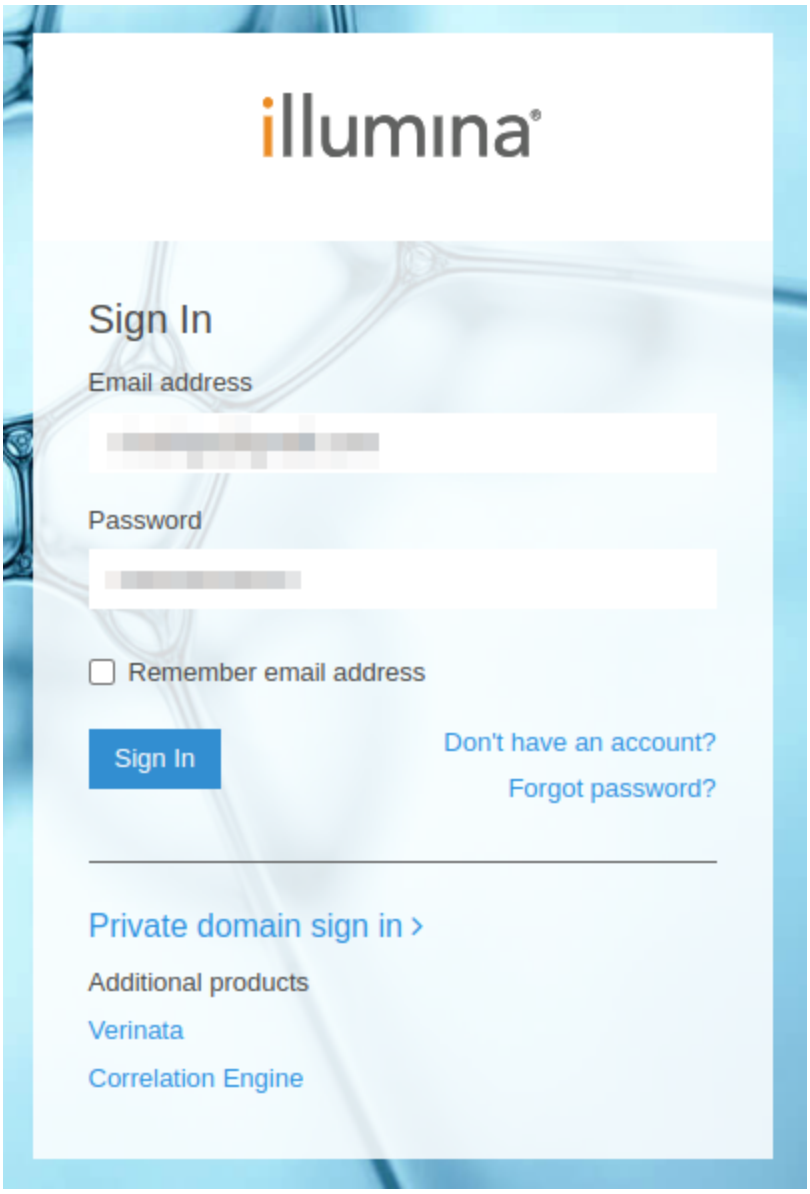
1.2 Upload FASTQ files from Illumina BaseSpace.

This feature can be accessed by clicking on the “Illumina BaseSpace” option displayed in the drop down menu shown when you hover over your user name.

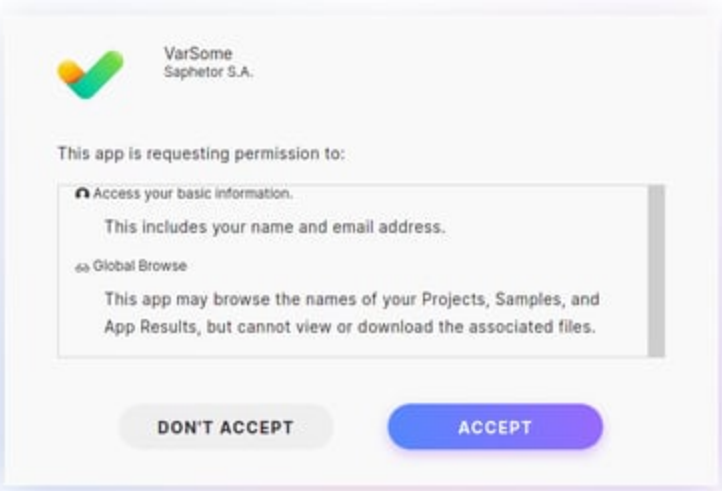


There are two options to connect your Illumina BaseSpace account depending on which BaseSpace Sequence Hub your data is allocated in (EU or US). After that, you will be automatically redirected to Illumina's site in order to login to your account.



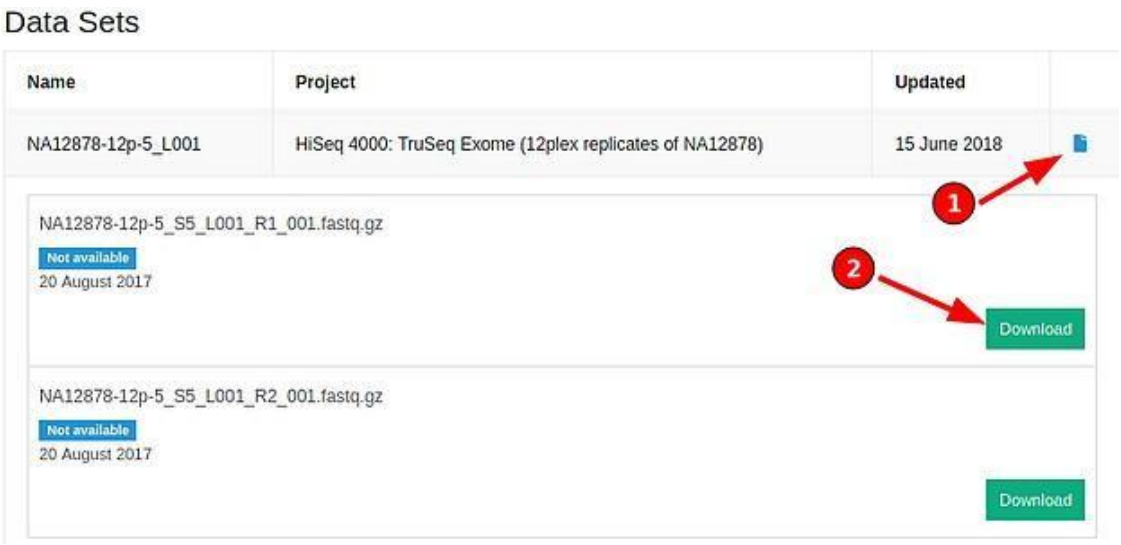


Your projects will then synchronize with VarSome Clinical after you have granted the necessary permits.

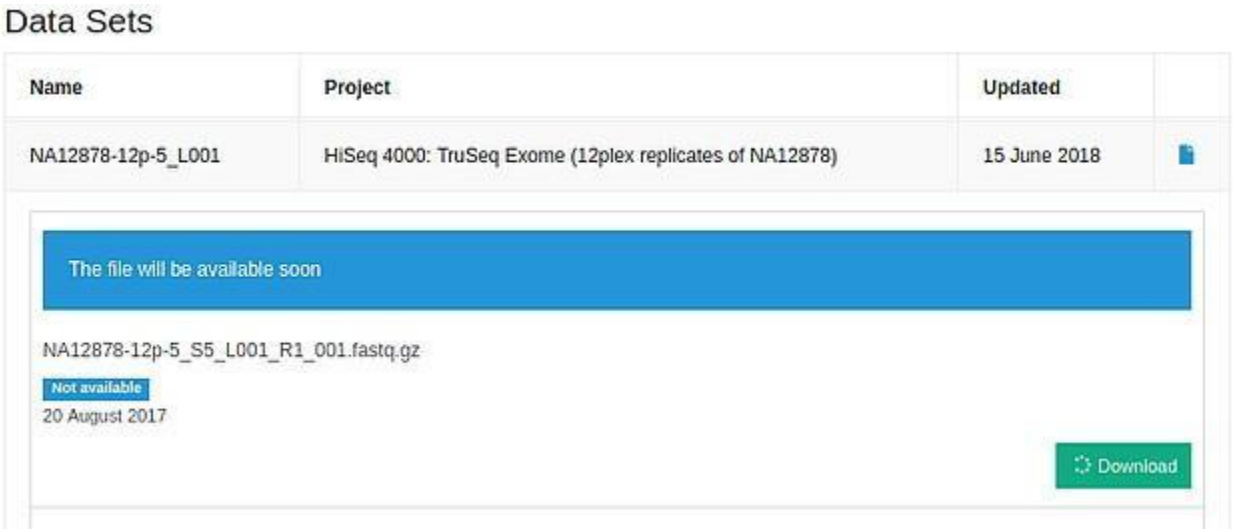


⚠️ Please note: if you do not grant access to VarSome, then your projects cannot synchronize and you will not be unable to transfer files.

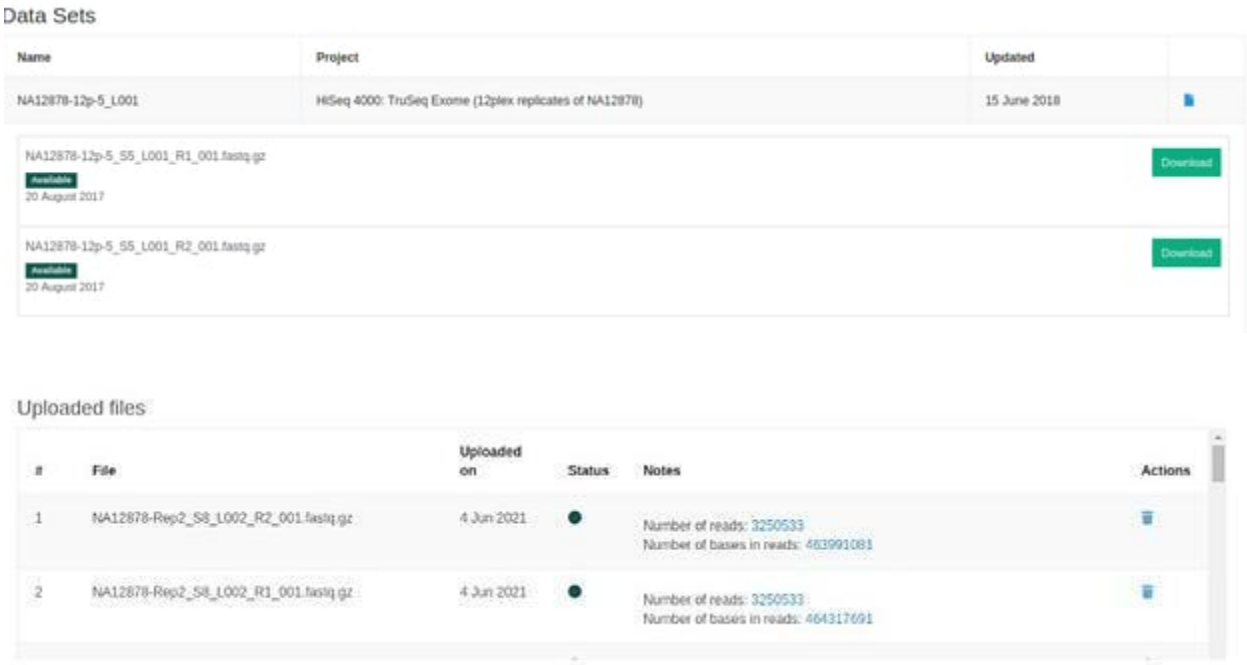
This functionality will allow you to download the FASTQ files directly to VarSome Clinical. To do this, please, click on the “Download” button to start downloading the reads into VarSome Clinical.



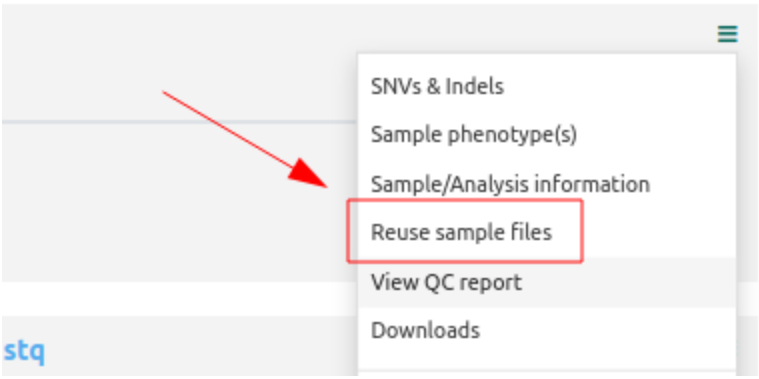
Please, wait a few minutes while the files are being downloaded. The downloading time will depend on the size of your data and on your internet connection.

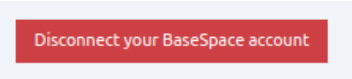


Once the download is complete, the status of the file will change from “Not available” to “Available”, meaning that the FASTQ files are accessible from the “Upload/View files” page and can be used to launch analyses (germline or somatic) from FASTQ.



⚠️ Please note that any file that has been previously imported from BaseSpace to VarSome Clinical will have a status of “Available”. If the file has already been used in an analysis, in order to analyze it again you will need to find the sample that contains it and select the option “Reuse sample files”.



Click on 

to disconnect the Illumina BaseSpace account from VarSome Clinical.

2. Manage

- Samples
- Workflows

Click on “Manage” to create, view, or edit your samples or workflows.

2.1 Samples

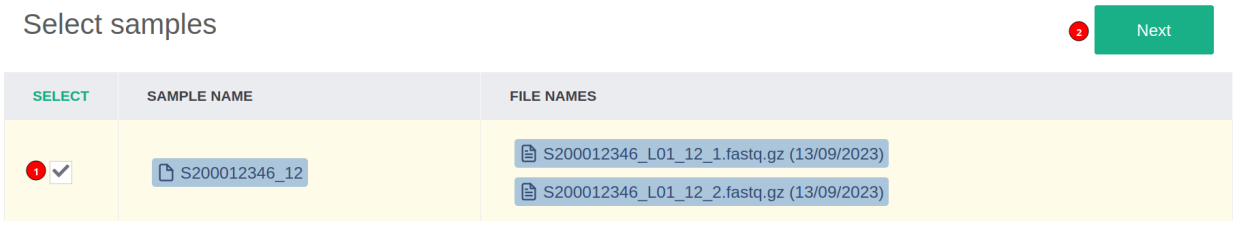
2.1.1 Create and define your samples

Go to Manage > Samples > Create new samples. **This step is required before using the samples for analysis.**

Once your files are uploaded, you need to create and define your samples. The sample creation involves the association of files to sample names and addition of optional sample metadata related to the patient (e.g. phenotypes).

Uploaded FASTQ files are recognized automatically in pairs (or groups) under the same filename prefix and according to the Illumina or MGI naming conventions. The table displays suggested sample names for each file or pair(s) of files. For VCF files, each file is associated with a suggested sample name.
The files shown in first place with a yellow background are the ones uploaded by the current user. Then, we display the files uploaded by other users (gray background). All files are sorted by their upload date.
Select the rows with suggested sample names in order to create your samples. You can select all by clicking on the “Select” option at the top of the table.

Once the files and suggested names are selected click on “Next”.



In the following menu you can optionally:

- Modify the **sample type** (germline or somatic). Click on the “germline” or “somatic” boxes if you wish to change the sample type.
- **Files**: remove or add files.
- **Sample name**: you can edit the suggested name of your sample here. Please remember that the sample ID should not contain information about patient identity.
- **Description**: add a description of your sample. Please remember that, like the sample ID, this field should not contain any information that can be used to identify a patient.
- **CNV VCF file for CNV sub-analysis**: you can choose along with the VCF/FASTQ files of your sample a VCF file that contains only CNVs for the same sample to be annotated. The CNV results will be displayed as a sub-analysis accessible from the main analysis of the sample.
- **Repeat Expansion VCF file for Repeat Expansion sub-analysis**: you can optionally add a VCF file containing repeat expansion (RE) variants for annotation. The RE results will be displayed as a sub-analysis accessible from the main analysis.
- **BAM file for alignment visualization** (VCF samples only): you can optionally attach a BAM file that will allow you to visualize the alignment of the reads and access to the IGV or JBrowse from the variant table. Please note that this file won’t be used for variant calling, only for visualization purposes.

You can define your sample as germline or somatic:

Depending on the sample type, there will be the following optional fields:

Germline samples:

- **Phenotypes**: to select phenotypes of interest, start typing a term and the associated phenotypes will appear as a list that can be selected. Phenotypes shown in the dropdown list can be limited based on their source. When selecting “All” you will get terms from HPO, MONDO and OMIM® databases, while selecting “Only OMIM®” they will be retrieved solely from the OMIM® database. These terms can be used to create a gene list and filter the variant analysis for the genes matching the selected phenotypes. For more details see: Phenotype Matching.

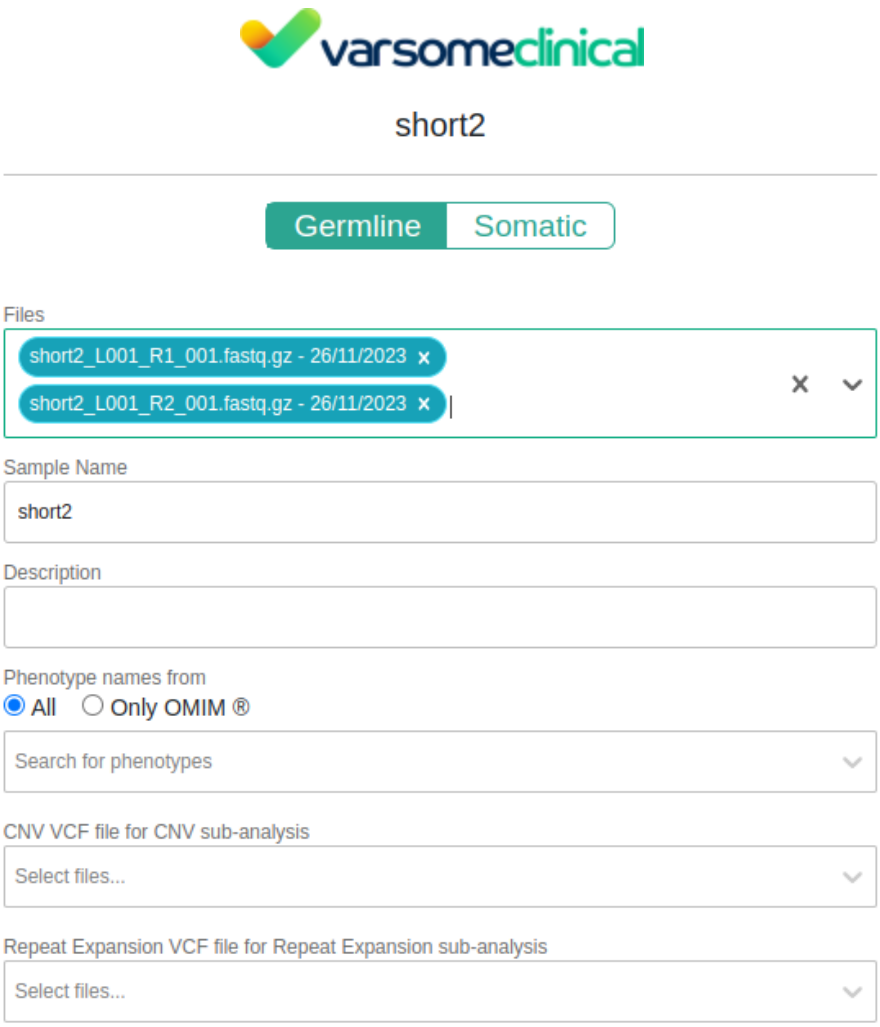
Somatic samples:

- **Tissue type:** specify the tissue type of your sample. This field is optional, however we encourage you to fill it in, as this information is taken into account when annotating the variant with information from cancer databases.
- **Cancer type:** specify the sample's type of cancer. This field is optional, however we encourage you to fill it in, as this information is taken into account when annotating the variant with information from cancer databases.
- **Age (years):** specify the sample's individual age (in years). This field is optional, however we encourage you to fill it in, as this information is taken into account when annotating the variant with information from cancer databases.
- **Sex:** specify the individual's sex. This field is optional, however we encourage you to fill it in, as this information gets considered for the annotation with the cancer databases.

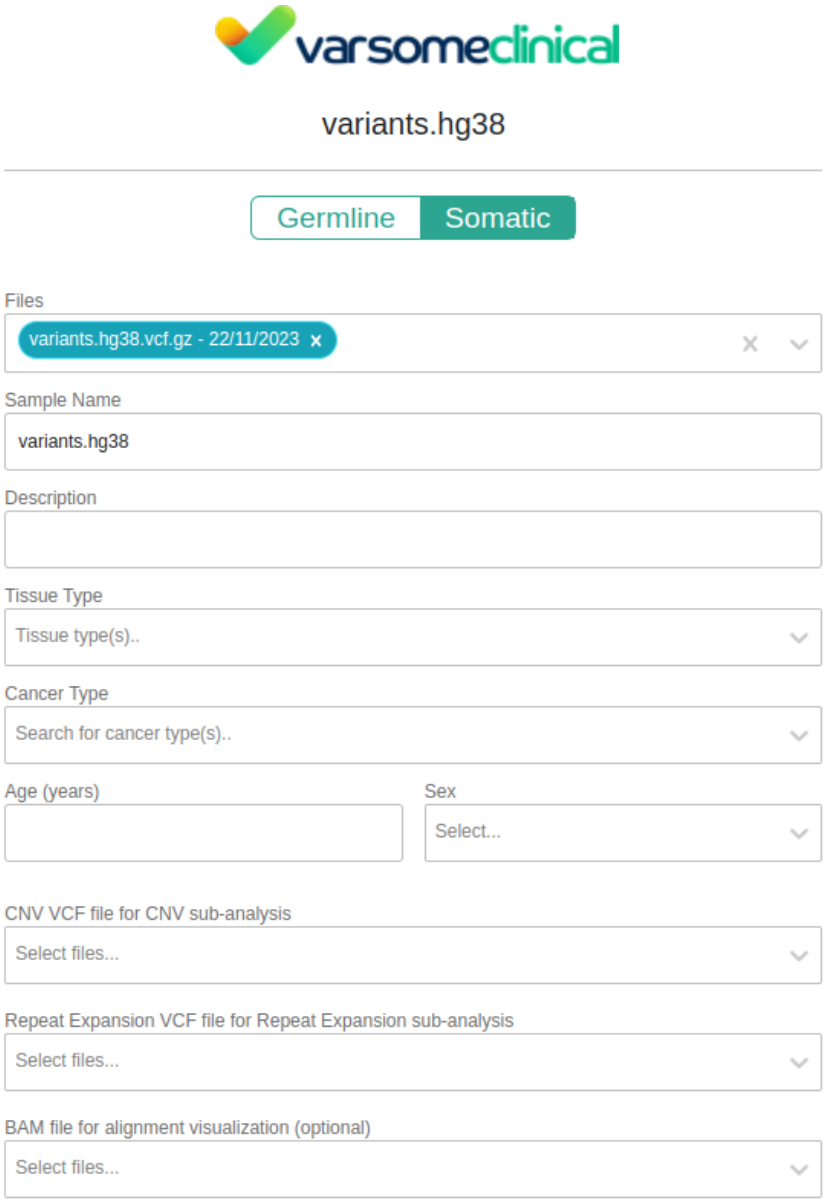
Please note that while all of the fields above are optional, we encourage you to fill them all in because this information is taken into account when annotating the variant with information from cancer databases.

Click on “Create” to create the samples. Your samples are now ready to launch a new analysis with them. Go to “Launch new analysis” to find more details.

FASTQ germline example



VCF somatic example



2.1.2 View/Edit your samples

Go to Manage > Samples > View/Edit new samples

This table shows all created samples. If the current user has already created some samples, they will be shown first (yellow background) while the samples created by other users of the same group are displayed last (gray background). All samples are sorted by the upload date.

In the actions column you can view/edit the sample information and remove the sample. When a user removes a sample, the associated files become available again to create a new sample (Create new sample menu).
Please note that you can only edit and remove your own samples. These actions are not granted for samples created by other users of your group.

2.2 Workflows

Go to Manage > Workflows

2.2.1 Workflow table

The table shows the workflows created within your group, the user who has created them, the date of their last update and the available actions to perform:

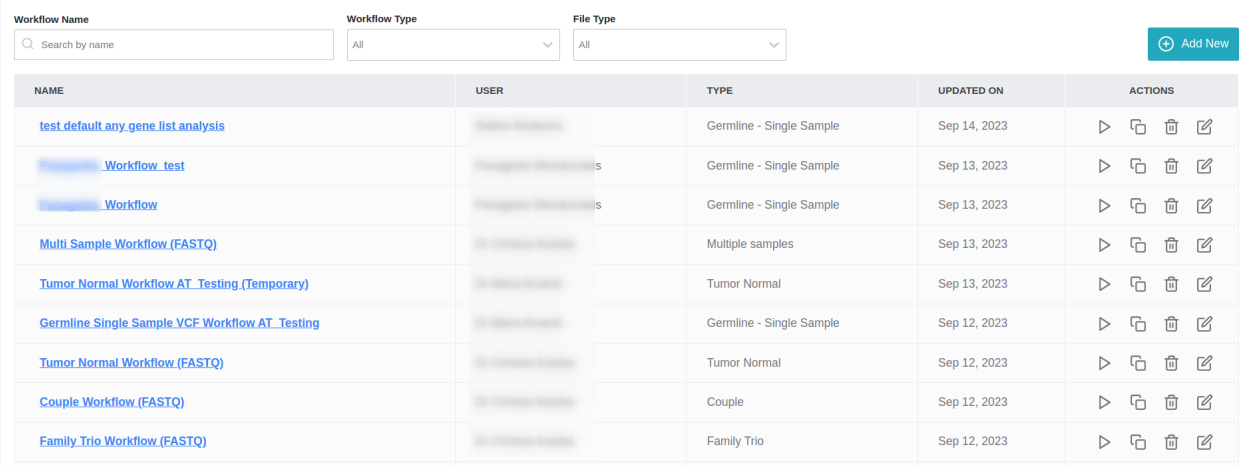
- Use workflow (▶️): this is to launch an analysis using this workflow. When selected, you will be directed to the launch new analysis page where this workflow is already pre-loaded. You can optionally modify any workflow parameter (if permitted) and then click on “Next” to select the samples to analyze using this workflow.
- Copy a workflow: create a new workflow by copying the values of an existing workflow. Please note that the workflow names must be unique.
- Remove a workflow (🗑️): remove an existing workflow.
- Edit a workflow: edit the values of an existing workflow.

⚠️**Only the group supervisor can add new workflows, edit, duplicate or remove the existing ones**. Any user can view the workflow details and use the workflows for launching a new analysis.

2.2.2 Create a new workflow

As the group supervisor, you can create a new workflow.

Click on “Add new” to create a new workflow, then select the input file type (FASTQ or VCF).



You can start filling the workflow fields. The workflow form is divided into two main parts.

Main analysis parameters

The first part correspond to the main analysis parameters:

- **Name:** workflow name. You can not use a workflow name if it already exists.
- **Analysis type:** select the type of the analysis from the drop-down list. This field can not be changed once the workflow has been created.
- **Assay**: select the capture, amplicon, whole genome library preparation method or kit corresponding to your analysis. This field is required if you are starting your analyses from FASTQ files. The assay's details will be used to calculate the coverage of the coding regions included in the kit. This information will be shown later in the Quality Control (QC) Report. If you are creating a VCF workflow this field is optional and has no effect other than having the assay mentioned in the QC report.
- **Keep variants** (only in FASTQ workflows)**:** Do you really want to see all the variants? If you also need the variants that did not pass the quality filters, then you should choose "Variant list will contain all variants". By default, only the high quality variants will be reported. Selecting "Variant list will contain all variants" will increase the amount of annotated variants and it could slow down the analysis.
- **Sequencer** (only in FASTQ workflows)**:** Select the sequencing technology used to obtain the FASTQ files. Two available options: Illumina or MGI.
- **Targeted mode** (only in FASTQ workflows)**:**
    - **ON:** this mode is used to limit the results to the assay's target regions. In order to achieve this, the variant caller will discard any reads in the alignment BAM file that do not overlap with the assay's regions. If at least one base of a read overlaps with one of the targeted regions, then the read is kept and included in the analysis. If there is no overlap with any of the targeted regions, the read is discarded and not taken into account when calling the variants. The end result of this approach is a faster and more specific analysis. ⚠️ Any variant not falling in the target regions defined by the selected assay will not be called when using the targeted mode.
    - **OFF (untargeted mode):** all variants will be reported regardless of whether they fall within the target regions of the sequencing assay used. You may want to consider using the untargeted mode, especially for whole-exome data, since that

will ensure you see all variants that may be of interest, including those that may be just at the very edge of a target region (e.g. intronic variants).

- **Genome:** choose the reference genome the reads will be aligned to. For VCF-based analyses run against **hg19**, if the VCF file contains variants reported on "chrM" (the name of the mitochondrial sequence in the hg19 genome), then those will be annotated with respect to the **NC_001807.4** sequence, the original mitochondrial sequence of hg19. If the variant is instead reported on "MT" (the name of the mitochondrial sequence in the hg38 genome) then they will be annotated with respect to the **rCRS**. If the analysis is launched from FASTQ sample(s), using either hg19 or hg38, any mitochondrial sequences will be aligned to the standard mitochondrial genome (GenBank number: **NC_120920.1**), which is included in the hg38 human genome.

hg19: "chrM" = NC_001807.4, "MT" = NC_012920.1 (rCRS)

hg38: "chrM" = "MT" = NC_012920.1 (rCRS)

- **Inheritance from:** which sources you want to use for the assessment of the mode of inheritance of a gene, which consequently affect the Germline Variant Classification. "All" includes OMIM®, CGD, GenCC, Gene2Phenotype, PanelApp and Clingen disease validity and "only OMIM®" includes only data from OMIM®. In case no information about the mode of inheritance is available in the aforementioned databases, the mode of inheritance is selected from DOMINO.
  Please note that PM2, BS2 and BP1 rules may change between both modes of annotation because the ACMG implementation uses different thresholds to evaluate these rules depending on the mode of inheritance.
- **Ethnicity:** you can optionally specify one of the ethnicities proposed by Gnomad. When selected, the values shown in the "Frequency" column of the variant table will correspond to the provided ethnicity. When the ethnicity is set to "Not specified" we will display the overall frequency.

Filters and sub-analysis options

Once filled with the above information, you can optionally fill the bottom section. It contains fields related to filters and sub-analyses.

- **Filters**: which (dynamic) filter(s) you wish to apply to the results. When provided, the dynamic filter will be automatically added to the variant table. Please remember that you will need to click on "Apply" to apply the filter in the variant table and get the filtered variants, find more details in section Create a Filter Set.
- **Run a gene list sub-analysis**: you can provide one or multiple gene lists to filter the results. When provided, a new sub-analysis will be run on top of the main analysis where only variants falling in genes belonging to the list will be shown in the table. If multiple lists are provided, we will create a combined gene list containing all genes included in each of the lists. This option is not currently available for tumor-normal workflows.
- **Run a gene list sub-analysis based on sample phenotypes**: when selected, a gene list will be built using the genes associated with the sample phenotypes. You must choose to filter by genes associated with all the sample phenotypes or with any of them. This

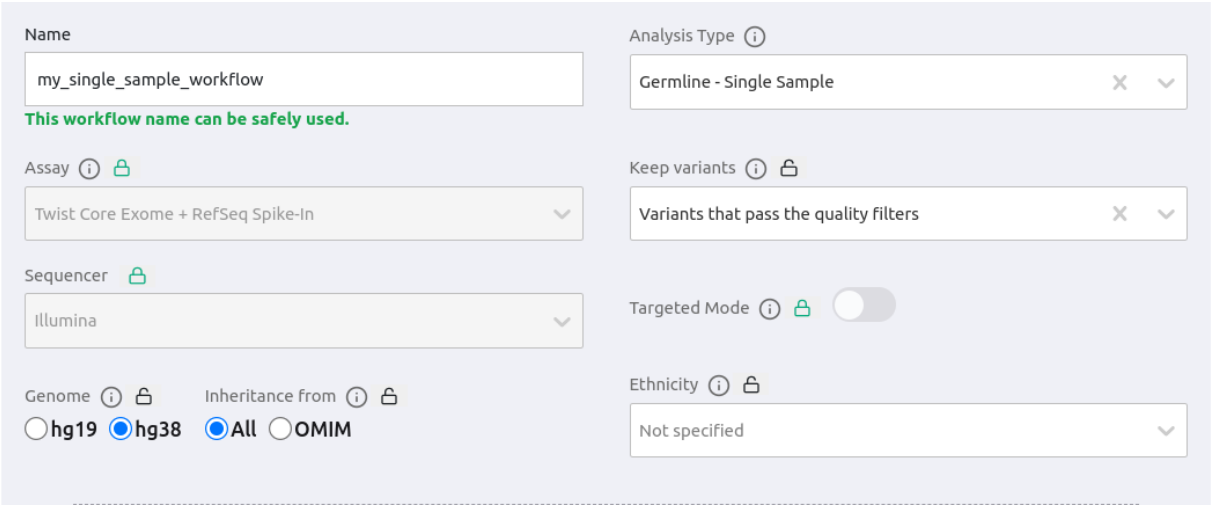
option is not available for somatic analysis. Please note that if you provide samples without phenotypes, this sub-analysis won't run as there won't be any phenotypes to create the gene list.

- **Show only gene list sub-analysis results**: when selected, the main analysis results will be hidden by default and we will only display the gene list sub-analysis results.
- **Run algorithmic filter sub-analysis**: select one or multiple algorithm filters from the list to run a sub-analysis. Each algorithm filter will be run as a single sub-analysis. Click on "Options" to modify the filter parameters.

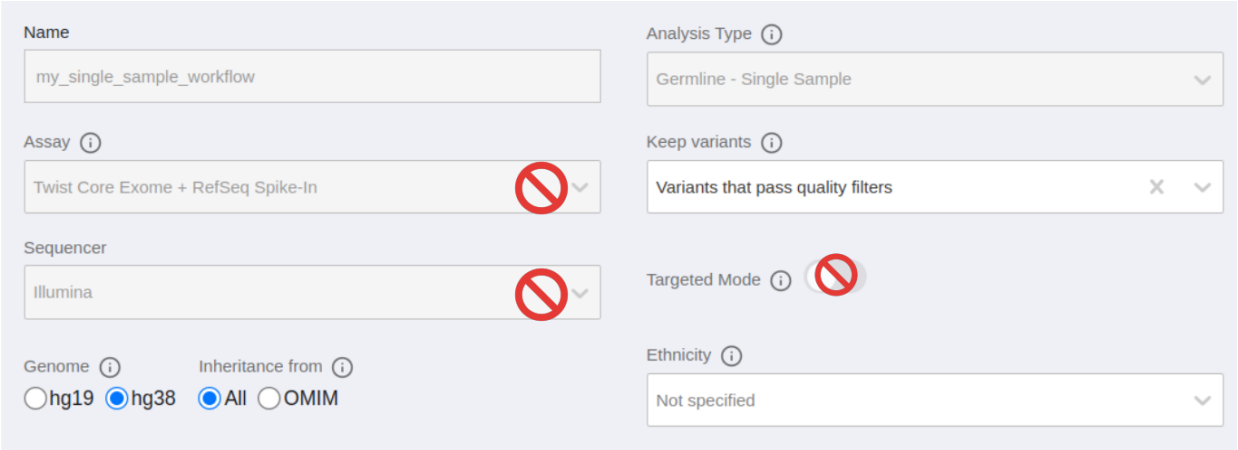
Lock workflow values

As the group supervisor, you will be able to create workflows that can be used by you or any other member of your group. While creating the workflow, you can decide if any of the fields has a fixed or an open value that can be changed by the person launching the analysis later. This is done using the lock icon (🔒).

When a field is set as open, it will have the default value set by the supervisor when creating the workflow, however this field can later be modified by the user using the workflow for the analysis.
If the lock icon is closed, then this field will have a fixed value that can not be changed when someone uses this workflow. This option is available in all workflow fields except the workflow name and analysis type which can not be changed when the user is using the workflow.



For example, in the picture above, we are creating a workflow where the values assay, sequencer and targeted mode are locked. The other values, such as genome, will have a default option (e.g. hg38) but can be changed when this workflow is selected for an analysis.



3. Launch analysis

- New analysis
- New cohort analysis

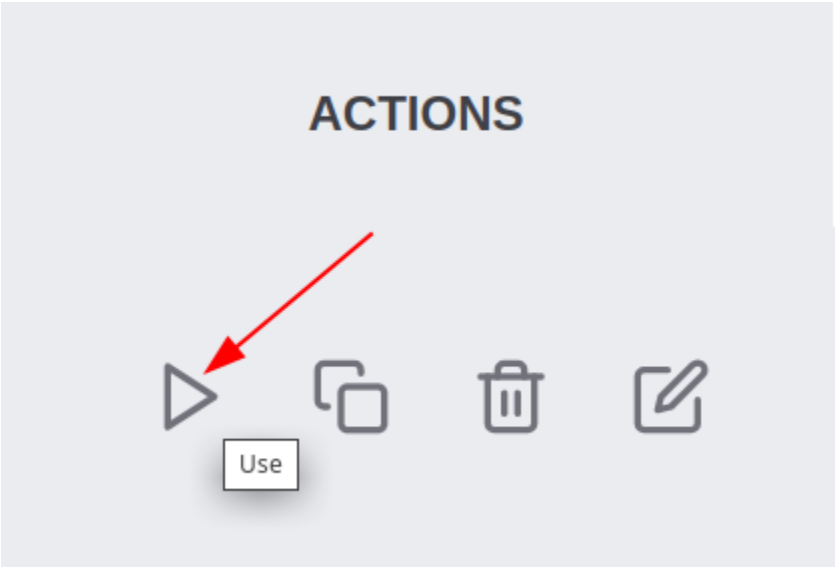
3.1 New analysis

You need to create your samples before launching a new analysis. If you have not defined your samples yet, please go to the Sample Definition section to find more information.

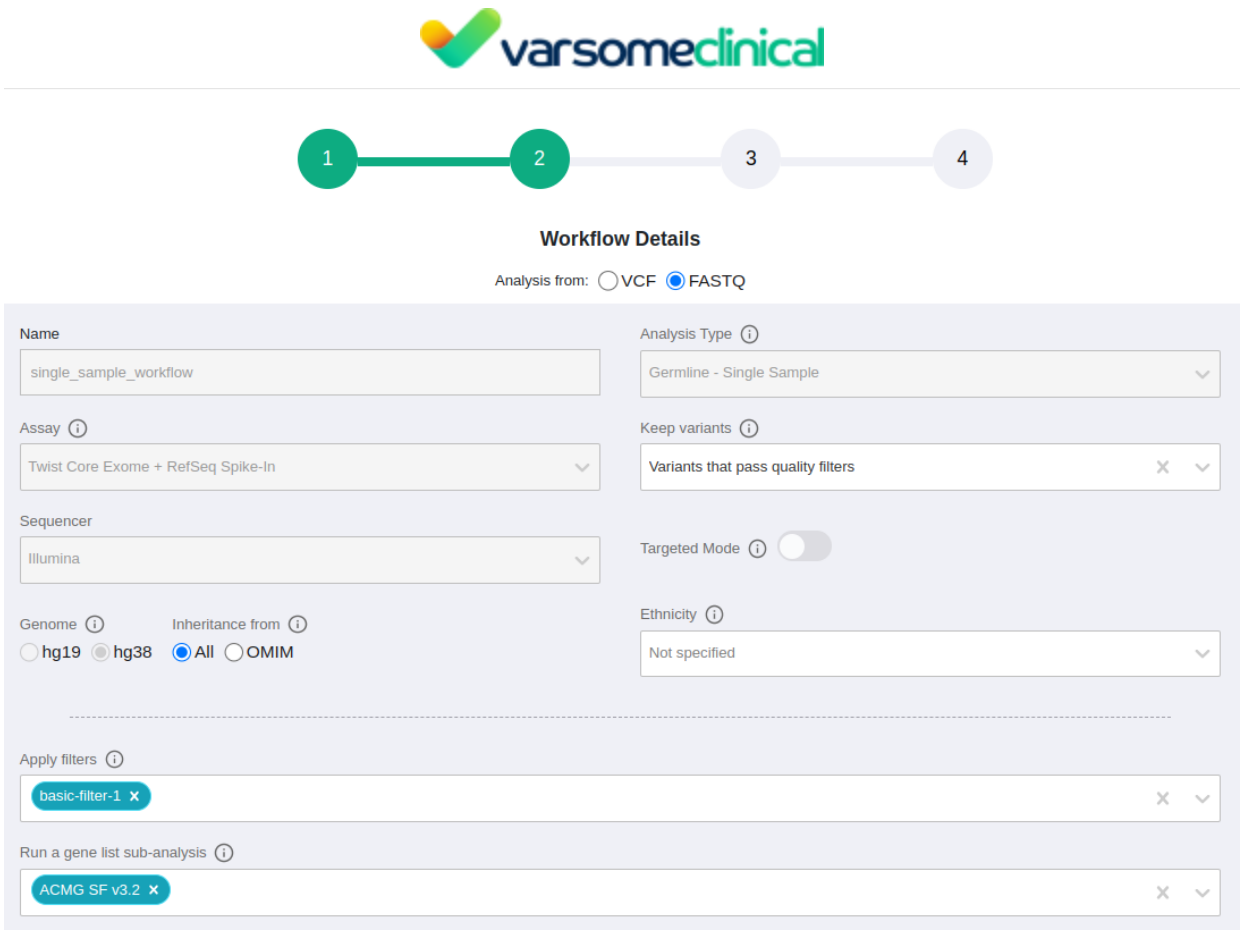
3.1.1 Use an existing workflow

You can start a new analysis from an existing workflow:

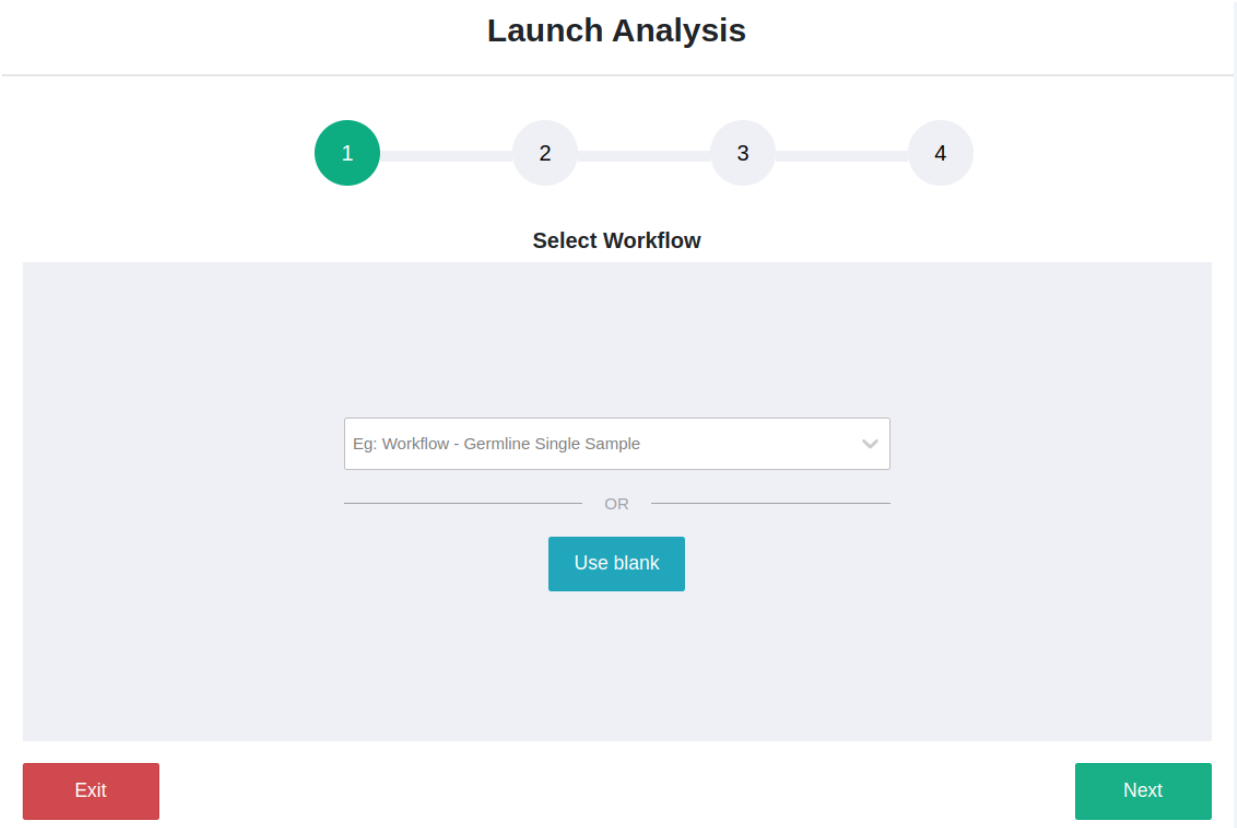
- Workflows table: click on the "Use" option (▶️) from the "Actions" column of the workflow of interest to launch an analysis using the selected workflow.



This option will take you to the "Launch analysis" menu where all parameters of the selected workflow are already loaded.



- Launch analysis > New analysis



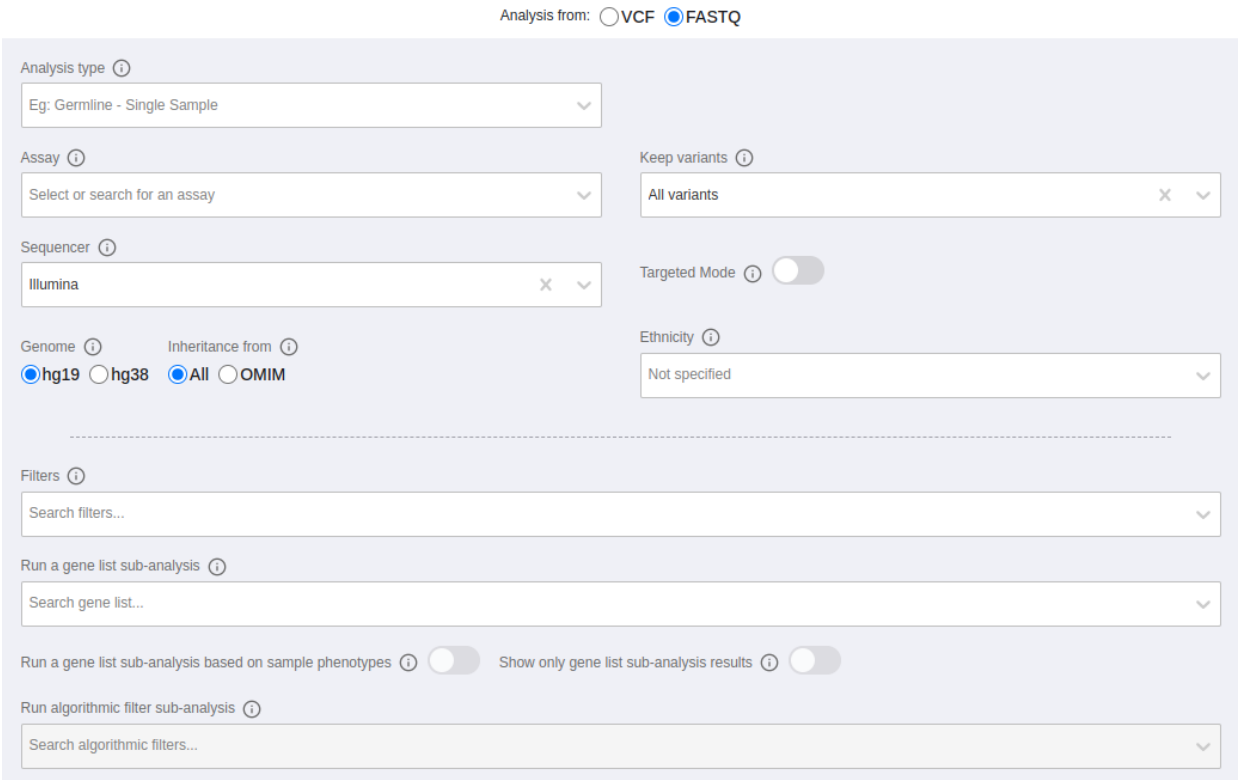
Here you can select a workflow from the available list of workflows shown in the drop-down list. When selected, all workflow parameters will be loaded.

⚠️ Please note that when launching an analysis using an existing workflow, if you modify any of the default workflow parameters defined by your group supervisor (unlocked fields) this won't result in the modification of the stored workflow. This change only applies to the current analysis.

3.1.2 Launch an analysis without a workflow

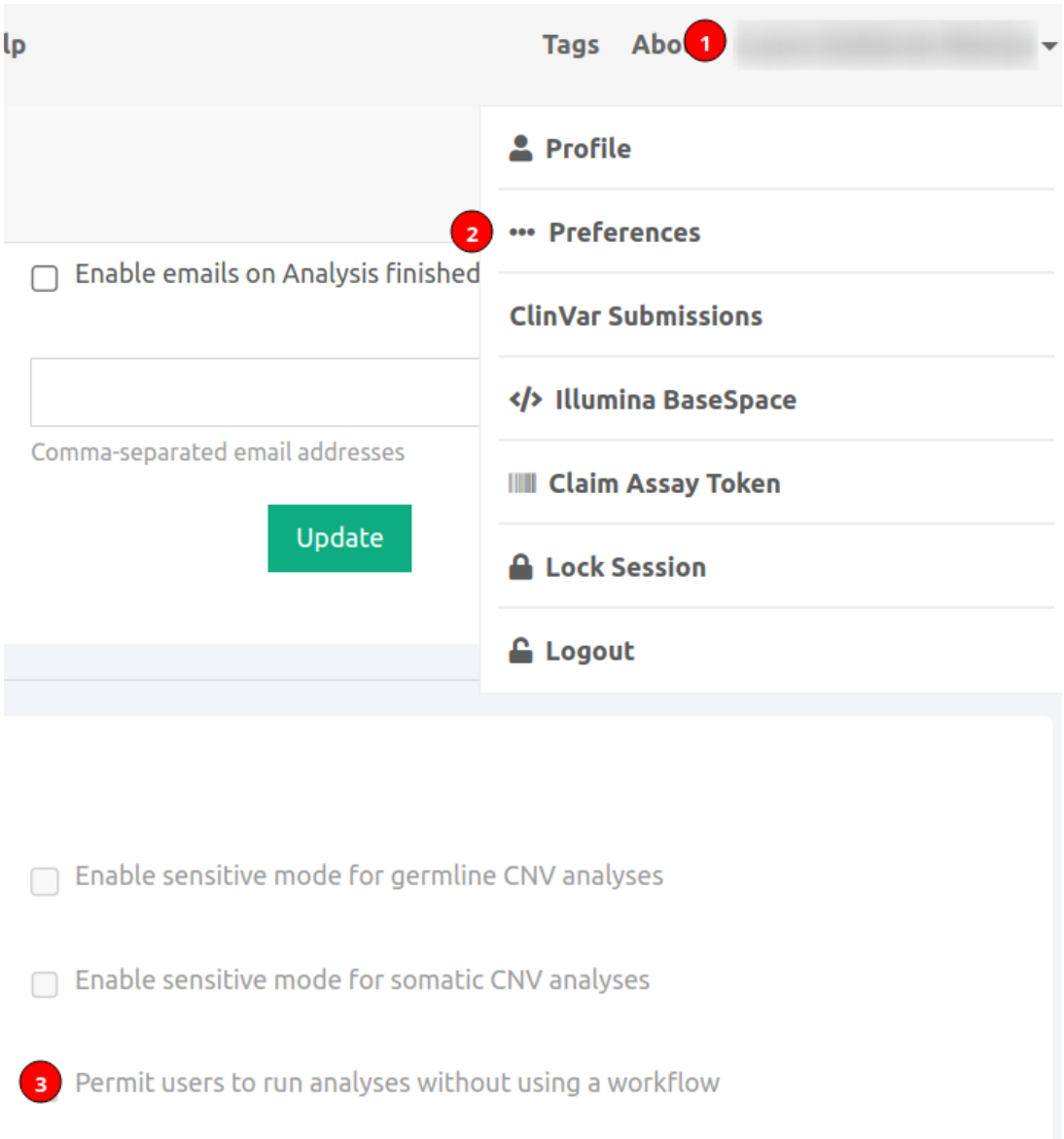
This option can be used to define all analysis parameters without using an existing workflow. Please note that the **analysis parameters defined when selecting this option option won't be saved as a workflow for future analyses**. It will be used only for the analysis launched at that moment. If you wish to create a workflow and use it for future analyses please check the Create a new Workflow section.

To use this option, go to "Launch analysis" > "New analysis" and click on "Launch an analysis without a workflow". You will need to fill in all the details before selecting the samples. All fields (except the workflow name which is not applicable here) are similar to those presented when a workflow is created. Please follow the instructions given in the Create a new Workflow section to define the analysis parameters.



How to limit the "Launch an analysis without a workflow" option:

By default all users of your group are able to launch analysis and define the analyses parameters themselves. As a group supervisor, if you would like that your colleagues only use the workflows you have created, you can do it by going to "Preferences" and unchecking the box "Permit users to run analyses without using workflows".

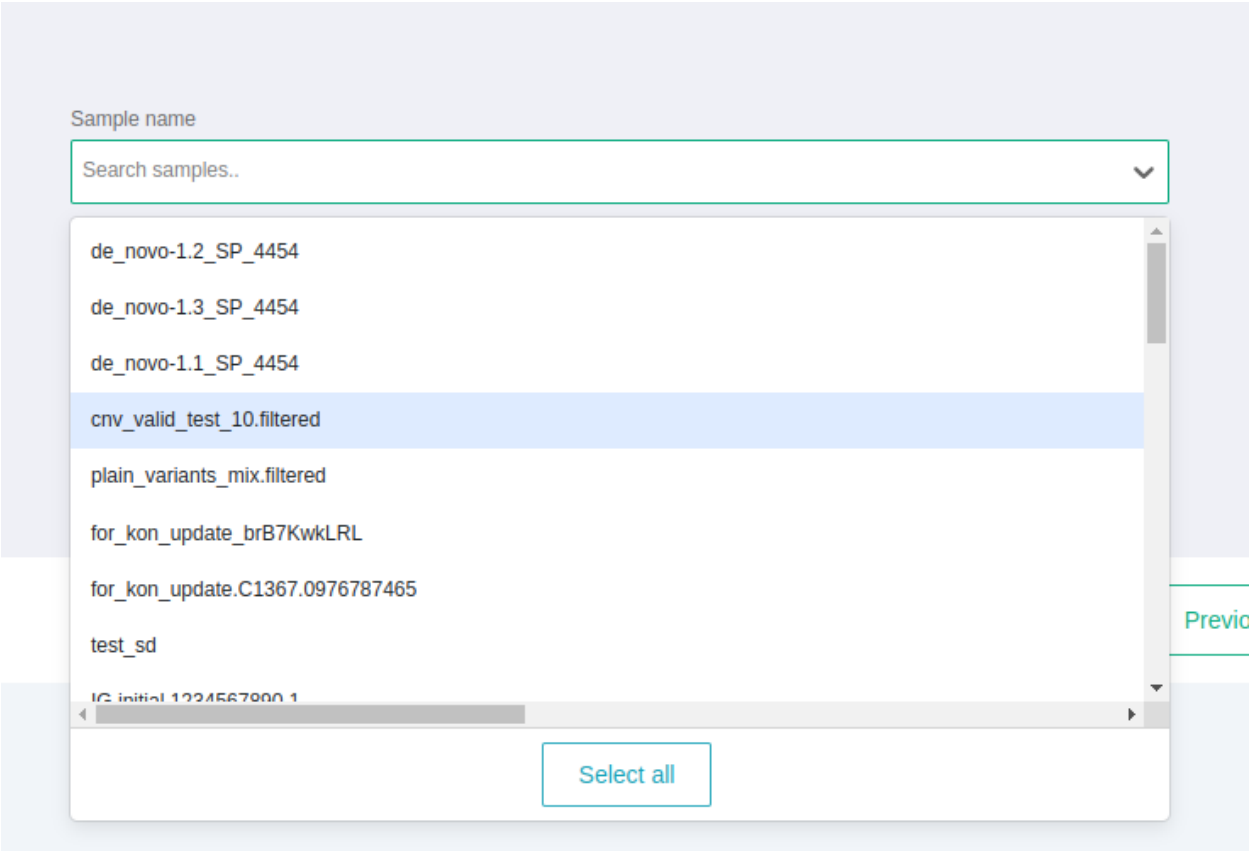


3.1.3 Select samples for analysis

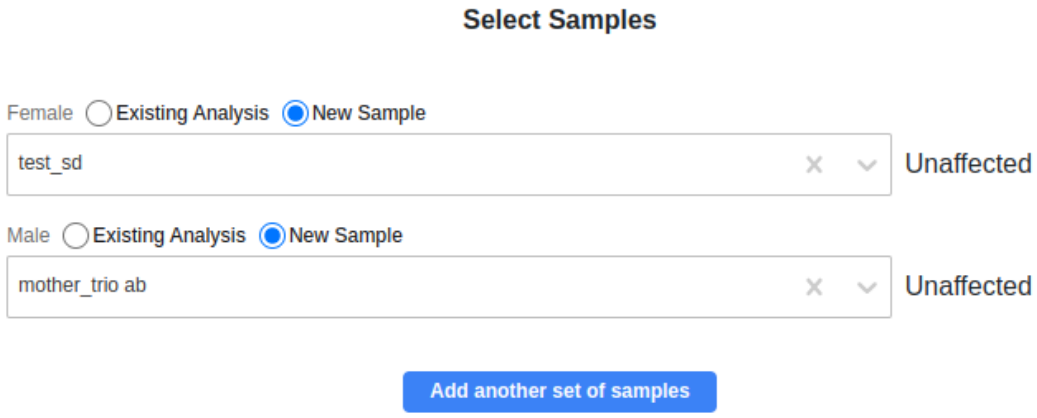
Once selected the analysis parameters, click on Next to select the samples. **Please remember that the samples need to be created before this step**. Go to the Create samples section for more information.

Once specified the analysis settings, you can select which samples you wish to analyze

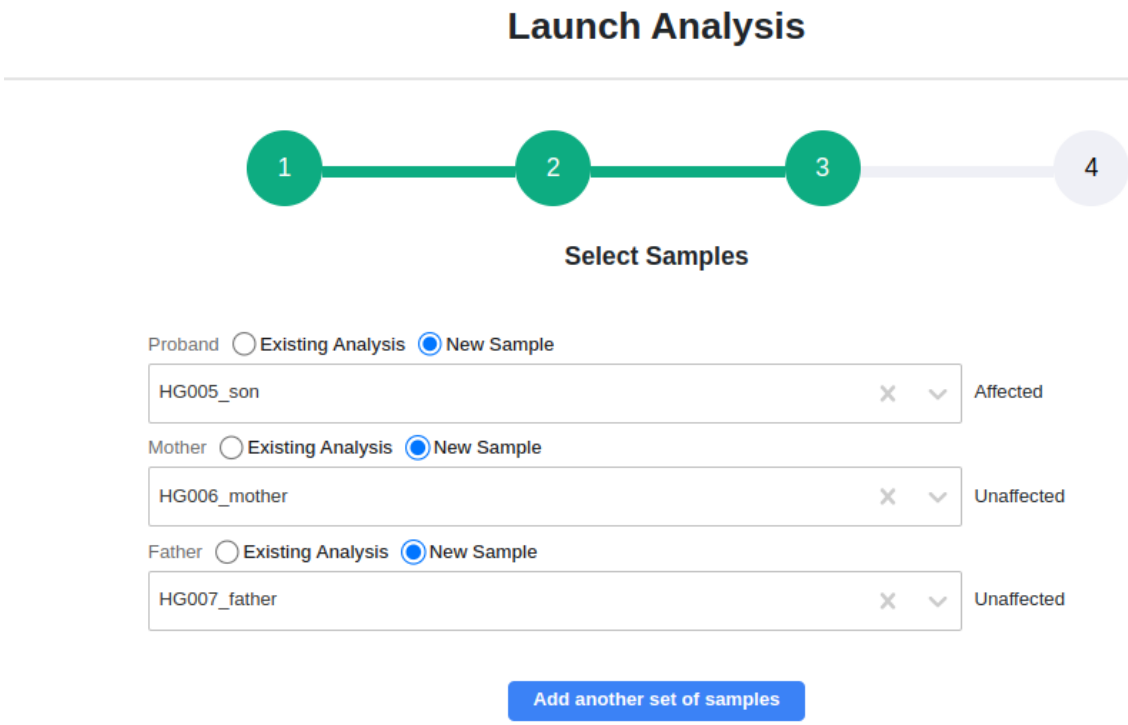
- **Single sample (germline or somatic): analysis of several independent samples**. Select the samples from the dropdown. All selected samples will be analyzed independently. Click on “Select all” to select all samples in one click.



- **Couple (for carrier risk analysis)**: analysis of two unaffected samples (male and female). Select one sample per field. Click on “Add another set of samples” to add another couple.

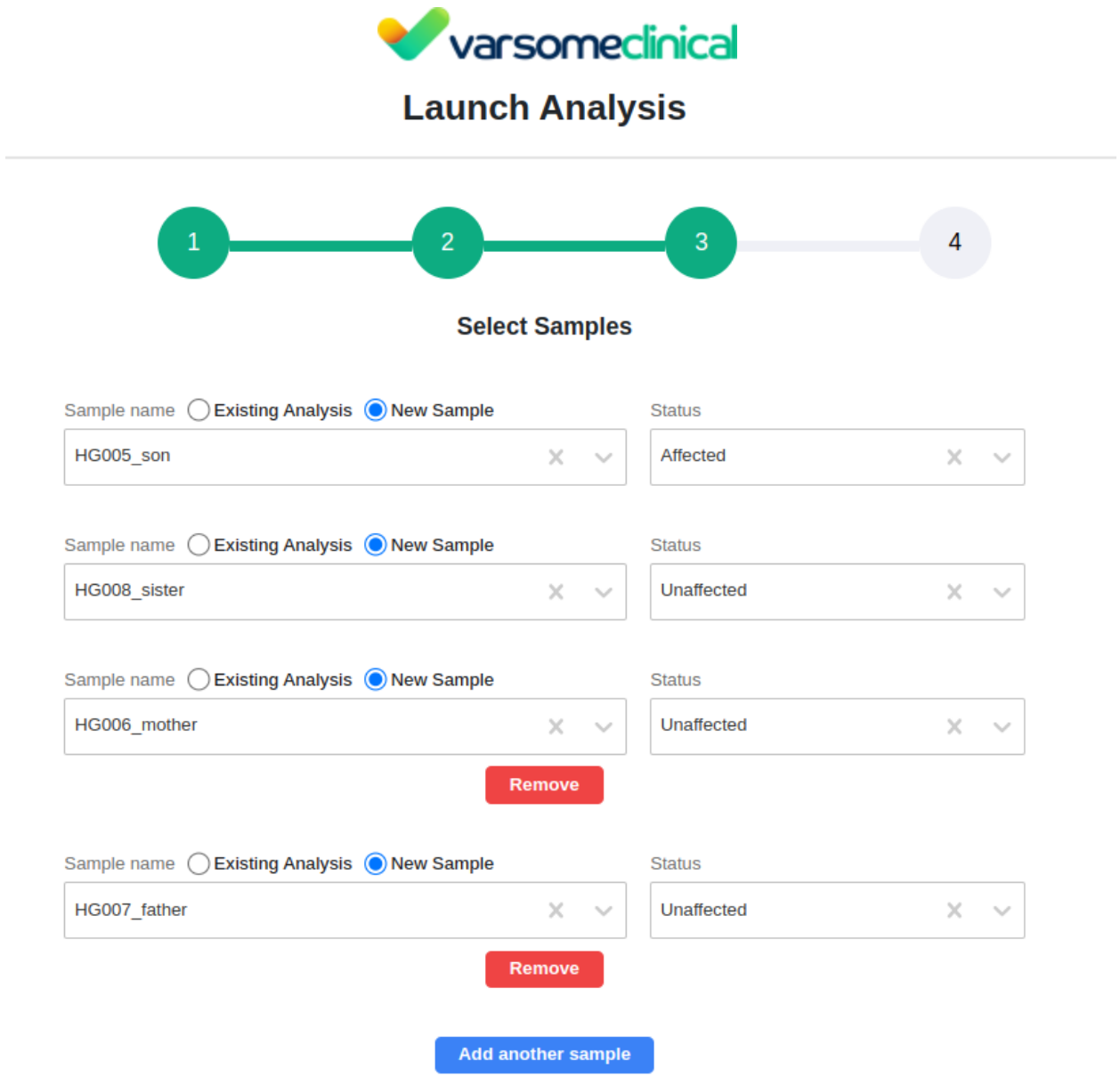


- **Family trio**: analysis of one affected child (proband) and two unaffected parents. Click on “Add another set of samples” to add another trio.

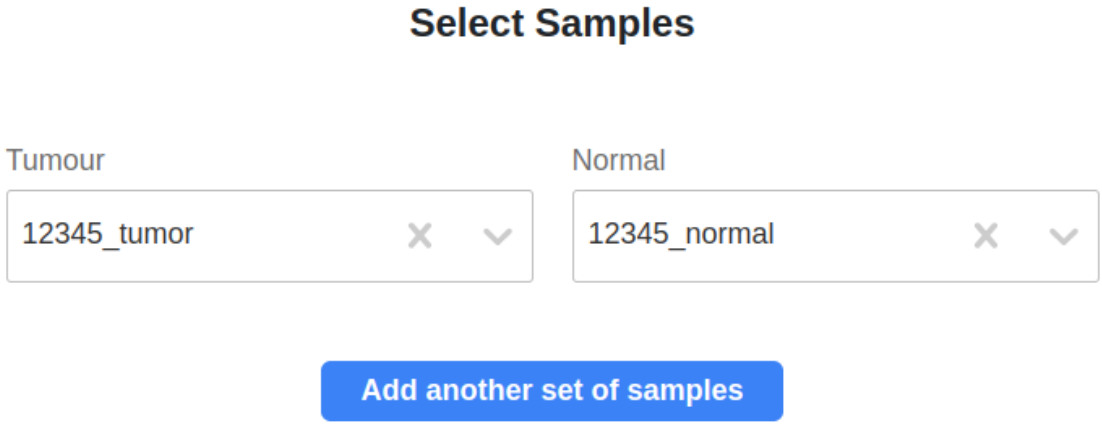


- **Multiple samples**: this is for multi-sample germline analysis that are not couple nor family trio (e.g. extended family like two parents and two siblings). You need to select the affected status for each member of the multi-sample analysis. Click on “add another sample” to add more samples to the analysis. Click on “remove” to remove any of the samples.

Note: If you wish to perform any type of multi-sample analysis with one or more samples already analyzed in VarSome Clinical, you don’t need to upload them again. For multi-sample analysis you can click on the “Existing analysis” radio button to browse and select samples already analyzed in VarSome Clinical.

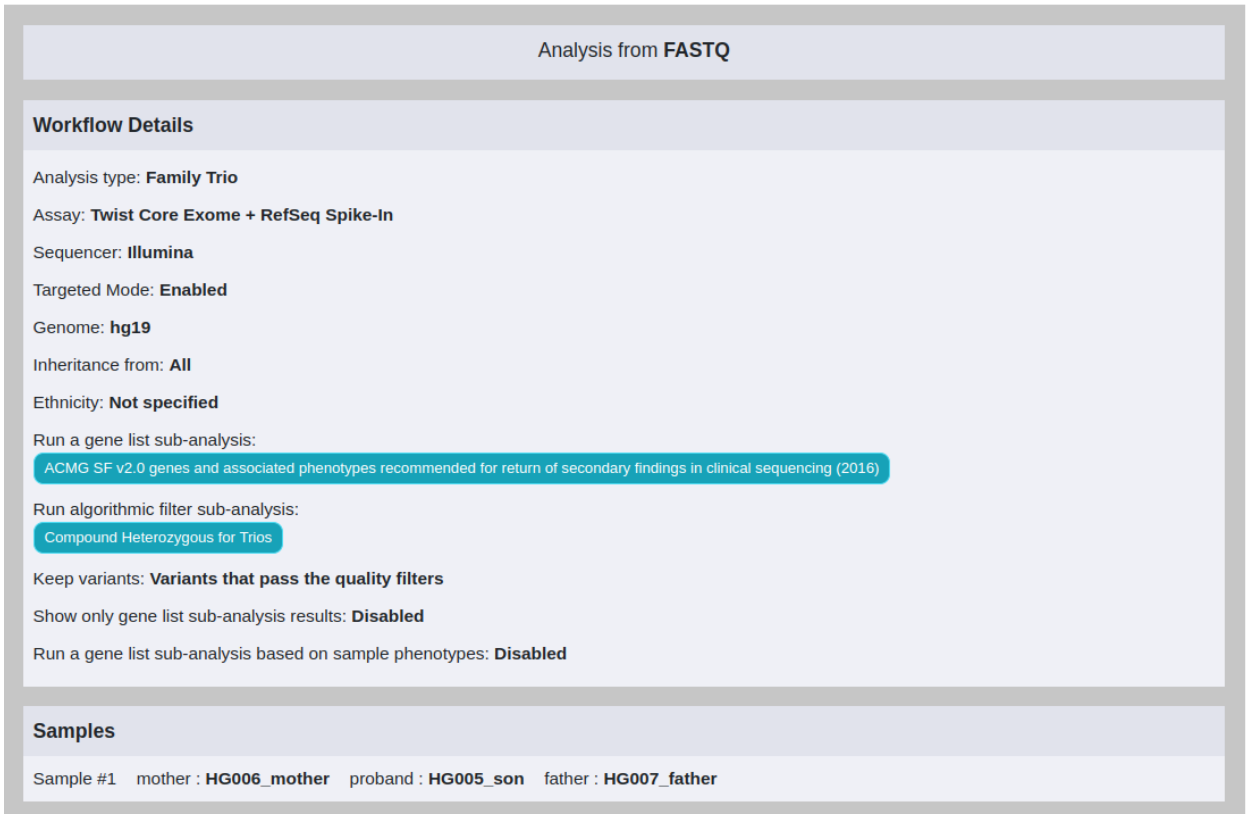


- **Tumor-normal** (only for analysis starting from FASTQ): one tumor sample and its matched normal. Please note that you have defined your samples previously (the tumor one as somatic and the normal one as germline). Click on "Add another set of samples" to add another tumor-normal pair of samples for the analysis.



3.1.4 Preview and launch analysis

In this step you will see a summary of the analysis settings (workflow details) and selected samples. You can also add tags to the analysis by clicking on the Tags drop-down and select one of the available tags.



You can click on “Previous” to go back to any of the previous steps and modify any of the analysis parameters.

3.2 New cohort analysis (multi-sample VCF)

If you have a multi-sample VCF file with more than 30 samples inside, you should go to this page in order to launch the analysis.

Sample information (left side of the menu)**:**

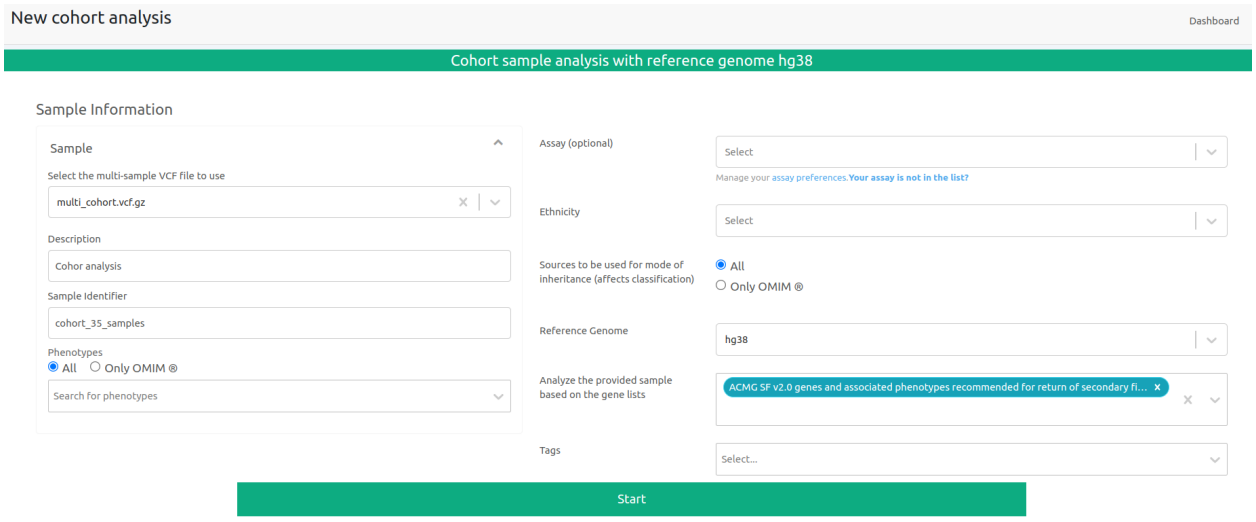
- Select the multi-sample VCF file to use
- Description: optional description
- Sample Identifier: user sample name
- Phenotypes: optionally provide phenotypes to the cohort analysis

Analysis information (right side of the menu)**:**

- Assay (optional)
- Ethnicity: ethnicity of the samples of the cohort
- Sources to be used for mode of inheritance: all or only OMIM
- Reference genome: reference genome version used to obtain the VCF (hg19 or hg38).

- Analyze the provided sample based on the gene lists: provide a gene list to run a gene list sub-analysis once the main analysis is done.
- Tags: provide tags to label your samples.

Click on “Start” to launch the analysis.



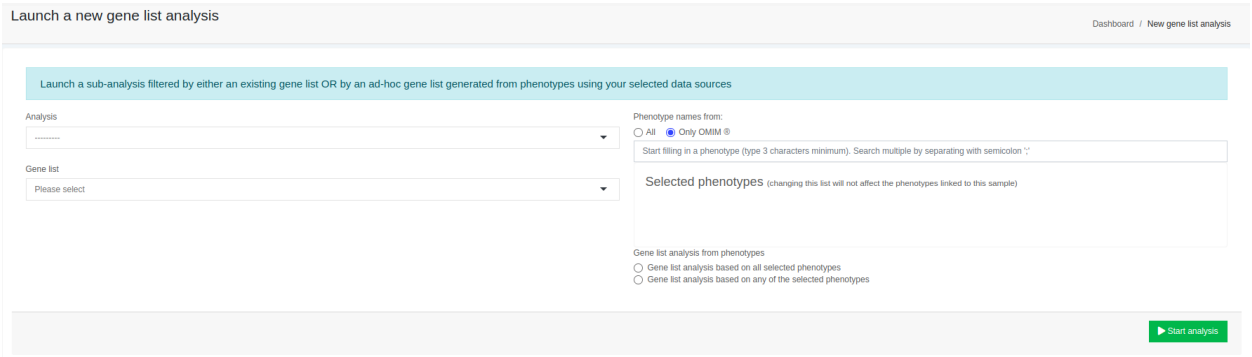
4. Launch sub-analysis

- Gene list analysis
- Algorithmic filters
- CNV/SV analysis
- Repeat expansion annotation from VCF file

4.1 Gene list analysis

Note: The methodology to generate a gene list from phenotype(s) has been changed. Previously, when making a gene list from phenotypes, we would include only those genes that are directly annotated with that phenotype. We have now extended this and instead first collect all diseases linked to the phenotype and then all genes linked to those diseases, as well as any genes directly linked to the phenotype. We already worked this way when adding phenotypes to analyses, so this change ensures we are consistent and also makes sure we don't miss any genes when creating gene lists.

Filter the results of a finished analysis to show only those variants falling within (including 500 base pairs up and downstream) the genes given.

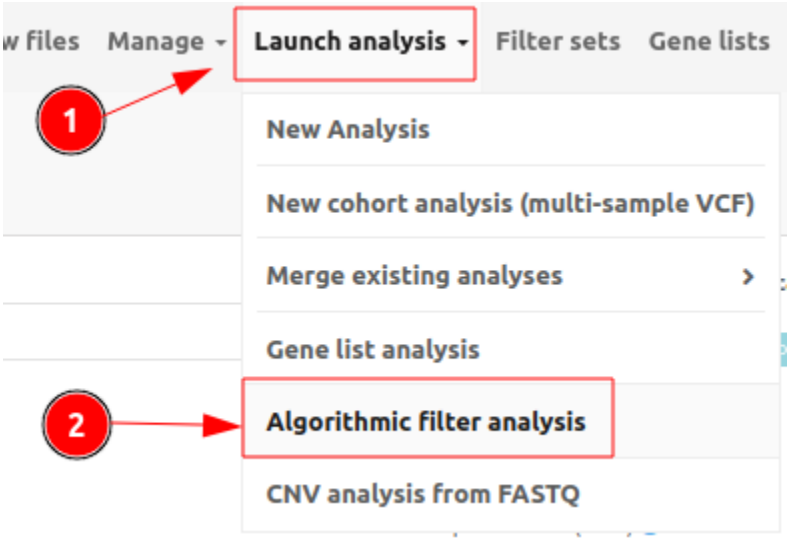


- **Analysis:** choose the analysis to be filtered.
- **Gene List:** choose a gene list to filter by.
- **Phenotypes:** choose a list of phenotypes. A gene list will be built using the genes associated with the chosen phenotypes. When filtering by phenotype, you can also choose to filter by genes associated with *all* the phenotypes or with *any* of them.

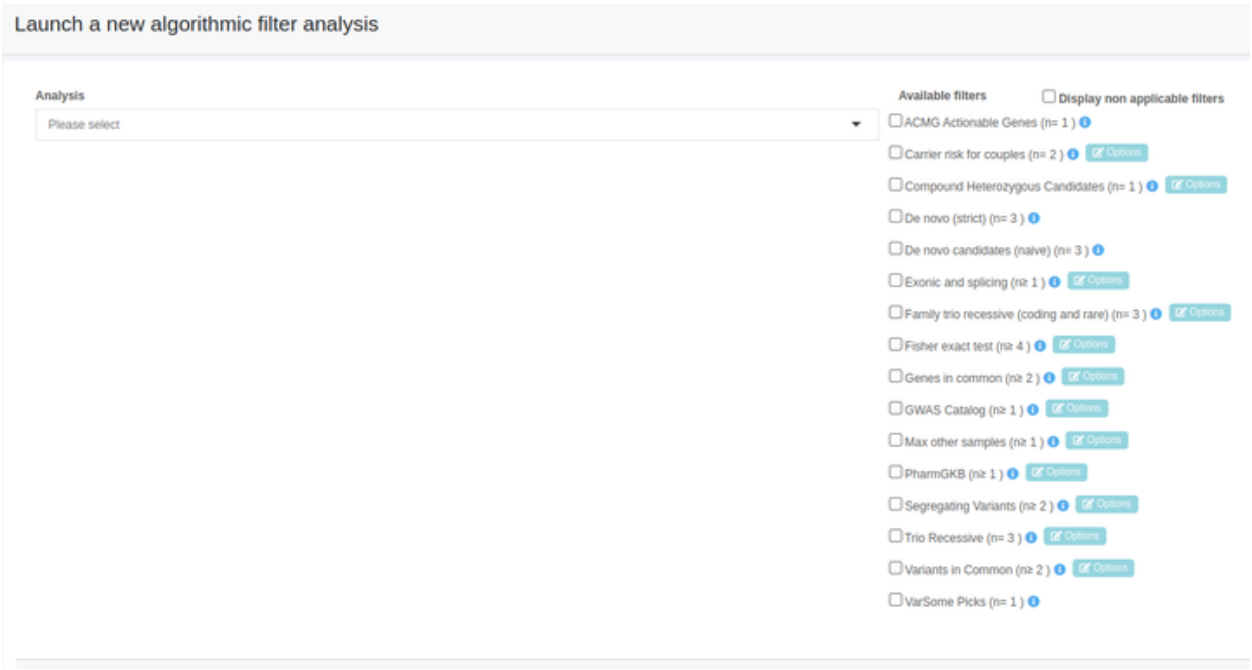
4.2 Algorithmic filters

You can apply sophisticated filters to perform more complex variant filtering to finished analyses. For several of the existing filters we now provide the option to change specific parameters that used to be fixed. That way you can customize each of those filters according to your needs.

You can access Algorithmic filters under Launch analysis > Algorithmic filter analysis as shown in the picture below.



To start an algorithmic filter, choose the target analysis, select an algorithmic filter, and then click on **Start** analysis.

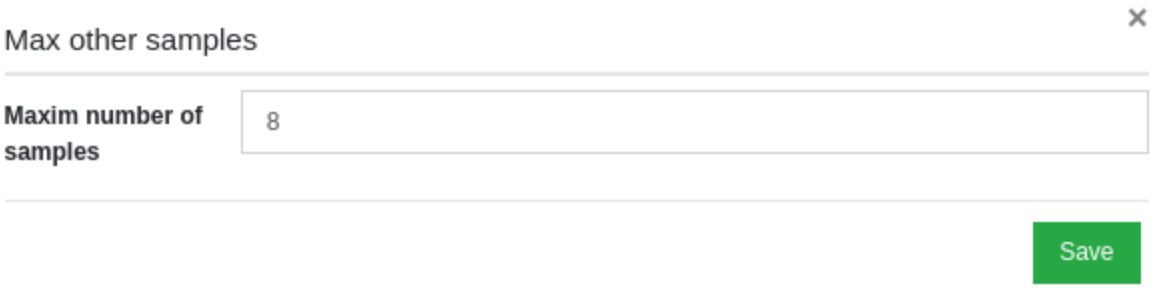


For further details on the algorithmic filter and the provided parameters please click on the info icon next to the “Options” button.

As you can see for several of the existing filters, we provide the option to change specific parameters. That way you can customize each of those filters according to your needs.

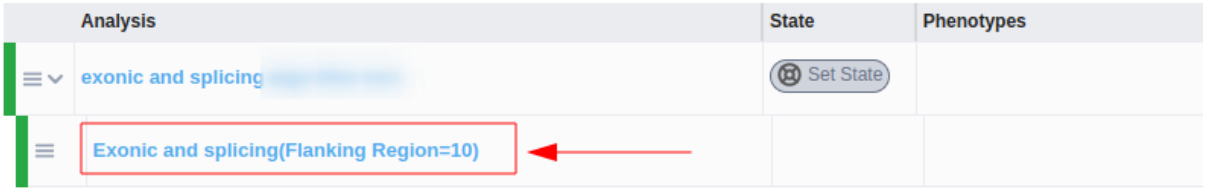
If you want to change a parameter click on the light blue “Options” button:

For example, if you want to change the default parameters for the filter "Max other samples (n≥1)", after you choose from the drop-down the analysis to which you wish to apply one or more filters, you then need to select the filter "Max other samples (n≥1)" by clicking the relevant check-box, which will in turn activate the Options box.

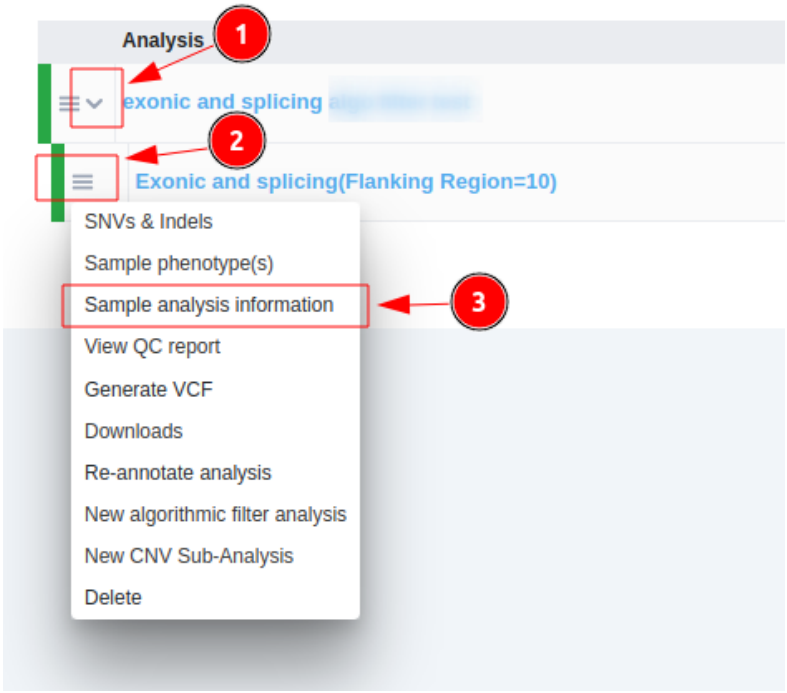


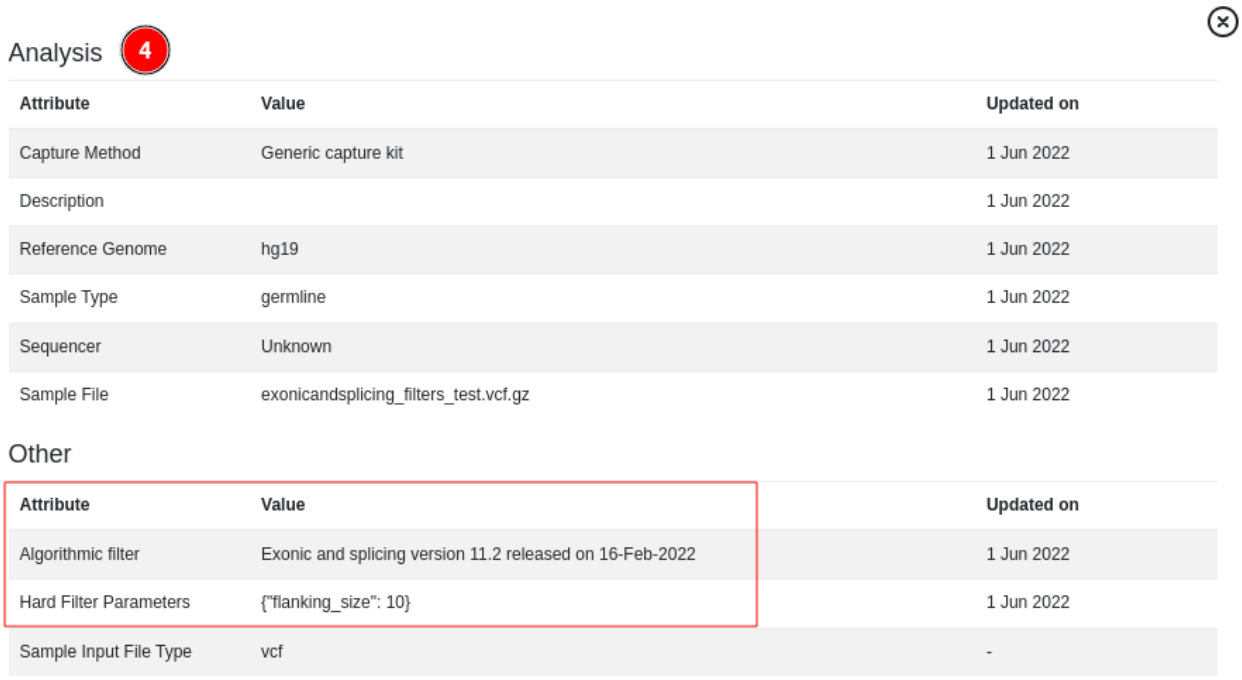
Now you can change the default value for this filter, by changing the number in the *Maxim number of samples* field and then by clicking on "Save".

Once the analysis has run, you can view the filter options selected for the algorithmic filter analysis both in the sub-analysis name



and in the “Sample analysis information” option of the “Analysis actions” menu:





Analysis 4

Attribute	Value	Updated on
Capture Method	Generic capture kit	1 Jun 2022
Description		1 Jun 2022
Reference Genome	hg19	1 Jun 2022
Sample Type	germline	1 Jun 2022
Sequencer	Unknown	1 Jun 2022
Sample File	exonicandsplicing_filters_test.vcf.gz	1 Jun 2022

Other

Attribute	Value	Updated on
Algorithmic filter	Exonic and splicing version 11.2 released on 16-Feb-2022	1 Jun 2022
Hard Filter Parameters	{"flanking_size": 10}	1 Jun 2022
Sample Input File Type	vcf	-

Available Algorithmic Filters

4.2.1 ACMG Actionable Genes

Variants meeting the criteria described in ACMG SF v2.0: Kalia et al., Genetics in Medicine, (2017) 19, 249–255 (doi:10.1038/gim.2016.190).

4.2.2 Carrier risk for couple

Pathogenic variants for which both members of a healthy couple are heterozygous and so can be passed on to any offspring. Included are genomic (not mitochondrial) variants that are classed as pathogenic, likely pathogenic or of uncertain significance, are not homozygous in either prospective parent and either:

- Both individuals are heterozygous for the variant
- The variant is on the X chromosome of the mother
- The variant falls in a gene that has at least two variants where one is present and heterozygous only in one individual and the other is present and heterozygous only in the other individual. In such cases, all pathogenic, heterozygous variants for that gene are shown as they are candidates for compound heterozygosity.

Options:

- **Strong VUS:** If selected, keep variants of Unknown Significance (but only if one of the strong pathogenic Germline rules has fired for this variant). VUS variants qualify only if they trigger one of the Germline rules: PVS*, PS* or PP5. If not selected, keep all VUS variants (irrespective of Germline rules).

Strong rules are PVS1, PS1, PS2, PS3, PS4 and PP5. In addition, any rule whose strength has been raised to "strong", "very strong" or "stand alone" will be considered a "strong" rule, even if its original strength as per the Germline guidelines was lower. For example, PM1 can be raised from "supporting" to "strong" if the variant is located in a dense hot-spot.

⚠️ Please note that we recommend that you further filter by genes with recessive mode of inheritance or loss-of-function pathogenicity.

4.2.3 Compound Heterozygous Candidates

Variants classified as pathogenic, likely pathogenic or of unknown significance for which all of the following apply:

- they are heterozygous variants in genes that carry at least one other heterozygous variant and no homozygous pathogenic variants.
- are not in mitochondria

Options:

- **Strong VUS**: If selected, keep variants of Unknown Significance (but only if one of the strong pathogenic Germline rules has fired for this variant). VUS variants qualify only if they trigger one of the Germline rules: PVS*, PS* or PP5. If not selected, keep all VUS variants (irrespective of Germline rules).
  Strong rules are PVS1, PS1, PS2, PS3, PS4 and PP5. In addition, any rule whose strength has been raised to "strong", "very strong" or "stand alone" will be considered a "strong" rule, even if its original strength as per the Germline guidelines was lower. For example, PM1 can be raised from "supporting" to "strong" if the variant is located in a dense hot-spot.
- **Homozygous also**: If selected it will also filter for variants which are homozygous.
- **Phased Mode**: If selected, it will only filter those variants with phasing information to identify compound heterozygous variant pairs in the same gene and phasing group and in different zygosities (1|0 vs 0|1).

4.2.4 Compound Heterozygous for Trios (n=3)

This filter is aimed to support the identification of compound heterozygous variants in an affected child when the genome of the two unaffected parents is also provided (Family Trio analysis (n=3)).

The filter will keep variants in the child classified as pathogenic, likely pathogenic or of unknown significance for which all of the following apply:

- they are heterozygous variants in genes that carry at least one other heterozygous variant and no homozygous pathogenic variants.
- are not in mitochondria
- If such variant pairs are detected, it will look for compound heterozygous pairs in the same gene in each of the parents.
- If either parent does have a compound heterozygous pair in the same gene (not necessarily the same pair as the child), then we discard the pair identified in step a, and

move to the next candidate. The assumption is that parents are unaffected, so the child would also be unaffected if one of the parents have a compound heterozygous pair in the same gene.

Options:

- Use phasing information (**phased mode**). When this option is enabled the filter will identify compound heterozygous variant pairs in the child—that is heterozygous variants in the same gene and phasing group and in different zygosities (1|0 vs 0|1).
- **Strong VUS**: If selected, keep variants of Unknown Significance (but only if one of the strong pathogenic ACMG rules has fired for this variant). VUS variants qualify only if they trigger one of the ACMG rules: PVS*, PS* or PP5. If not selected, keep all VUS variants (irrespective of ACMG rules). Strong rules are PVS1, PS1, PS2, PS3, PS4 and PP5. In addition, any rule whose strength has been raised to "strong", "very strong" or "stand alone" will be considered a "strong" rule, even if its original strength as per the ACMG guidelines was lower. For example, PM1 can be raised from "supporting" to "strong" if the variant is located in a dense hot-spot.
- **Homozygous also**: If selected it will also filter for variants which are homozygous.

4.2.5 De novo (strict)

Variants present in the proband and absent in both parents, where neither parent has any reads supporting the variant but only counting positions where the parents have a minimum coverage of 8.

4.2.6 De novo candidates (naive)

Variants likely to have arisen in the child from unaffected parents. Variants must meet either of the following conditions:

- the child is homozygous for the variant, but the variant is only present in one parent OR
- the variant is present in the child, but not present in either parent.

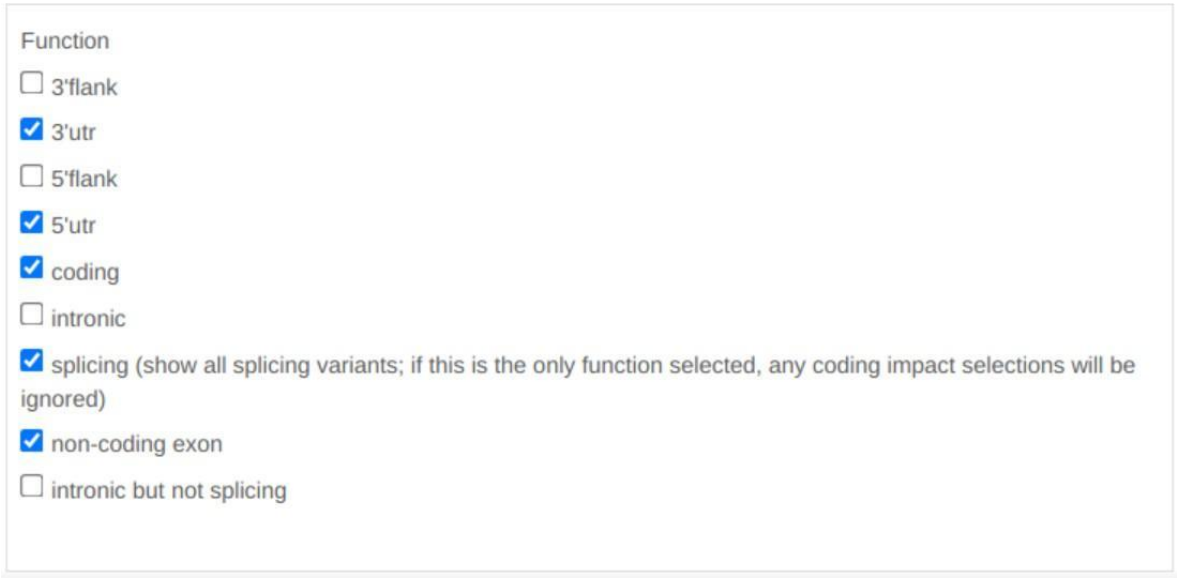
⚠️ Please note that we recommend that you further filter for pathogenicity, coverage, frequency, and mode of inheritance.

4.2.7 Exonic and splicing

A special case of algorithmic filter is “Exonic and splicing” which only keeps exonic (including UTR and other non-coding exons) and splicing (no more than 10 nucleotides from a known splice site) variants. This filter comes with the option to change the distance from a known splice site according to your needs.
The filter may be launched manually, just like any other, but it will also be run automatically on any analyses with more than 500 000 variants.

This algorithmic filter provides the same results as would occur if you made a dynamic filter with the following:
Coding

Splicing
Non-coding exon +3' utr
Non-coding exon +5' utr



The aim of this filter, and the reason it will run automatically for large analyses, is to provide a smaller subset of results to the user which will be far quicker and easier to sort through. Since, even with WGS analyses, the variants of interest tend to be those that can affect the protein sequence, we feel that this filter will help our users quickly identify and focus on the variants of interest even on larger samples such as WGS.

⚠️ Please note that while the filter will be run automatically for such large analyses, the full result set will still be available as usual. The filter will run as a sub-analysis and will not affect the results of the main, parent analysis in any way.

3’ and 5’ flank
The 3' and 5' flank refer to the transcript(s). Using Dynamic filter for Function, VarSome Clinical will mark a variant as flanking if it is within 500nt of a transcript's start or end position. UTRs are part of the transcript, and such shouldn't be an issue.
If you have a particular set of regions, we could filter the result for those regions. Alternatively, you could run a gene list analysis, limiting the results to only the list of genes.
Another possibility, if this is something you will often be doing, is setting up an Algorithmic filter. Algorithmic filters are more sophisticated filters that run as separate analyses and there we can write one catering to whatever specific filtering you require. In that case, we could search for coding variants in all genes and also include any found in the upstream regions of the target genes.

splicing
Splicing information is used by our Germine Variant Classification implementation to decide whether some rules should be applied or to boost their strength.

We use the scSNV as well as MaxEntScan databases for splice-site prediction. This is only available for single-nucleotide variants. Variants above the 'ADA Boost Splicing' threshold (0.708) and 'Random Forest Splicing' threshold (0.515) are considered candidate splicing variants. The scSNV splice-site prediction is used in the following rules:
BP7: A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice site consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved. (Benign, Supporting). The splicing is checked as follows:
The variant is found more than 2 bases away from the next splice site.
It is not predicted splicing according to the scSNV database.
PP3: Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.) (Pathogenic, Supporting). We exceptionally boost the strength to 'Strong' if the variant is predicted splicing and rule PVS1 was not triggered.
Splicing is also considered in rule:
PVS1: null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multi exon deletion) in a gene where LOF is a known mechanism of disease. (Pathogenic, Very Strong).
In PVS1, intronic variants within 1 or 2 nucleotides from the exon of the transcript splice site are considered null variants as defined in this paper.

4.2.8 Family trio recessive (coding and rare)

The following criteria are applied in this order.

1. Keep only, either:
   a. Coding variants that are frameshift or missense or nonsense or stop-loss or exon deletion or in frame or start loss or splice junction loss
   b. Splicing variants (+/-10 bp from exon ends)
2. Remove variants where allelic balance < 0.2 or coverage <= 5
3. Remove variants present in both parents except if child is homozygous and both parents are heterozygous.
4. Remove variants with gnomAD population frequency over 1% based on the ethnicity of the proband. If ethnicity is not provided, the general population frequency is used.
5. Remove variants that are homozygous for the alternative allele in either parent.
6. Remove variants in mitochondria and chromosome Y.
7. Keep only variants where EITHER of these criteria apply:
   a. Variant is homozygous for the alternative allele in the child.
   b. Child is heterozygous and the following criteria BOTH apply:
      i. There are two or more variants in the same gene. To qualify, a variant must be in a coding transcript of a gene with a Transcript Support Level consistent with the sample's settings. The variant must be Pathogenic, Likely Pathogenic or of Uncertain Significance (VUS).
      ii. The variants did not all come from the same parent; some variants on the gene may have come from the mother and some from the father or are de

novo.

Options: All of the following fixed values originally set for the filter can now be changed:
• Maximum distance from splice site
• Minimum distance from splice site
• Minimum coverage
• Minimum allelic balance
• Maximum frequency

This filter specification was kindly contributed by Dr Erica Davis of Lurie Children's Hospital in 2021.

4.2.9 Fisher exact test

This filter will select any variants that are found more often (fisher exact test p-value <=0.05) in the affected samples than in the controls.

4.2.10 Genes in common

All variants classified as pathogenic, likely pathogenic or of unknown significance (but only if one of the strong pathogenic Germline rules has fired for this variant), which are found in genes with at least one such variant in all samples of a merged analysis.

The filter works in the following way:
- It will identify all genes with at least one pathogenic, likely pathogenic or VUS variant in all samples.
- It will then return ALL pathogenic, likely pathogenic or VUS variants falling in that set of genes.

Options:
- Pathogenic only: If this option is selected, variants classified as Benign(B) or Likely Benign(LB) or VUS/LB-B will be excluded from the results.

4.2.11 GWAS Catalog

All variants with GWAS Catalog data.

Options:
- Pathogenic only: If this option is selected, variants classified as Benign(B) or Likely Benign(LB) or VUS/LB-B will be excluded from the results.

4.2.12 Max other samples

Remove variants that exist in more than the selected number of other samples from your group, and those with more than 5% population frequency.

4.2.13 PharmGKB

Variants associated with genes that have PharmaGKB information. For more information see in the section Pharmacogenomics Knowledge Base (PharmGKB).

Options:

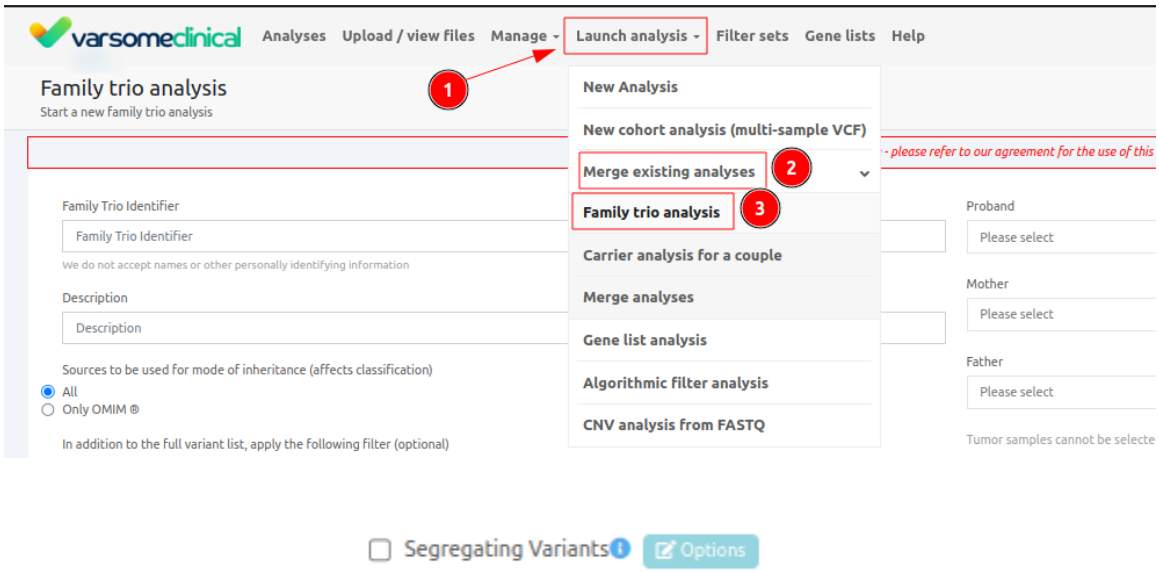
- Pathogenic only: If this option is selected, variants classified as Benign(B) or Likely Benign(LB) or VUS/LB-B will be excluded from the results.

4.2.14 Segregating variants

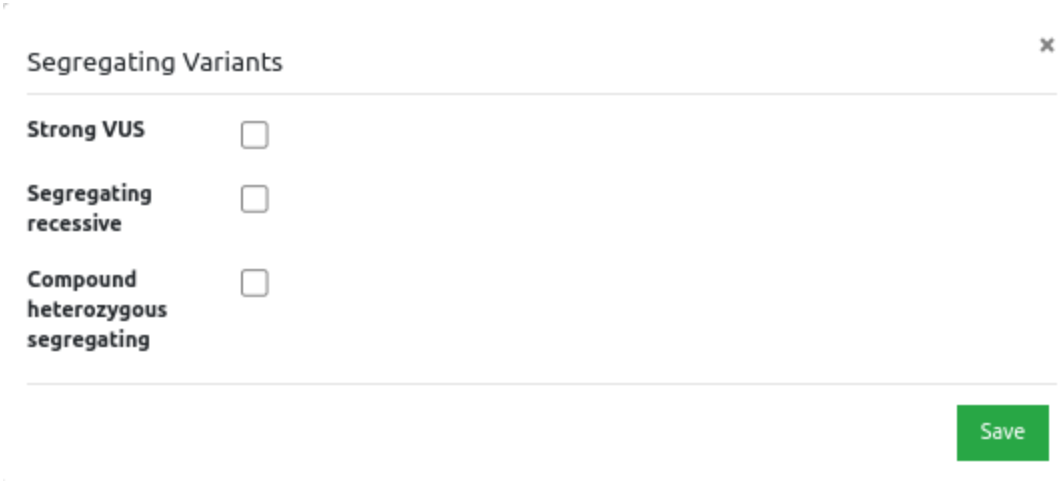
For some of these filters, in order to add the parameterization functionality we have integrated similarly functioning filters into one. For example, you can now find under the filter "Segregating Variants", the filters previously known as:

- Segregating Variants (dominant, all VUS),
- Segregating Variants (dominant, strong VUS)
- Segregating Variants (recessive, all VUS)
- Segregating Variants (recessive, strong VUS)
- Compound Heterozygous Segregating Variants (all VUS)
- Compound Heterozygous Segregating Variants (strong VUS).

The segregating variants analysis falls in the category of algorithmic filters that can be performed on Family trio analyses as shown below:



It is a parameterizable algorithmic filter with the following options:



Segregating dominant variants

Variants annotated as pathogenic, likely pathogenic, or of unknown function (but only if one of the strong pathogenic Germline Variant Classification rules has fired for this variant) that are present (in any zygosity) in all affected samples and absent from all healthy samples. Select the "Strong VUS'' option.

When selected, "Strong VUS'' will keep variants of unknown significance (VUS) only if one of the strong pathogenic ACMG rules has fired for this variant. If not selected, any VUS that meets the filter's other criteria will be kept (irrespective of ACMG rules).

Strong rules are PVS1, PS1, PS2, PS3, PS4 and PP5. In addition, any rule whose strength has been raised to "strong", "very strong" or "stand alone" will be considered a "strong" rule, even if its original strength as per the ACMG guidelines was lower. For example, PM1 can be raised from "supporting" to "strong" if the variant is located in a dense hot-spot.

Segregating recessive variants

Variants annotated as pathogenic, likely pathogenic, or of unknown function (but only if one of the strong pathogenic Germline Variant Classification rules has fired for this variant) that are homozygous in all affected samples but heterozygous or absent from all healthy samples. Select the "Segregating recessive" option.

Compound heterozygous segregating

Variants present (in any zygosity) in all affected samples and absent from all healthy samples and it will also find possible segregating compound heterozygous variants: those variants annotated as pathogenic, likely pathogenic, or of unknown function that are found in genes where all affected samples have at least two so annotated heterozygous variants. Select the "Compound heterozygous segregating" option.

This option cannot be used in combination with the Segregating recessive option.

Running the analysis with multiple affected samples

⚠️ Please note that this filter is not specifically designed for family trios, but for general multi-sample (e.g. cohort) analyses. When running it with multiple affected samples, it will look

for variants that are homozygous in all affected samples and absent or heterozygous in all healthy samples. The segregating variants (dominant) filter will look for variants that are present (het or hom) in all affected samples and absent in all healthy samples.

4.2.15 Trio Recessive

Pathogenic variants that may be causative of recessive disorders in the child of unaffected parents.

We follow this selection process:

1. We exclude variants that are homozygous for the alternative allele in either parent.
2. We exclude variants in mitochondria and chromosome Y.
3. We keep variants where EITHER of these criteria apply:
    a. Variant is homozygous for the alternative allele in the child.
    b. Child is heterozygous and the following two criteria BOTH apply:
        i. The variants did not all come from the same parent; some variants on the gene may have come from the mother and some from the father or are de novo.
        ii. There are two or more variants in the same gene. To qualify, a variant must be in a coding transcript of a gene with a Transcript Support Level consistent with the sample's settings. The variant must be Pathogenic, Likely Pathogenic or of Uncertain Significance (VUS).

Options:

- **Strong VUS**: If selected, keep variants of Unknown Significance (but only if one of the strong pathogenic Germline rules has fired for this variant). VUS variants qualify only if they trigger one of the Germline rules: PVS*, PS* or PP5. If not selected, keep all VUS variants (irrespective of Germline rules).
- **Missing from one parent**: Only keep variants that are missing from at least one parent (strict); in other words, neither parent has all variants that the child has in the same gene.
- **Maximum frequency for recessive**: We exclude any variants that are found with a frequency greater than the selected threshold. The default value is set to 1, so that no variants will be excluded based on the frequency.

⚠️ Please note that we recommend that you further filter by genes with recessive mode of inheritance.

4.2.16 Variants in Common

Variants annotated as pathogenic, likely pathogenic or of unknown function (but only if one of the strong pathogenic ACMG rules has fired for this variant) that are present (in any zygosity) in all samples.

Options:

- Common only to all affected: If selected, keep all variants that are found (in any zygosity) in all affected samples; their status in non-affected samples is ignored.

- Pathogenic only: If this option is selected, variants classified as Benign or Likely Benign will be excluded from the results.

4.2.17 VarSome Picks

VarSome Picks is an advanced algorithmic filter, designed to empower bioinformatic analysis of genomic variants using AI. This tool goes beyond conventional variant prioritization methods by considering essential parameters such as phenotype, gene, and variant data, all within a disease-specific context. Its primary objective is to rank potentially causative variants, aiding researchers and clinicians in identifying significant genetic associations related to specific diseases.

Main Features

Comprehensive Data Sources

VarSome Picks leverages a diverse set of authoritative data sources to enhance the accuracy of its variant ranking. These sources include:

- HPO (Human Phenotype Ontology) - is used to get the phenotype (HPO term) - gene association
- GenCC (Genetic Clinical Characterization)
- CGD (Clinical Genomic Database)
- ClinGen Disease Validity
- MONDO (Monarch Disease Ontology) - is used here to get the disease (MONDO term) - gene associations.
- OMIM® (Online Mendelian Inheritance in Man)
- Gene2phenotype
- PanelApp

By integrating information from multiple databases, VarSome Picks ensures a comprehensive and well-informed list of variants that have been classified as potentially causative for the disease, even those previously labeled as Variants of Uncertain Significance (VUS).

Genetic and Phenotypic Parameters

VarSome Picks incorporates crucial genetic and phenotypic parameters to prioritize variants accurately. These include:

- Phenotype(s) Selected by the User: VarSome Picks ranks variants found in the Top 10 genes associated with the selected phenotype(s) by the user. This feature facilitates tailored analysis, focusing on variants with higher disease relevance.
- Zygosity: Variant-related information on zygosity is considered to assess the impact of heterozygous or homozygous variants on the disease phenotype.
- VarSome Germline Classification: The algorithm applies the ACMG (American College of Medical Genetics and Genomics) criteria for germline variant classification, ensuring robust variant categorization.

- Mode of Inheritance: Gene-related information on the mode of inheritance is taken into account to identify variants associated with specific inheritance patterns.

Quality Parameters

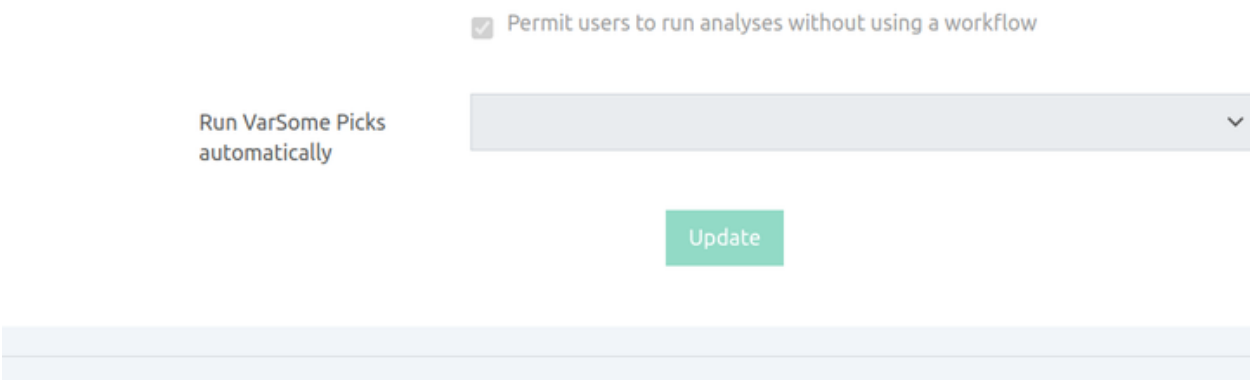
VarSome Picks considers essential quality parameters for variant evaluation:

- Allelic Balance: For germline variants, the algorithm applies a threshold for allelic balance to ensure reliable variant calling.
- Coverage in the Sample: Coverage information is evaluated to gauge the depth of sequencing and assess data quality.

How to use VarSome Picks

VarSome Picks supports germline analyses from FASTQ files and VCF files against hg38 and hg19 reference genomes. It is also available for gene list analyses.

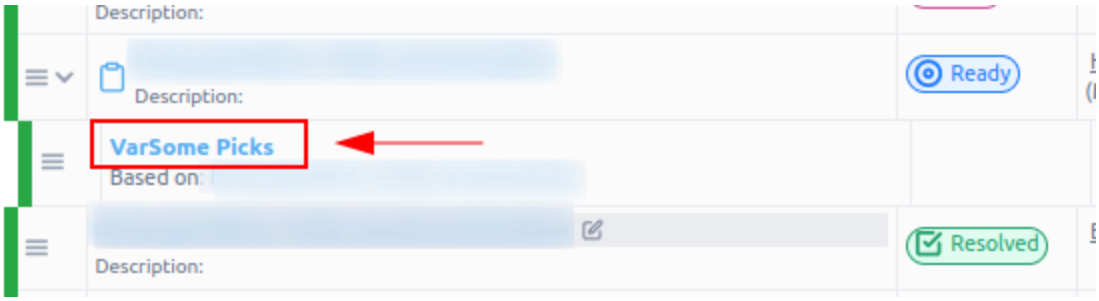
The filter can be run on-demand (Analysis actions > New algorithm filter analysis) once the main analysis has finished. It can also run automatically for all analyses if the group supervisor enables this option in the analysis preferences.



In this case, the algorithmic filter will automatically run only when phenotypes are provided by the user. The results will be shown as the result of any other algorithmic filter, and in this case, with the variants being ordered by the VarSome Pick's assigned priority.

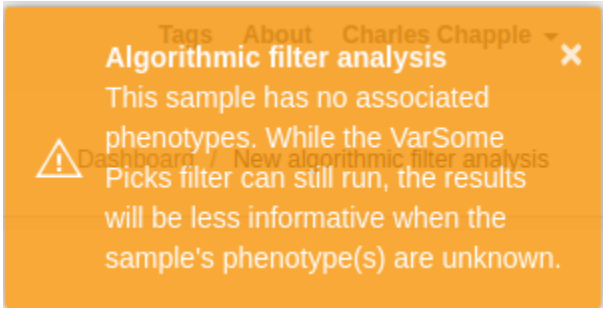
VarSome Picks runs only for:

- Single sample germline analysis
- Gene lists analysis



Whenever phenotypes are added or modified by the user in an existing analysis, if the supervisor has enabled the automatic analyses option, VarSome Picks will re-run automatically. The previous output is erased, ensuring up-to-date and accurate results for the latest phenotype selections.

It is also possible to run VarSome Picks even if phenotypes are not initially provided. In this scenario, the algorithmic filter will not be triggered automatically. However, a warning will be displayed to make sure our users are aware of the potential limitations of the analysis due to the absence of phenotype information:



Note to the User/Disclaimer

Experts and users should exercise discretion in evaluating the appropriateness of VarSome Picks for their specific use case, considering factors like non-specific or genetically heterogeneous phenotypes and idiosyncrasies in search functionality (e.g. “hearing loss” instead of a query using the term “deafness” may not return the relevant gene). Variable expressivity and age-related penetrance, while important considerations, are not currently accounted for by the algorithm. Therefore, the decision regarding the appropriateness of this approach has to be at the discretion of the experts/the users.

4.3 CNV/SV analysis

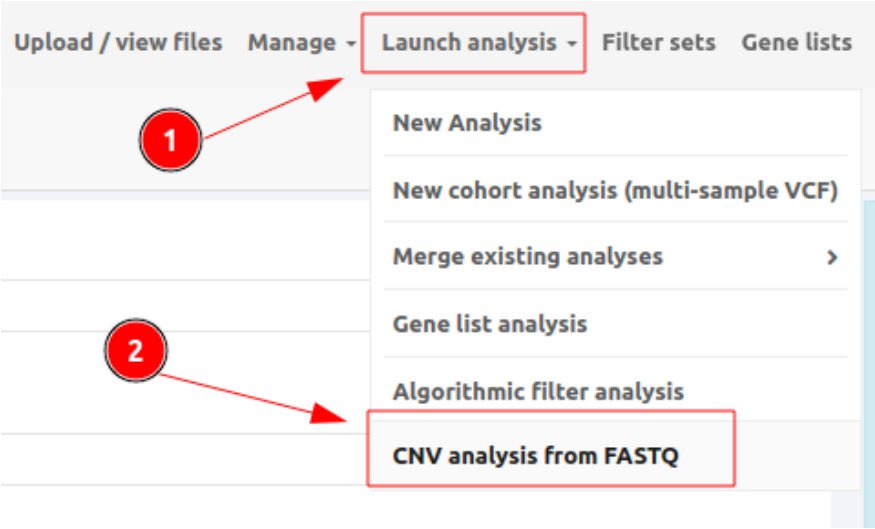
VarSome Clinical currently offers two CNV calling solutions for your samples that are suitable for all types of NGS samples: WGS, WES and panels.

4.3.1 CNV/SV calling (from FASTQ)

VarSome Clinical currently offers two types of CNV calling solutions:

- **Delly** - suitable for single WGS samples and WGS tumor-normal CNV analysis.
- **ExomeDepth** - suitable for cohorts of WES/panels and also for WGS samples.

To start a CNV analysis from FASTQ, please go to “Launch analysis” > “CNV analysis from FASTQ”.



Whole exome sequencing (WES) or targeted panel data

For such samples, we use the ExomeDepth CNV caller. The read depth based tool requires five or more (ideally between five and ten) germline or somatic samples that have already been analyzed on VarSome Clinical. These will be run as a cohort with each sample analyzed using the rest as a control. The samples should all have been sequenced using the same assay since CNV calls will only be made in the assay's target regions. For optimal results, the selected samples should:

- be from the same sequencing run
- come from individuals unrelated to each other and
- be of the same sex (either all male or all female). If the samples of the cohort are of not of the same sex, the CNV calls obtained for chromosomes X and Y will not be reliable

All samples will be analyzed together and the results (along with a visual display) of each sample will be shown as a sub-analysis of that sample.

⚠️ Please note that an inherent limitation of WES is that it produces reads only covering the ~2% of the human genome that falls in exons. Therefore, the full spectrum of CNVs and breakpoints may not be completely characterized. In addition, many large CNVs and

cross-chromosome events may not be detected. For optimal results, we suggest either sequencing the entire genome (WGS), or a different experimental approach such as array CGH. Nevertheless, CNV detection based on WES data may give a quick insight into CNV patterns for a specific disease or phenotype. For more details on the limitations of calling CNVs in such data, please see R. Tanner *et al.*, 2014.

Sensitive mode

CNV calling for **non-WGS** CNV analyses is also available in "Sensitive mode". Compared to standard mode, a lower CNV detection threshold is applied, resulting in more sensitive calling and typically in a higher number of calls. CNV detection can be particularly challenging; for instance single exon CNVs can be hard to call. Still, in a clinical setting, the ability to detect such CNVs is of paramount importance. Sensitive mode is optimized for the needs of clinical laboratories. It allows a shift to the trade-off between recall and false discovery to benefit sensitivity, compared to the standard mode.
You can enable this feature for either somatic or germline samples (or both) in Preferences. ⚠️Please, note that these settings are only available to the group administrator and any changes will be applied to all users of the group.

Whole genome sequencing (WGS) data

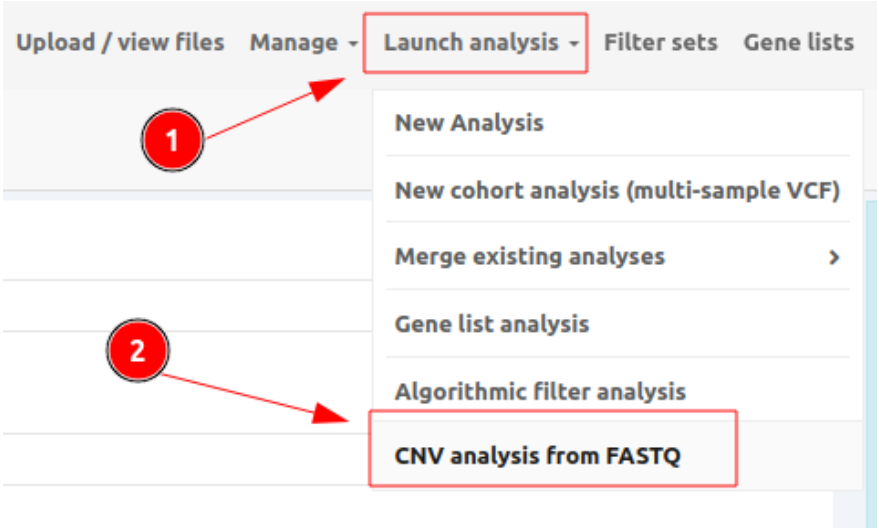
For WGS samples, VarSome Clinical offers two solutions:

- CNV calling for a single WGS sample. We use delly, an integrated structural variant (SV) caller tool that can detect both CNVs and other forms of Structural Variants (SVs) at single-nucleotide resolution in short-read genomic sequencing data. It combines 3 different approaches (paired-ends, split-reads and read-depth) to discover extensive genomic rearrangements. Quality passed CNV calls (deletions and duplications) are retained, while other types of SVs are currently not reported.
- CNV calling for a cohort of (2-5) WGS samples. The ExomeDepth caller has been adapted to also process WGS. The solution is suitable for samples with long CNVs (>50kb) that may not be reliably called by delly. For WGS, the assay target regions comprise the complete genome, split into 50Kb bins. As a result, this imposes a hard minimum size limit: no CNVs smaller than 50Kb can be detected using this approach. The requirements for non-relatedness between the samples and their processing by the same laboratory,

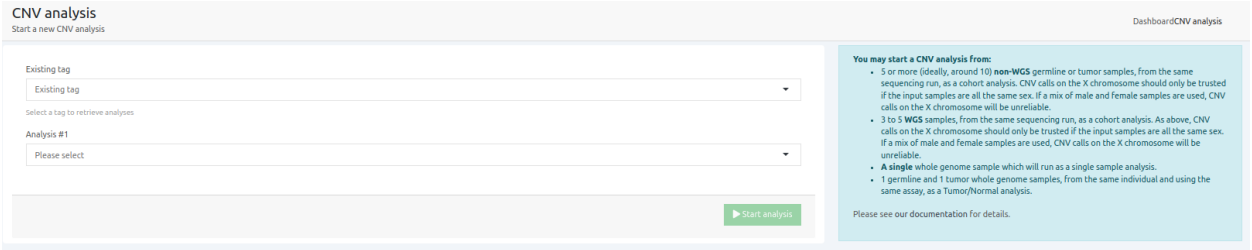
sequencer and ideally in the same batch, apply to WGS samples too. All samples are analyzed in a single CNV analysis and the results (along with a visual display) of each sample are shown as a sub-analysis.

A step-by-step example on how to run a CNV/SV analysis

Select “CNV analysis from FASTQ” from the “Launch analysis” drop-down menu on VarSome Clinical:



VarSome Clinical interface allows you to select a minimum of five (5) and a maximum of twenty five (25) already analyzed samples to be used as a cohort for CNV calling. For best results, we recommend you select **5-10 samples** from unrelated individuals of the same sex that were sequenced on the same sequencing run.



Each sample's results will appear as a sub-analysis of the main analysis.

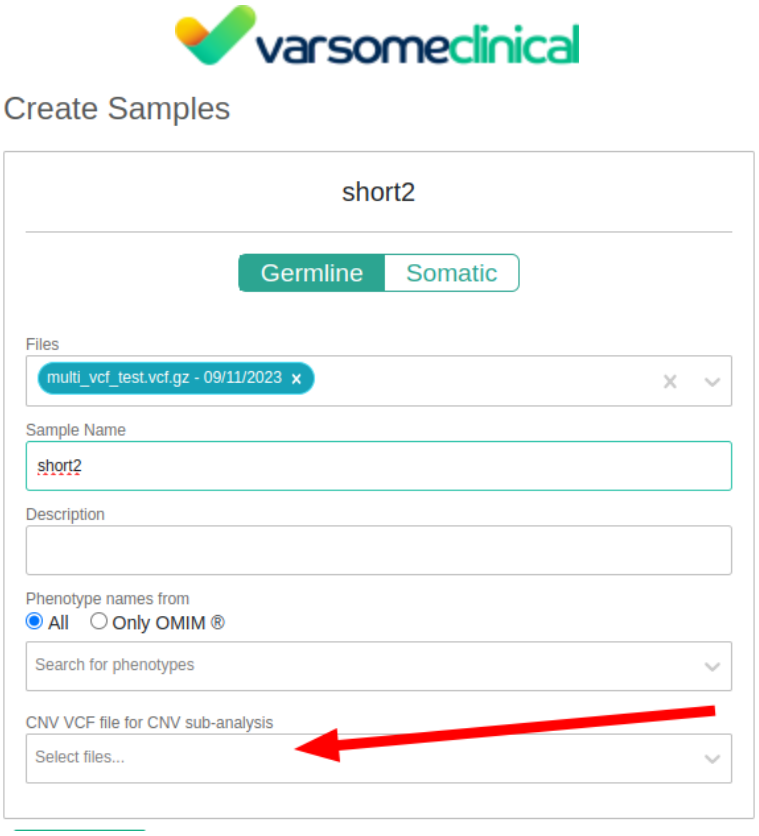
4.3.2 CNV annotation (from VCF)

VarSome Clinical provides a pipeline to annotate CNVs from VCF files. There are two ways to annotate CNVs from VCFs.

- Provide a valid VCF file that contains **both** copy-number-variants (CNVs) and SNPs / small INDELs when launching a new analysis either from FASTQ or VCF. Files that include both small variants and CNVs will be split to two separate files, one for small variants and one for CNVs (i.e **.filtered** and **.cnv**).
- Provide a valid VCF file containing only CNV variants.

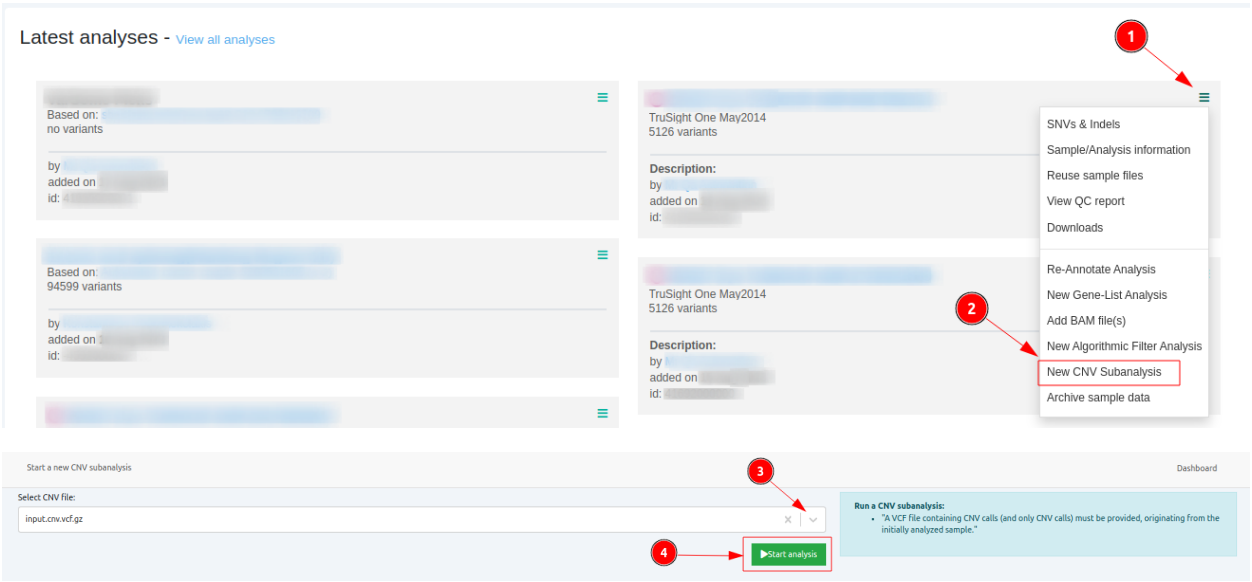
The CNV analysis from VCF is launched as a sub-analysis of the main analysis. You can launch a CNV annotation by:

- Adding a CNV VCF file when defining your sample.

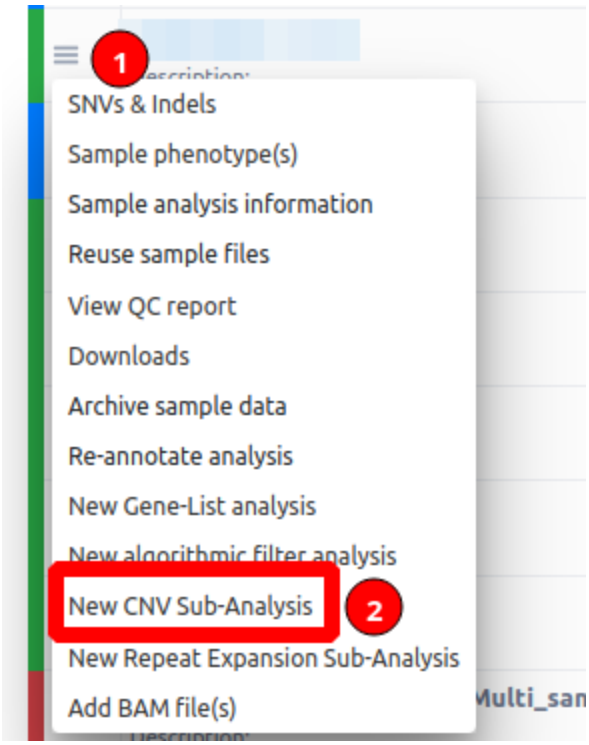


- Launching the analysis once the main analysis has finished as a “New CNV sub-analysis” either from single or multi sample analyses.

From the Dashboard menu:



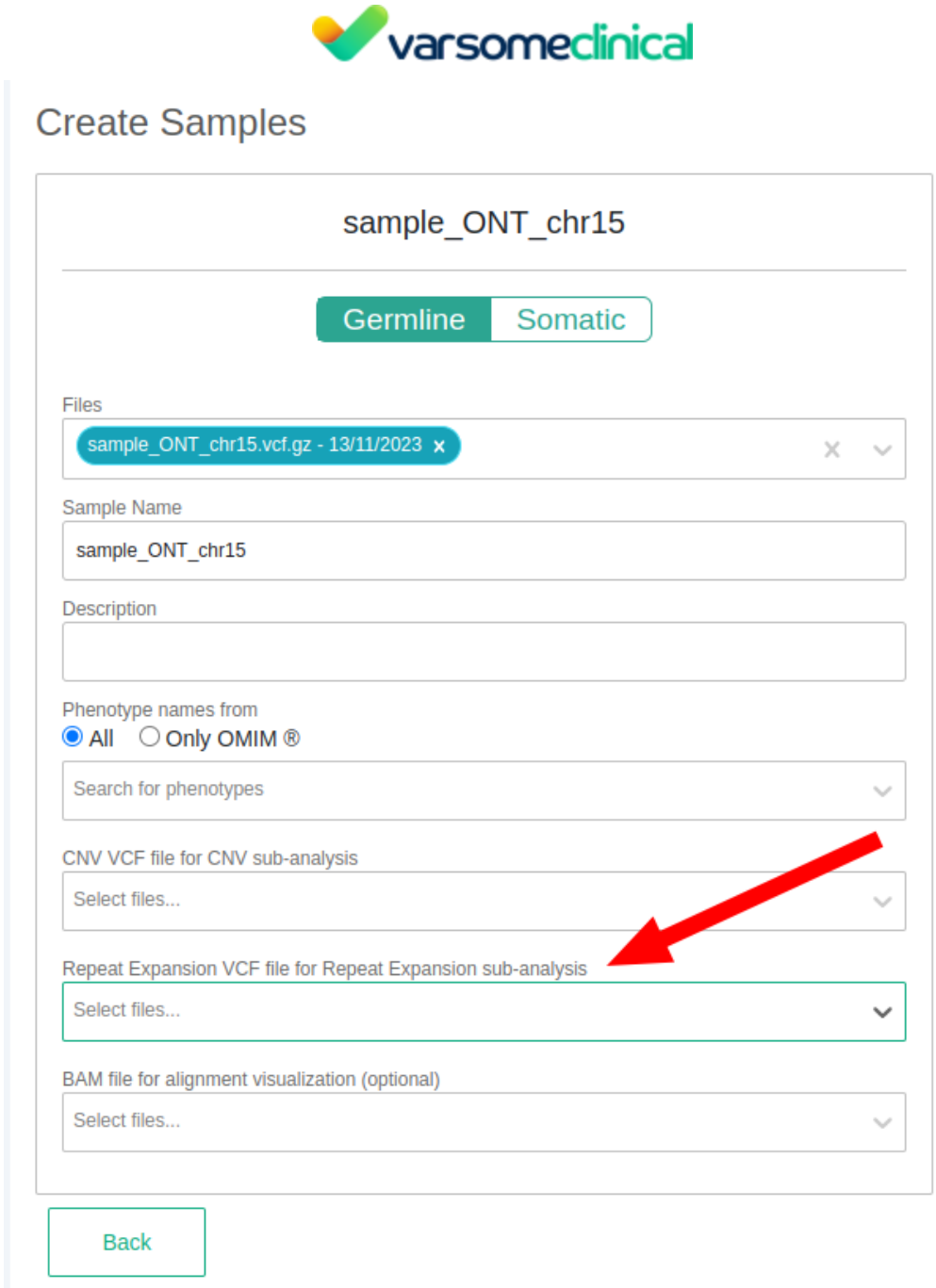
From the Analyses menu:



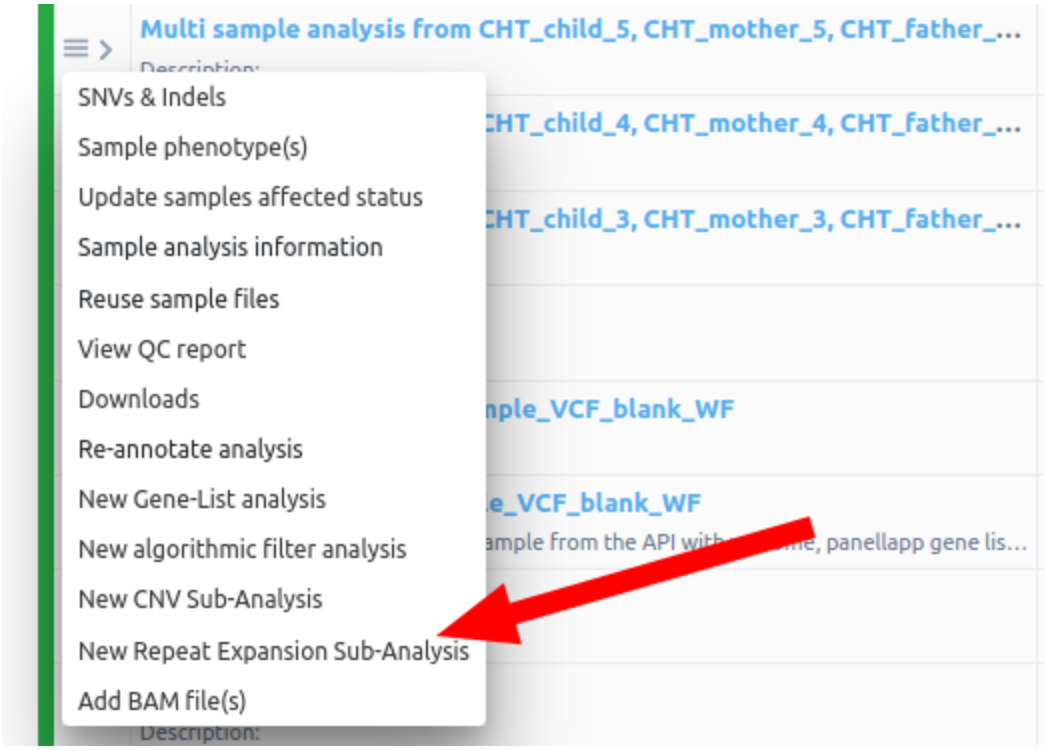
4.4 Repeat expansion annotation from VCF file

VarSome Clinical can annotate short tandem repeats (STR) VCF files. Please go to the "Requirements for Repeat Expansion VCF files" to find out more information.
The STR analysis from VCF is launched as a sub-analysis of the main analysis. You can launch a STR annotation by:

- Adding an STR VCF file when defining your sample.



- Launching the analysis once the main analysis has finished as a “New Repeat Expansion sub-analysis” either from single or multi sample analyses.

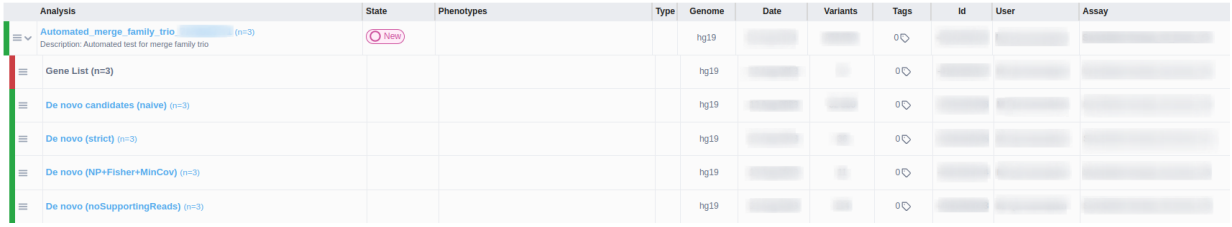


5. Merging analyses

- Family trio analysis
- Carrier analysis for a couple
- Merge analyses

To merge already finished analyses, click on ”Launch analysis” and choose one of the options under “Merge existing analyses”. When the analyses to be merged were run from FASTQ input files, VarSome Clinical will perform **joint calling** on all samples to be merged. The allelic balance and the coverage for each sample will be displayed in the resulting variant table. Clicking on the coverage numbers opens a window showing the read coverage (JBrowse) for the selected variant in that sample.
If any of the parent analyses were instead run from VCF input data, then joint calling cannot be performed and so VarSome Clinical will merge the results of the original analyses directly.

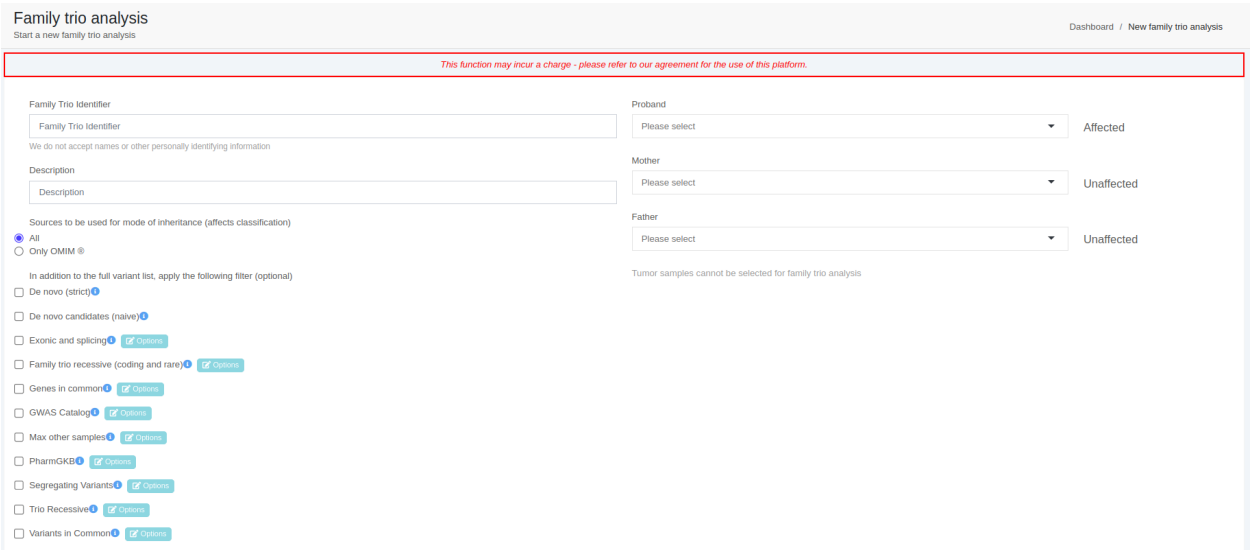
The finished analysis will be displayed as shown in the image below: For analysis containing more than one sample:



The first entry corresponds to the main sample (Automated_merge_family_trio (n=3)) , the merged analysis, and below it are shown any sub-analyses applied to the parent merged analysis (for example *De novo candidates (naive), De novo (strict)*, etc).

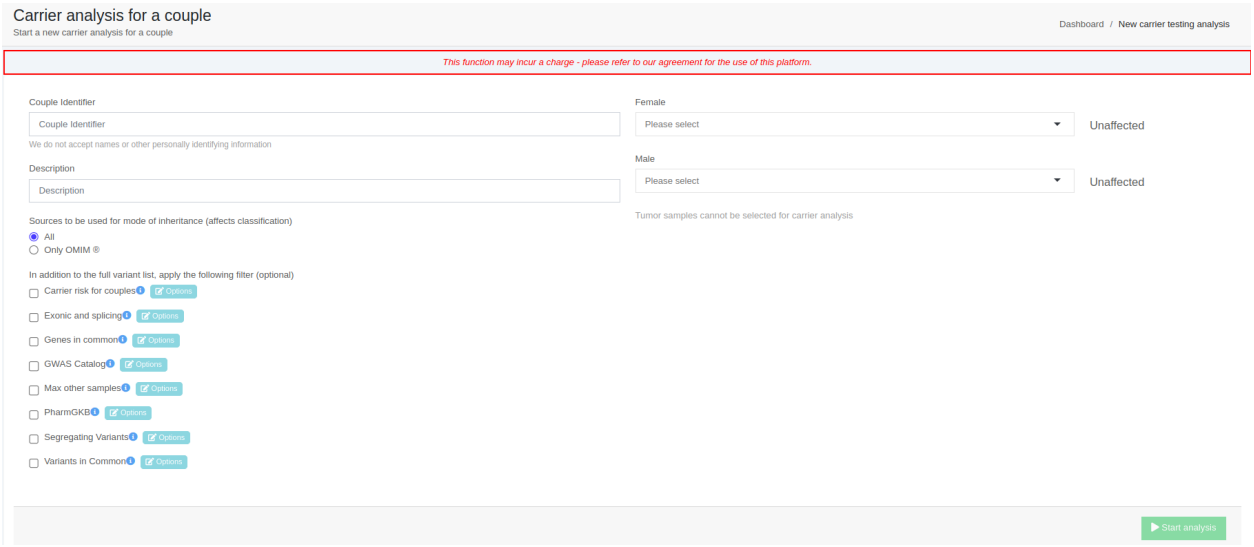
5.1 Family trio analysis

For a trio analysis, the mandatory fields are the analysis identifier, and the identifiers of the Proband, Mother and Father analyses. You can also select any of the listed algorithmic filters which will be applied to the analysis' results. Please click on the icon to get the description of each filter. The algorithmic filters (eg. De novo candidates) are optional, and can be started at any moment. Click on ”Start analysis” to launch the analysis.



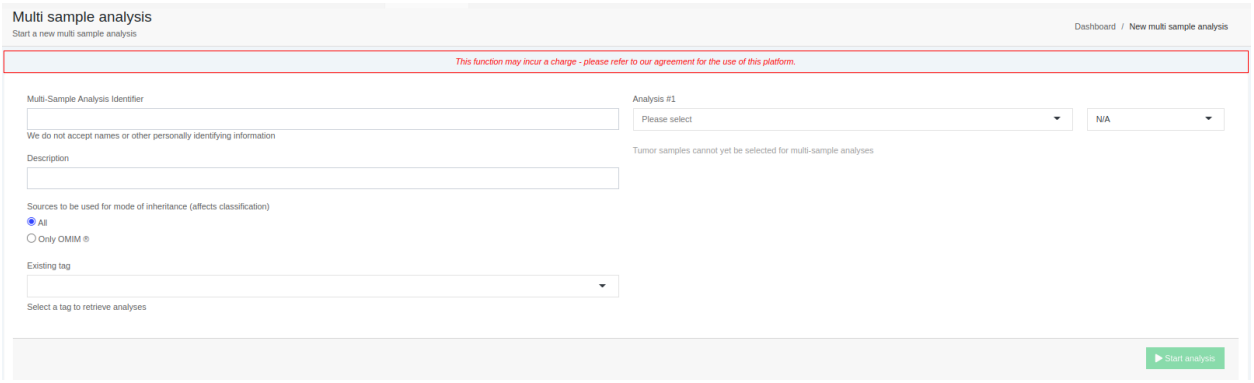
5.2 Carrier analysis for a couple

Choose an identifier for the analysis, and an optional description. Then, on the right hand side, select the female and male samples that will be analyzed. You can also select any of the listed algorithmic filters which will be applied to the analysis' results. Please, click on the icon to get the description of each filter. The algorithmic filters (eg. De novo candidates) are optional, and can be started at any moment. Click on ”Start analysis” to launch the analysis.



5.3 Merge analyses

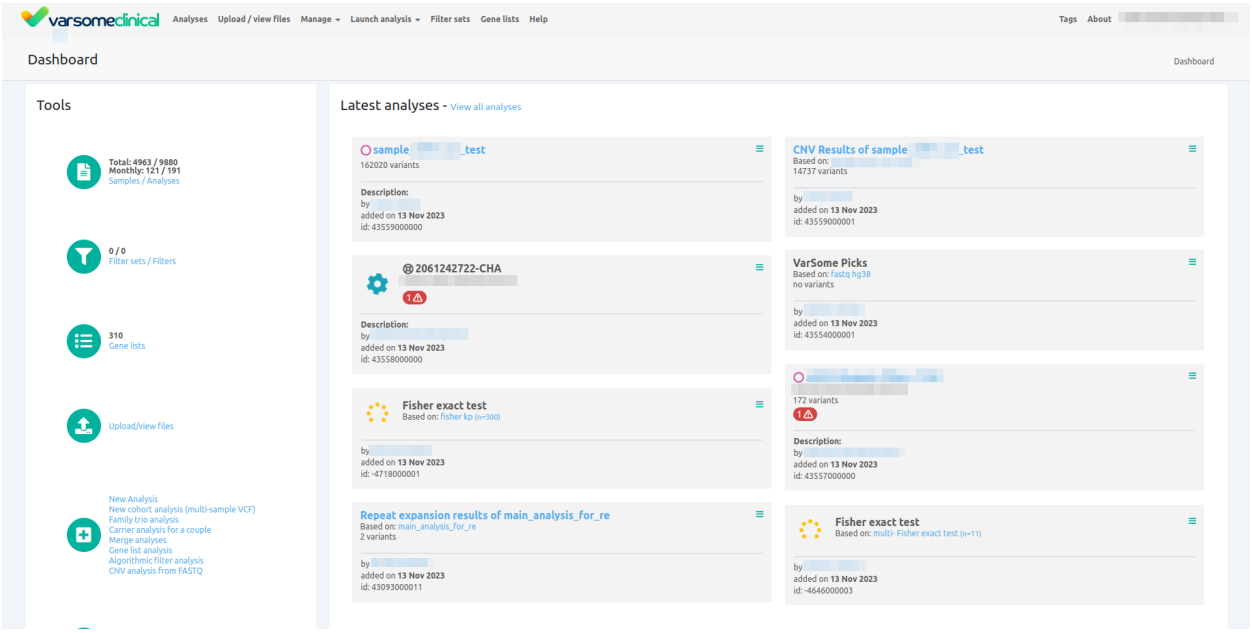
Select this option "Merge analyses" to merge analyses with no special settings. Simply set an identifier and select the analyses to be merged on the left hand side. Click on ”Start analysis” to launch the analysis.



6. VarSome Clinical Dashboard and Analyses pages

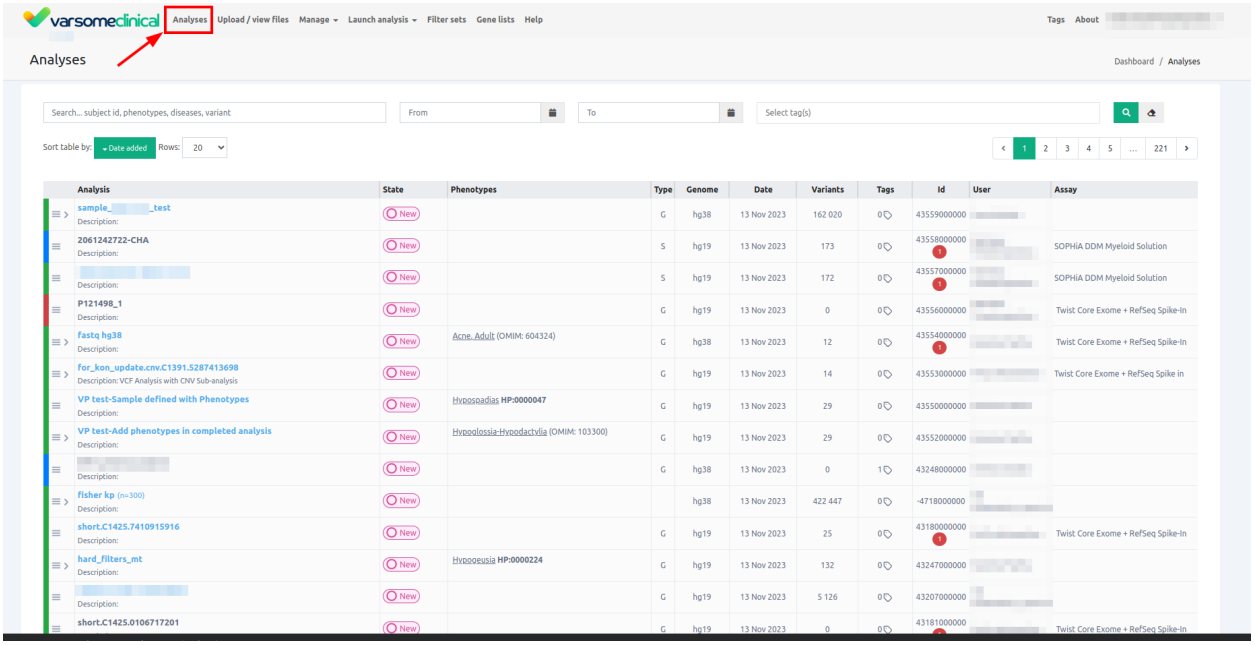
The **Dashboard** page displays your *latest* analyzed samples, while the **Analyses** window displays all the samples that you or your group have uploaded to VarSome Clinical.

In both pages you can see the analysis status (e.g. running). An email will be sent to you once the analysis finishes.



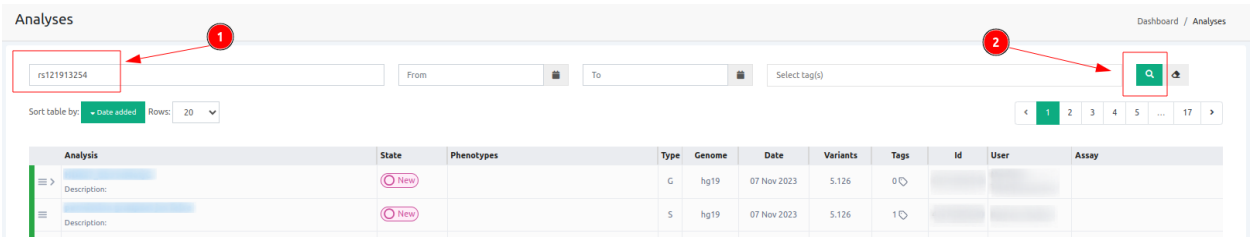
⚠️ Please note that when an analysis has no results to report, no variants will be displayed in the sample's grey box and no active link will be visible to access the results.
If however an analysis doesn't report any information on the variants called (or not), it doesn't necessarily mean that it has no results to show.

When clicking on “Analyses”, you will be redirected to the Analysis Table View, a table that contains all the analyses that you or members of your group have uploaded to VarSome Clinical. From the Analyses Table, you can see the status of each analysis you are running or have run, as well as access the main analysis information and select any actions on an analysis.



For each analysis, you can view some basic information like the name and the description of the analysis, the state, the phenotypes linked to this analysis (if any), the type of the sample (germline [G] or somatic [S]), the reference genome used for this analysis, the date it ran, the number of variants that were called, the number of tags added to this analysis, the analysis ID, the name of the user that launched the analysis as well as the Assay used for this analysis.

VarSome Clinical allows the user to search for any of the analyses in which a variant has been found. This can be done from the “Analyses” page, if you type the variant you are looking for and press enter/click on search (🔍):

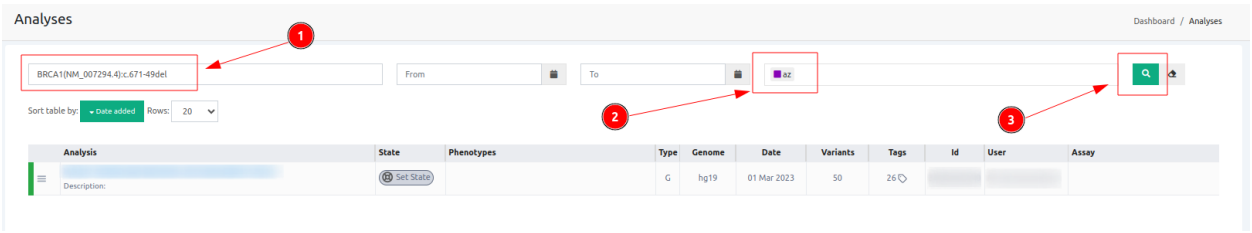


Only those samples containing the variant will be shown in the "Analyses" menu.

In addition, it is also possible to identify samples containing the given variant within a subset of analyzed samples. To do this, you can combine the previous feature with the Sample Tags feature.

First, you will need to label the samples of interest using the Sample Tags. Then, you can go to the “Analyses” menu and search for a specific variant in a subset of samples labeled with a particular Tag.

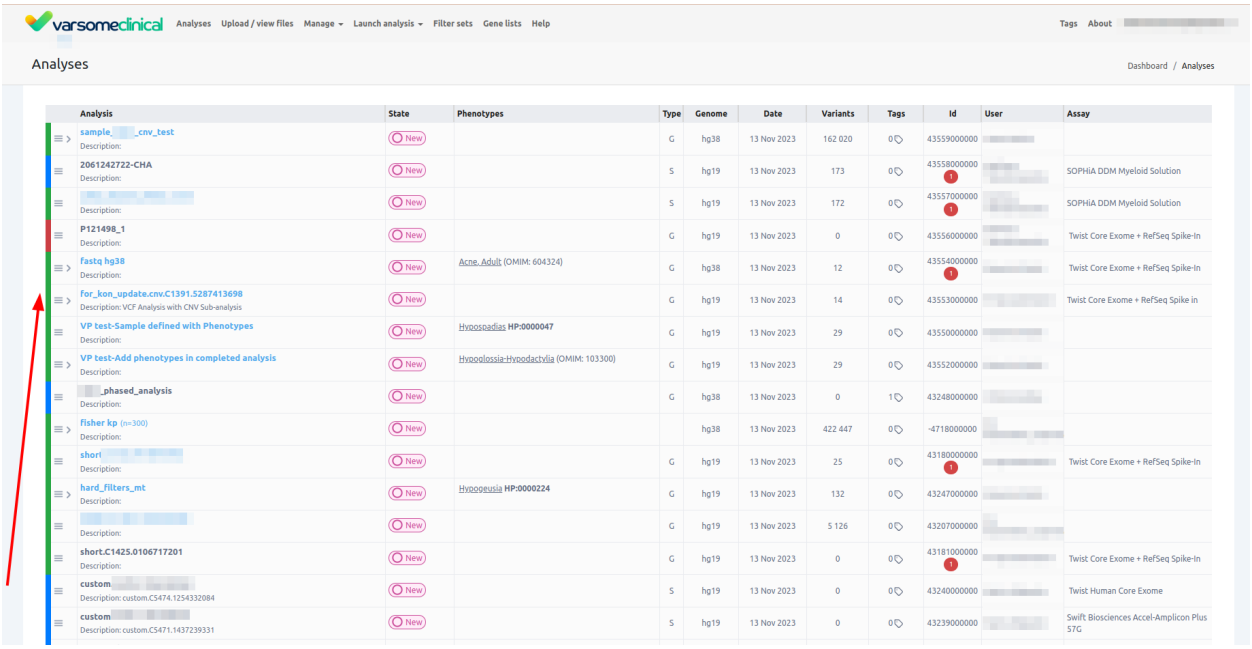
In the example shown in the picture below, only those samples labeled with an "az" tag and containing the specific variant will be shown.



The accepted formats for variant search queries are the following:

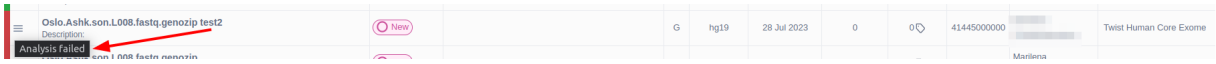
- Chromosome - position - ref seq - variant seq (e.g. chr7:140453136:A:T)
- HGVS DNA-level variants (e.g. BRAF:c.1799T>A, BRAF(NM_004333.6):c.1799T>A)
- HGVS single amino acid substitutions (e.g. BRAF Val600Glu, BRAF p.Val600Glu, BRAF V600E)
- dbSNP rs id (e.g. rs113488022)

You can view the current status of each of your analyses from the colored bar on the left and by hovering the mouse over the bar:

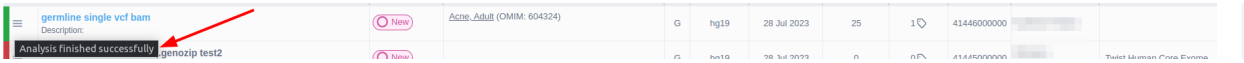


The different colors indicate different sample/analysis status that follow this pattern:

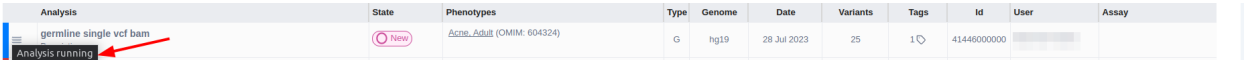
Red: the analysis has failed



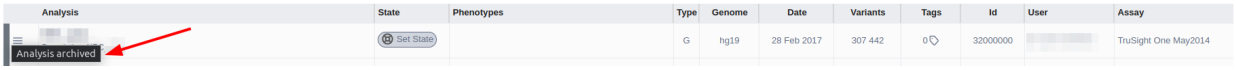
Green: the analysis has finished successfully



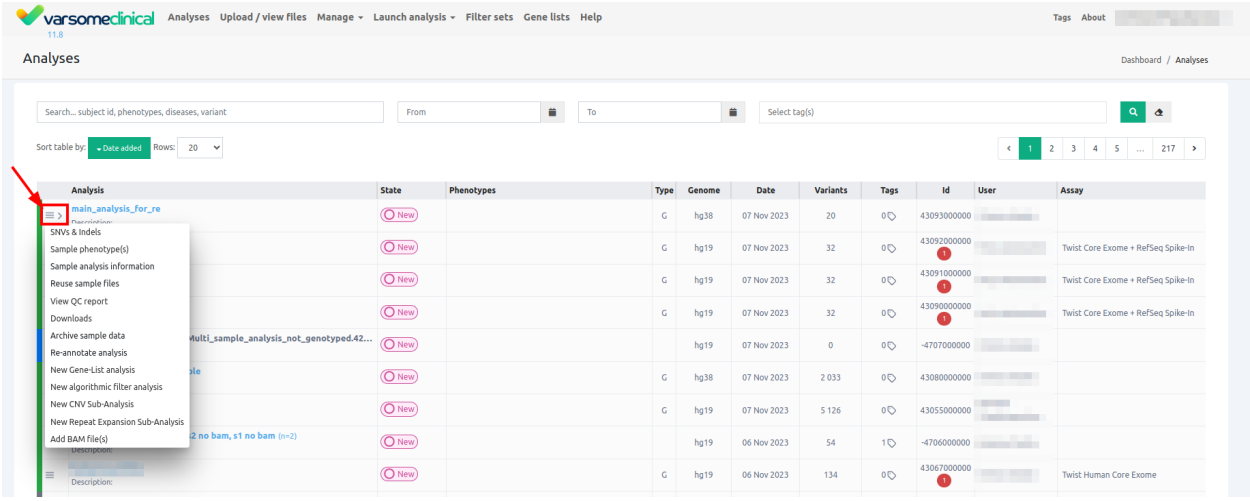
Blue: The analysis is currently running



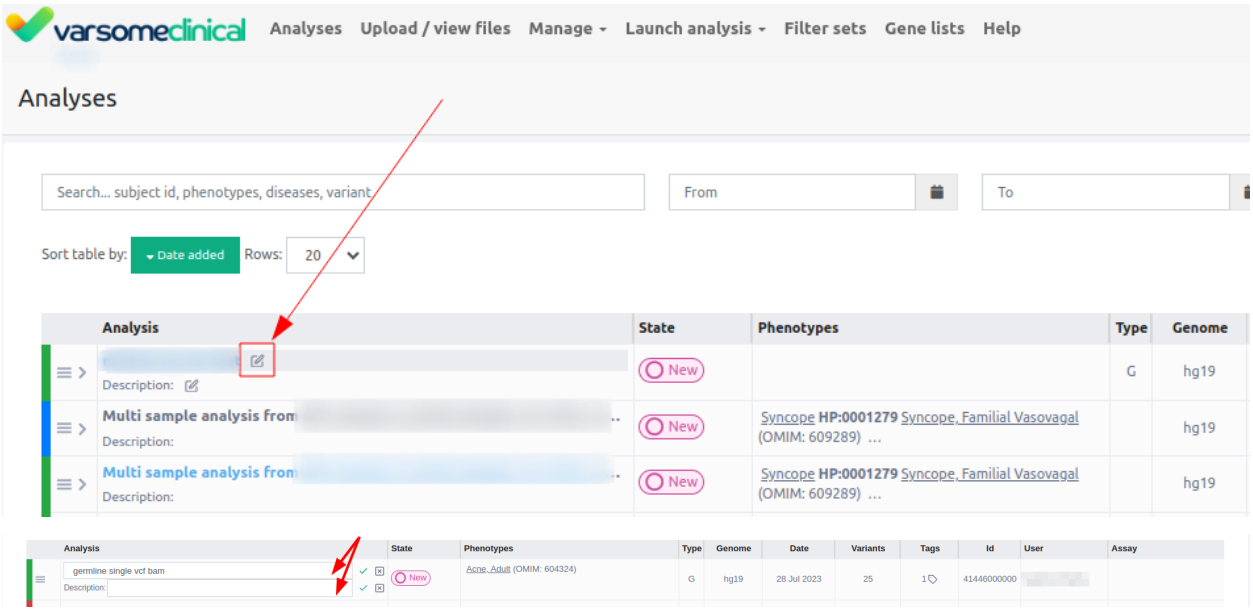
Gray: The analysis is archived



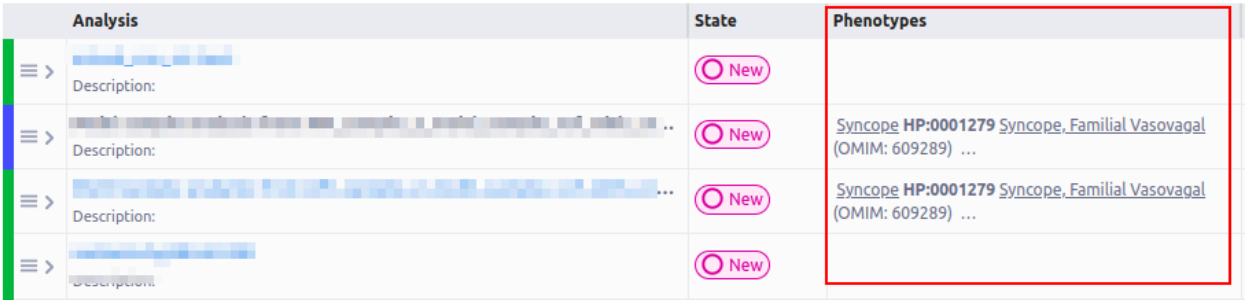
You can open the analysis actions menu by clicking on the three horizontal lines next to the name of the analysis:



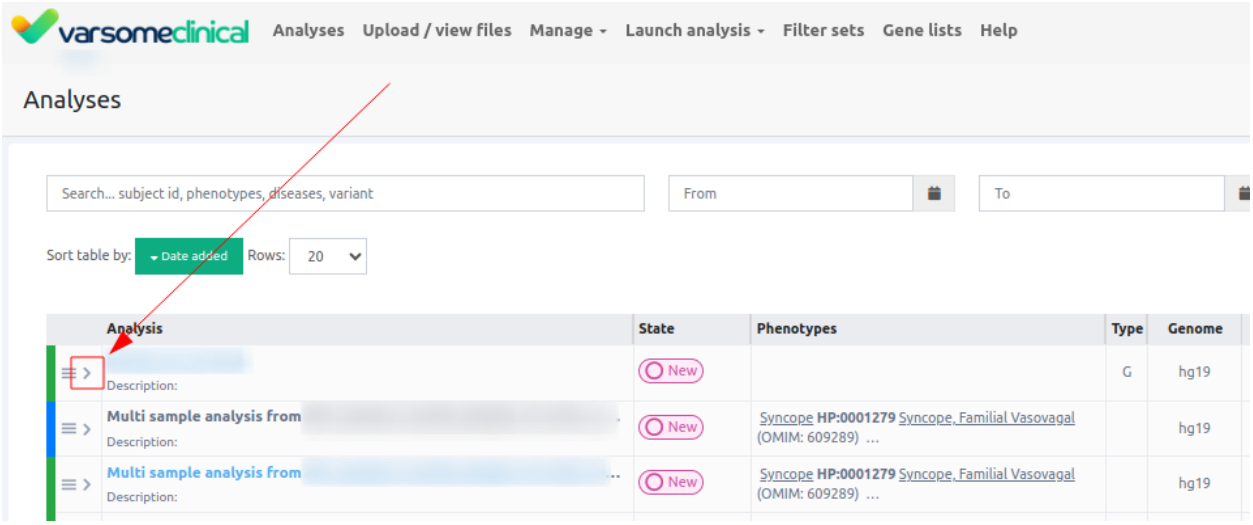
From the Analyses Table, you can also change the name or the description of your analyzed samples:

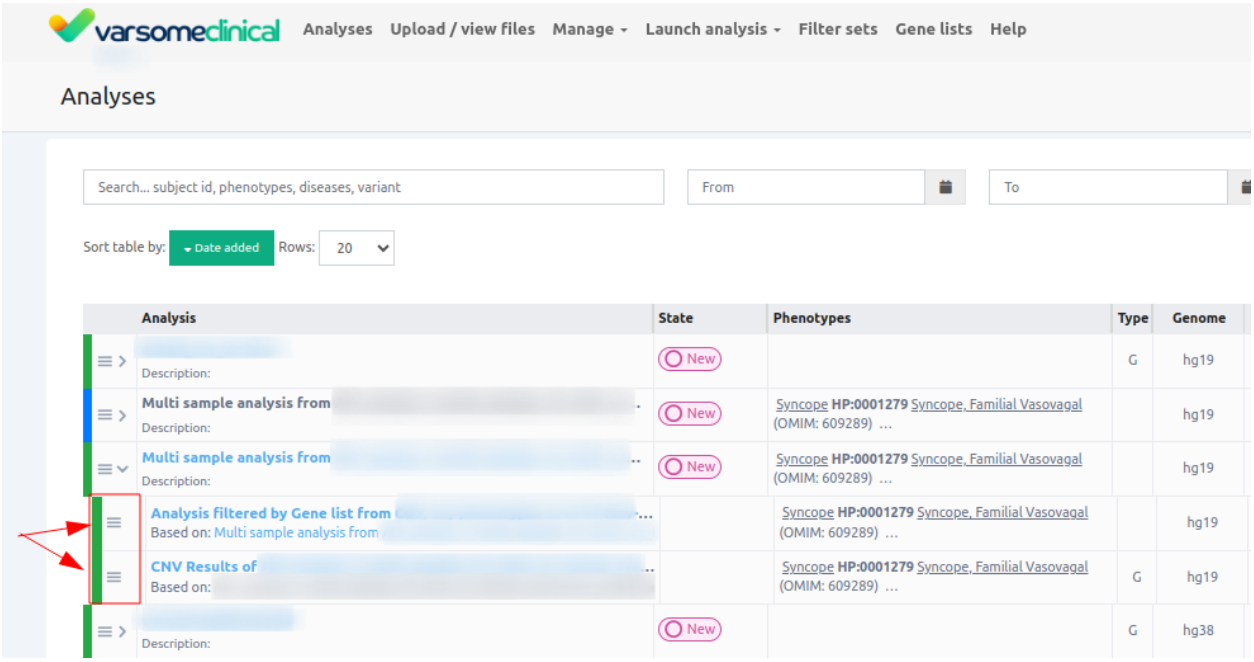


Additionally, when adding phenotypes to the analysis these will be displayed on the Phenotypes column.

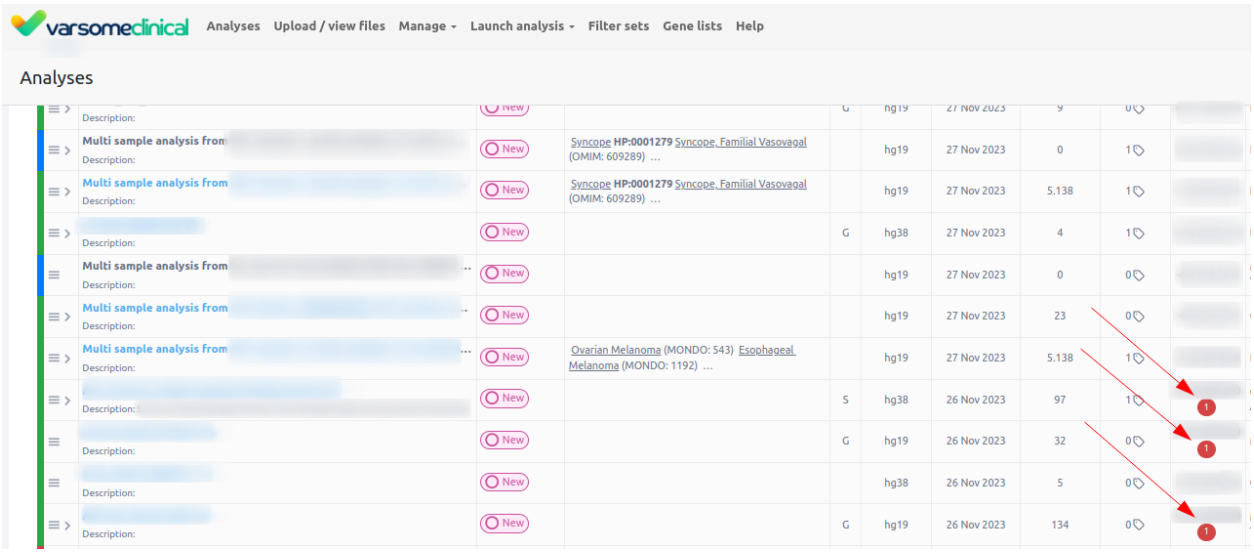


When you launch any sub-analysis on your main analysis, you can click on the arrow on the left of the analysis to view the sub-analysis:

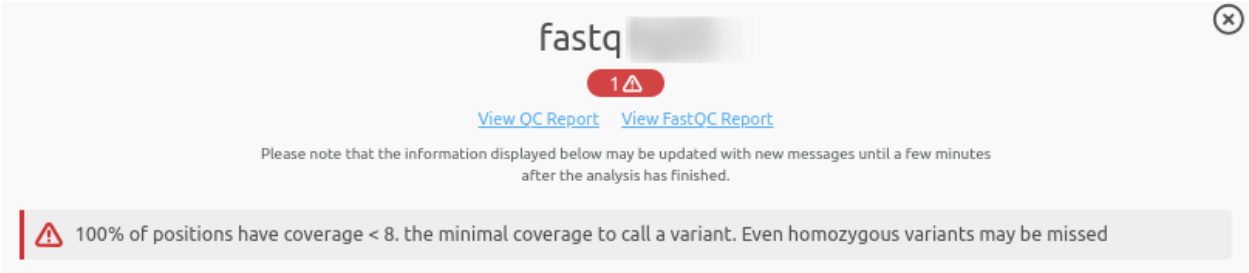




The Analyses Table view also displays possible issues detected in the sample by [FastQC or other quality control]measures applied:

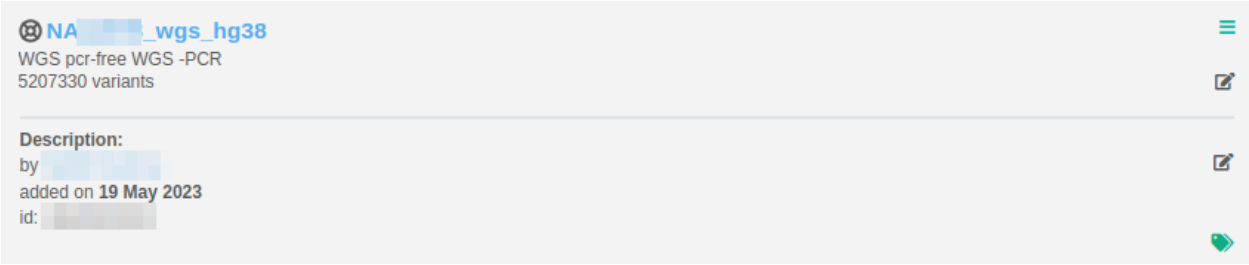


Clicking on the error or warning icon will let you investigate the issue.

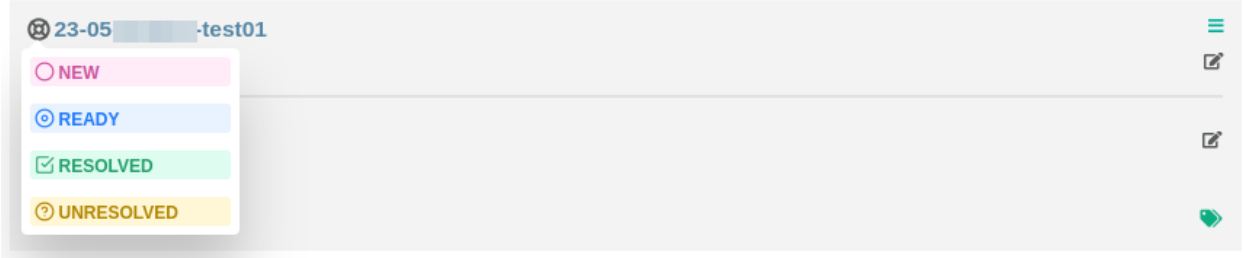


When an analysis is finished, it will appear on your **Dashboard** and, depending on the state of your clinical sample, you can select a predetermined status for it. The **Clinical Sample Status** is a way to organize analyzed samples within a team or between different teams that collaborate to yield a clinical result. In this way, when the analysis is finished and is ready for a curator team to examine the results and yield a diagnosis, the **Clinical Sample Status** can give information on the status of the analysis at a given moment by the previous team that worked on it. If a clinical diagnosis was completed for an analysis, the **Clinical Sample Status** can be changed to indicate whether a diagnosis was made (Resolved) or if no diagnosis could be reached (Unresolved).

For the analyses that were performed before 11.7.5 release a black status icon will appear next to the analysis name and by clicking on it, you can select a status for your older analyses:



After an analysis is finished, you can choose between four different statuses:



NEW : The analysis has finished running and its results can be viewed.

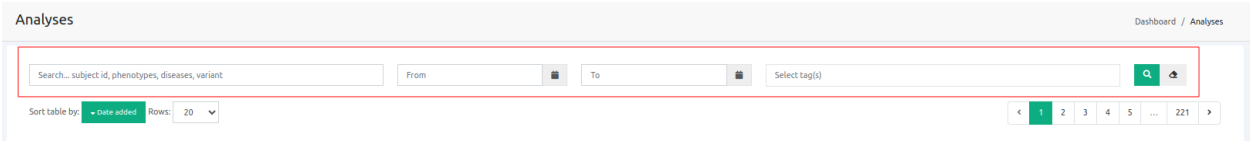
READY : The analysis is ready for review. This option is useful when multiple teams share the work of analyzing a sample, and the first team sets up basic filtering and then passes it on to

the next team. Setting the state to "Ready" indicates that the first pass has been done and the sample is now ready to be viewed by the next team.

RESOLVED : A diagnosis has been made.

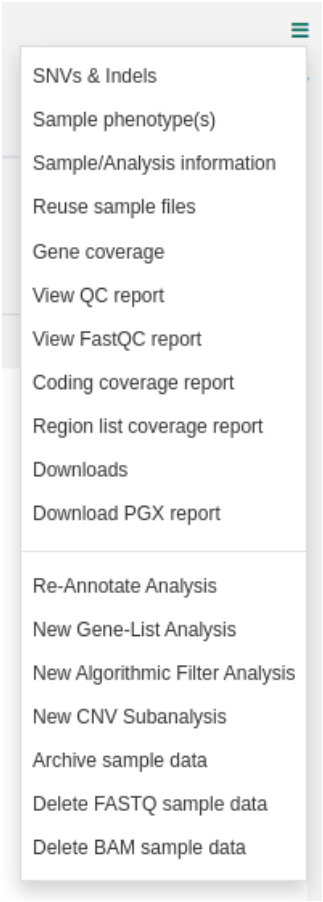
UNRESOLVED : The analysis was inconclusive: no diagnosis can be made at this time.

The Analyses tab displays all the analyzed samples. You can use the “Search” field to look for sample names, users, phenotypes or diseases. Gene list analyses or algorithmic filters are displayed as sub-analyses of the main sample.

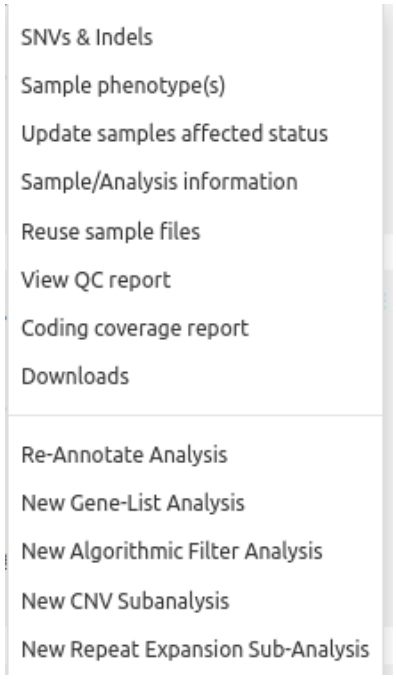


To see the list of available options for the analysis, click on .

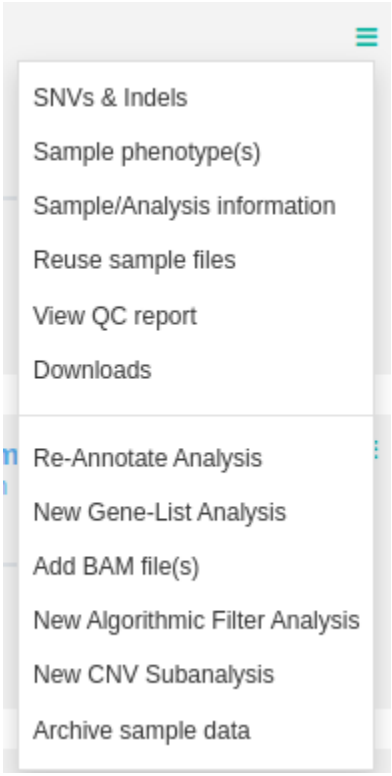
Analysis starting from FastQ:



Multi sample analysis:



Analysis starting from VCF:



6.1 “Analysis actions" options

You can choose among a series of actions that allow you to retrieve information or generate different reports, providing an overview of your results, or further explore the analysis. You can also access this menu from the *"Analysis Actions"* button of the upper right corner at the variant table page.

SNVs & Indels

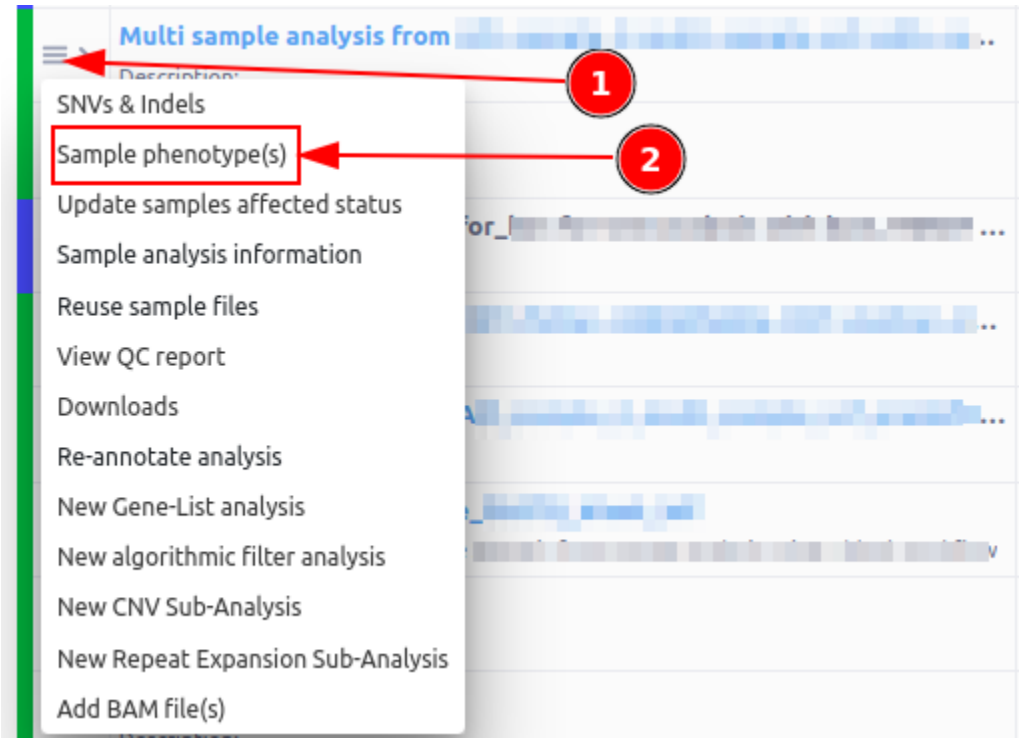
The results of the analysis. When clicked at the Dashboard leads to the Variant table.

Sample phenotype(s)

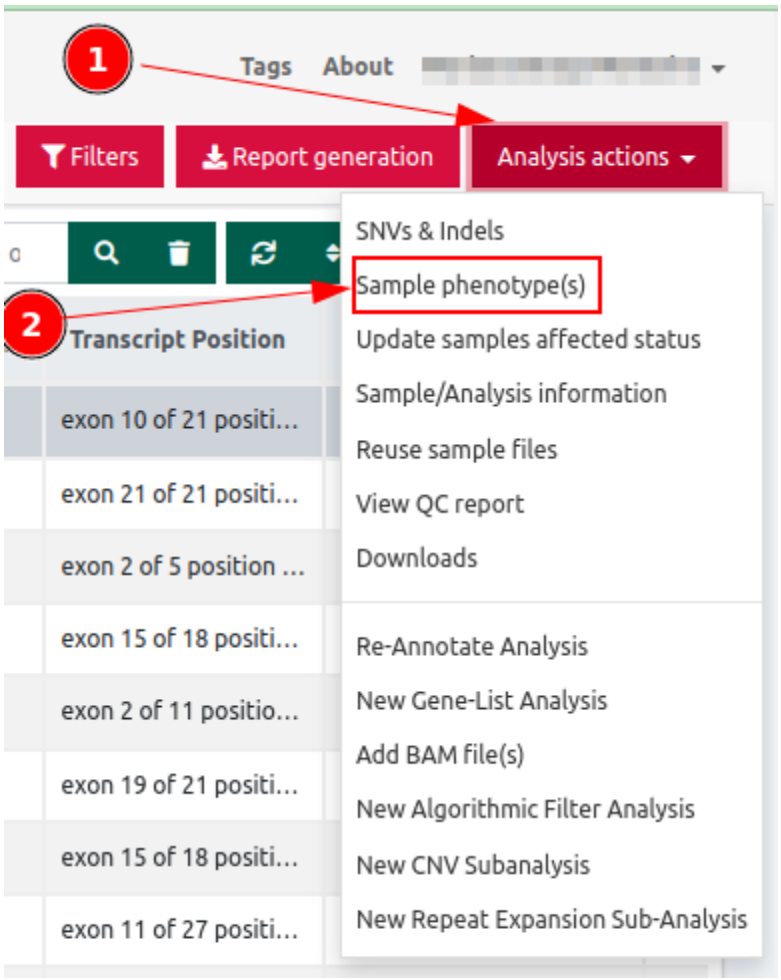
Using these fields it is possible to add and modify the clinical description of the sample as described below.. You can choose this option if you want to provide phenotypes associated with your case from a drop-down menu or view the already selected ones. The phenotypes will be matched to the resulting variants.

Add / Modify phenotypes

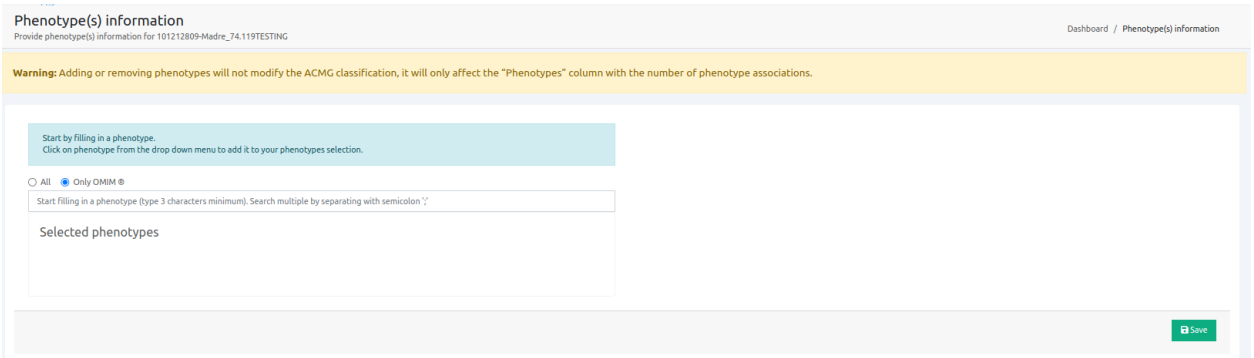
You can also add phenotypes in an already analyzed sample by clicking on the button of the Analysis actions menu either when viewing your sample in the Analyses page:



Or you can edit the Phenotype information from the results page of an analysis by clicking on the "Analysis actions" button:

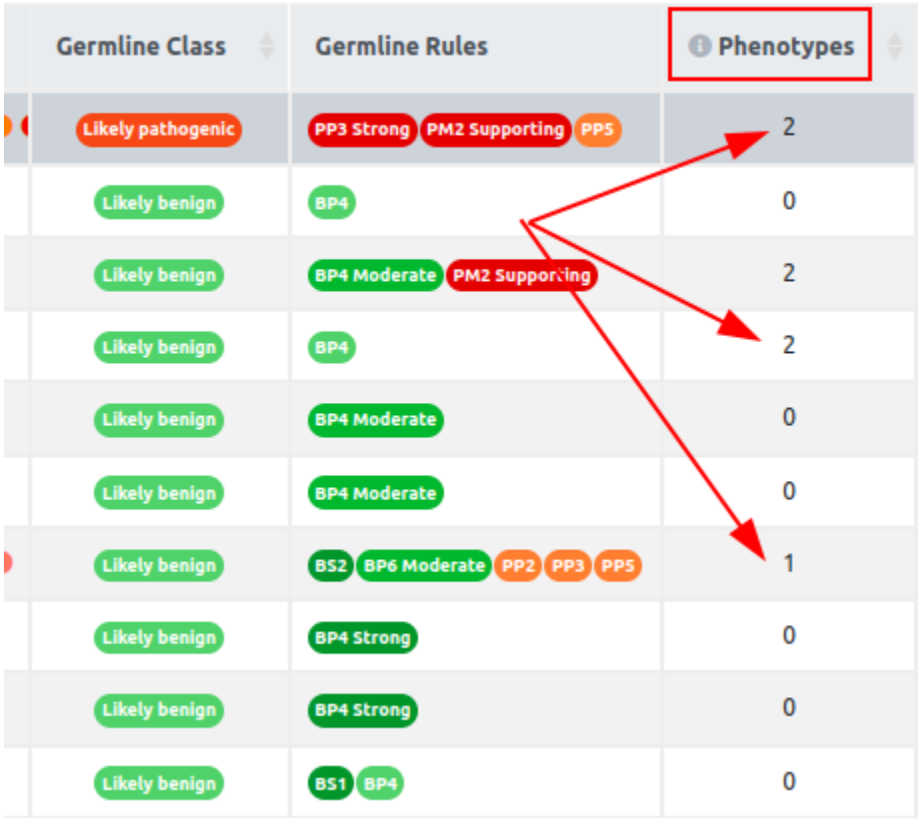


You will then be directed to a new screen where you can add the phenotype information and then click on "Save" to apply the changes. After a few minutes, the "Phenotypes" column will be updated



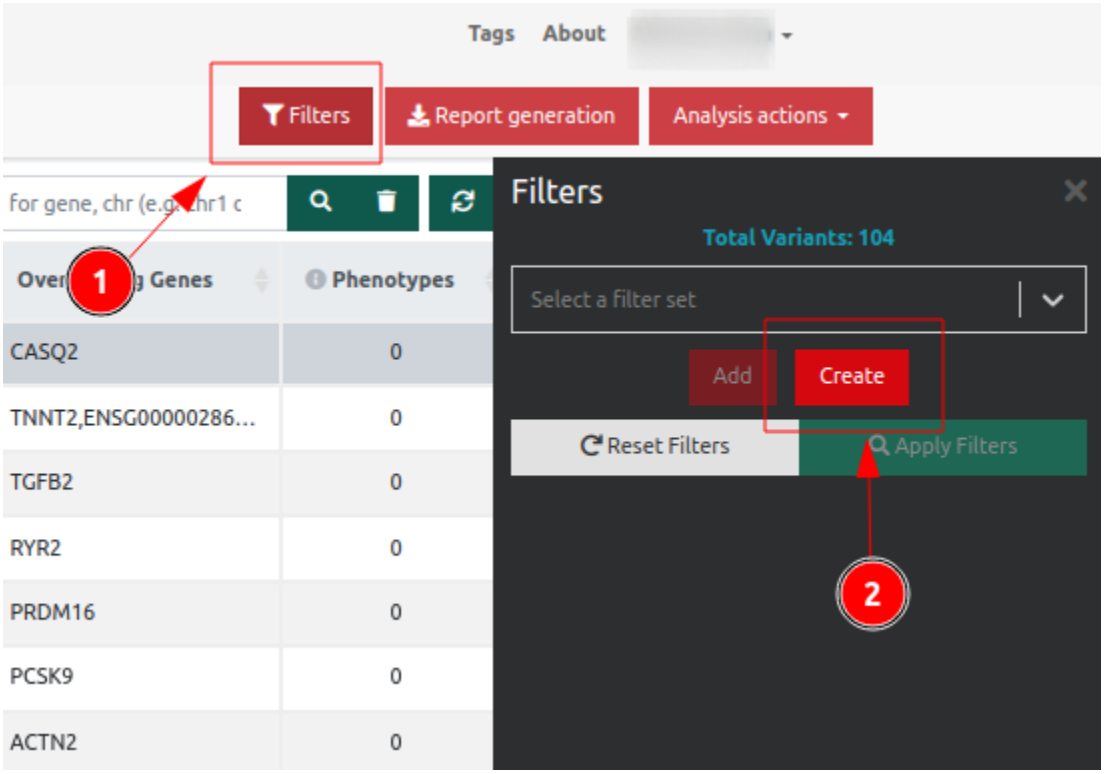
There will be a column named "Phenotypes" in the variant table with a value per each variant. This column will contain the number of user input phenotype(s) associated with the variant

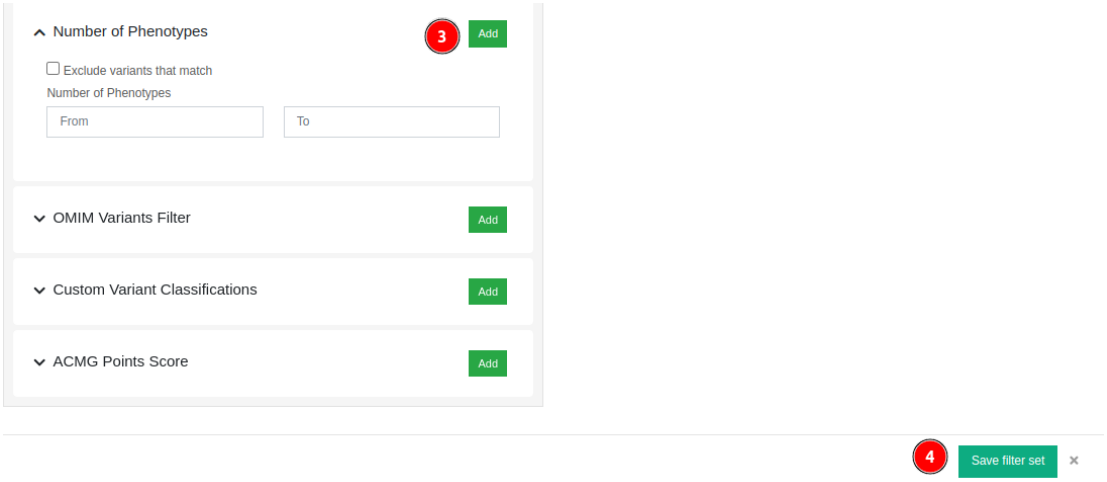
gene.



Hover over the number with the mouse to see the name of the matched phenotypes.

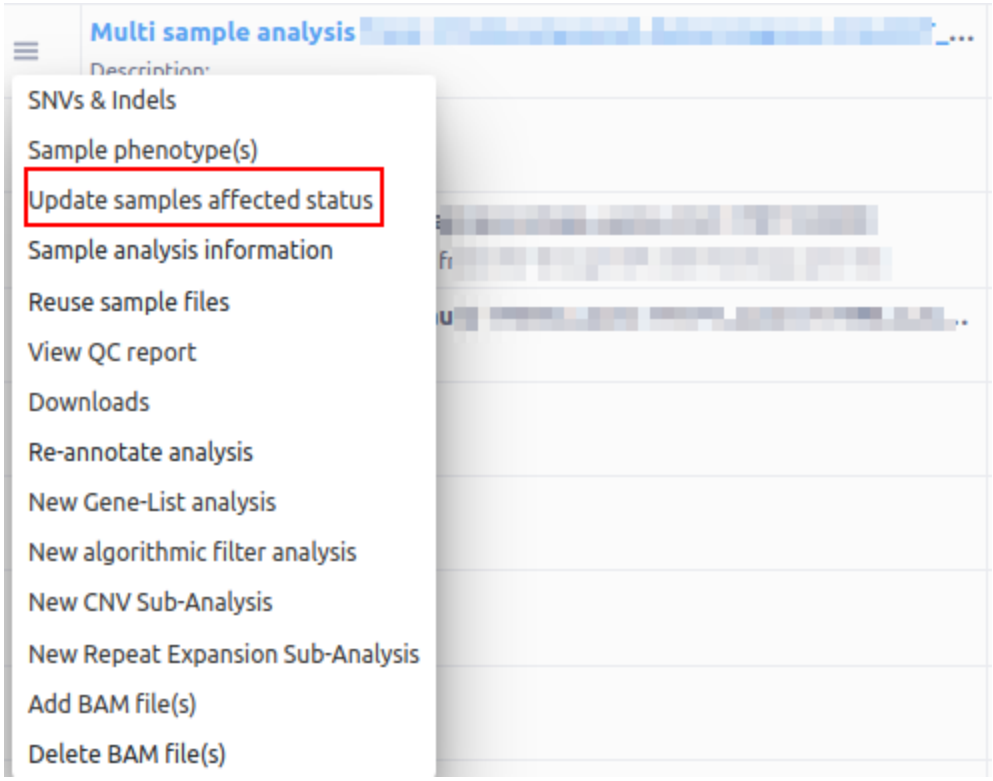
This column can be used to sort the table in descending/ascending order. You can also create a dynamic filter to filter variants based on these values following the steps below:



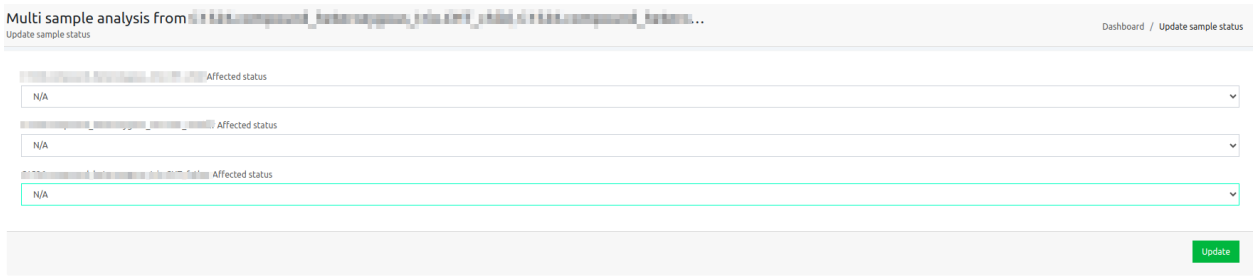


Update sample affected samples

It is possible to change the affected status of the merged analysis from the three horizontal lines next to the analysis name on the Analyses page.
This option is useful in order to run algorithmic filters for segregating variants, where you need to set which sample is the affected and unaffected respectively.



You will be directed to a screen, as the one shown below:



You can delete the filters displayed, then update the affected/unaffected status of your samples and run the algorithmic filters you would like for your merged analysis.

Sample/Analysis information

It opens a pop-up table with an overview of information about the analysis and the sample, the sequencing (e.g. capture method, number of reads and bases, sequencer), the analysis (e.g. the versions of the databases and the implemented pipelines and software), databases, classification (e.g. ACMG annotator version) and other useful metrics such as those mentioned below.

Useful metrics:

- Predicted Sex: As part of the main analysis, the sample sex is predicted based on zygosity in selected chromosome X SNVs. The algorithm assumes that females are XX and males are XY. Sex is predicted with a p-value (binomial test) <0.01, otherwise sex is undetermined. INDELs are not considered in the calculation due to their higher

false positive rate.

- Rare Homozygous Count: This metric provides the fraction of rare variants in homozygosity. A minimum of 50 genotypes are required to report results.
  - An SNV is considered rare when its frequency is below 1% or if it is unknown.
  - InDels are not considered in the calculation due to their higher false positive rate.

  The number of rare SNVs and their fraction in homozygosity is reported.
- Runs of Homozygosity: Samples are scanned for the presence of extended segments of homozygosity. Runs of homozygosity (**ROH**) are computed directly from the VCF file. These calculations are based on the information contained in the VCF file (either the user's or the one generated by VarSome Clinical) and specifically on the either genotype likelihoods (PL) or genotypes (GT). By default, genotype likelihoods are expected. ROH is calculated for different minimum region sizes (100Kb, 500Kb, 1.5Mb and 5Mb). For each size, we calculate the number of regions of that size or greater which show evidence of ROH.

  Please note:
  - ROH are computed for germline samples, but not for somatic
  - Only bi-allelic sites are considered in ROH calculations.
  - ROH are only reported for autosomal chromosomes.
  - All ROH have a minimum of 50 variants
  - INDELs are not considered in the calculation due to their higher false positive rate.
  - Minimum ROH quality (Phred score) is 20. ROH quality represents the probability of the state assignment being incorrect. Larger values indicate a more confident call.
  - ROH is calculated for all analyses. Please note, however, that ROH results for small, targeted panels are less reliable because of the sparseness of the target regions.

  The following ROH related metrics are reported for each length threshold in Analysis Actions > Sample/Analysis Information:

  NROH: The number of ROH sections detected in the sample

  SROH: The total length of ROH in bp

  maxROH: The length of longest ROH segment in bp

  FROH: The fraction of autosomal genome in autozygosity

  ROH are visible also in the Sample View.

Reuse sample files

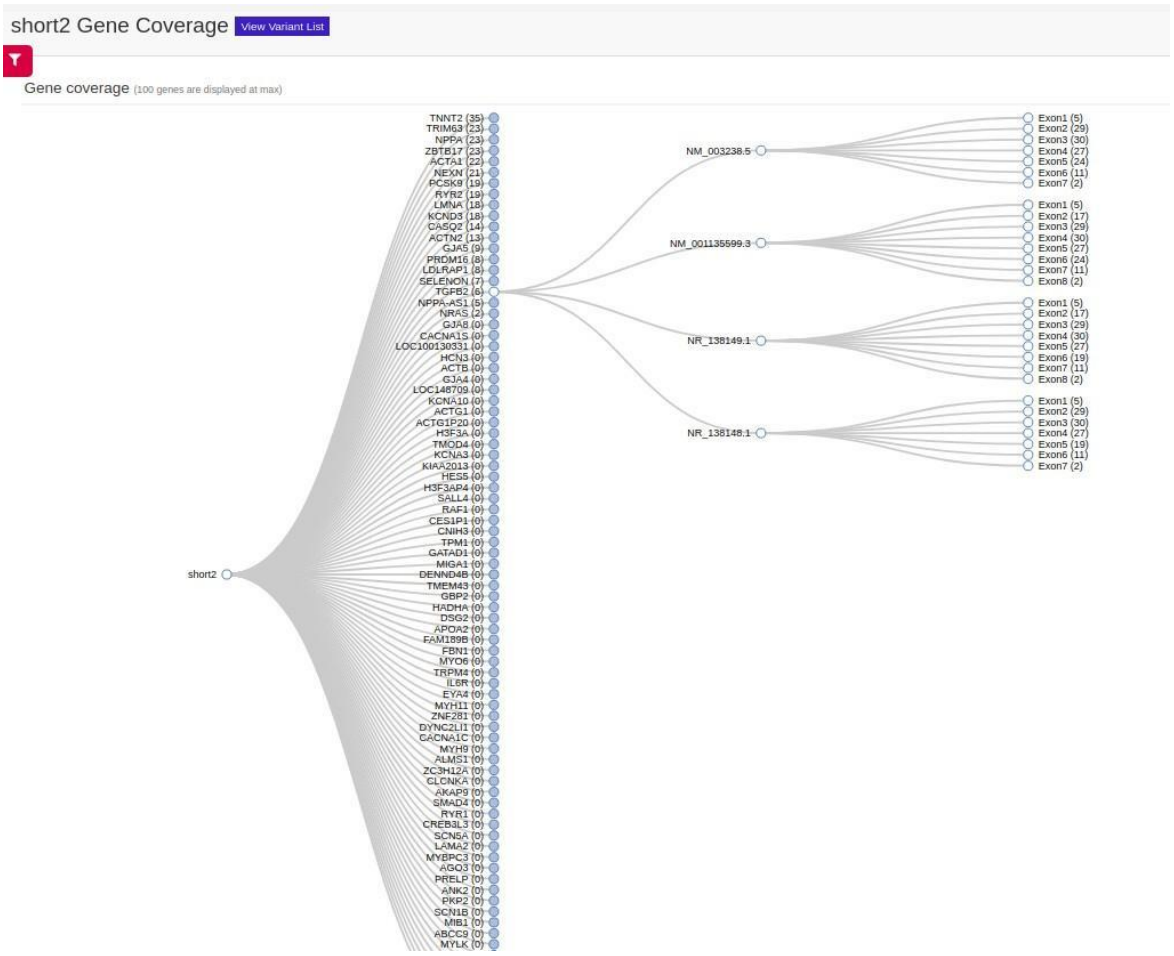
This option is available if you wish to reuse an existing sample for a new analysis. A blue box message in the upper right corner of the page will let you know that you will receive an email when the input file(s) of the sample is/are again available for new analysis and which you can view in the "Upload / view files" page.

Gene Coverage

When clicking on Gene Coverage, a new window will open showing a tree with the coverage per gene. ⚠️ Please note that only information for 100 or fewer genes is shown since the depiction of more genes would be impractical. If you wish to see coverage information for analyses with more than 100 genes, please use a gene list to limit the displayed results to 100 genes or fewer.

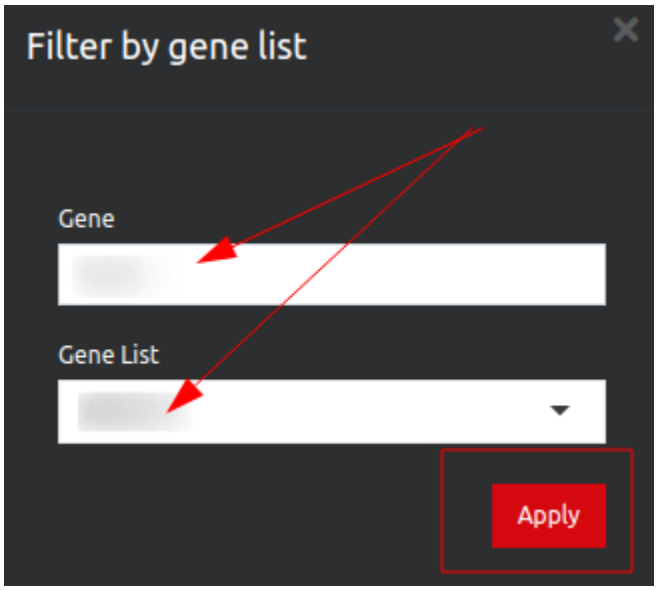
The gene list can be selected when the new window opens by clicking on 

. If this option has been selected when analyzing a gene list, the tree will contain only the genes from that gene list.



Clicking on any gene will result in a tree of its transcripts and each transcript is also expanded to a tree showing the coverage of individual exons. Clicking on an exon will open a new tab, showing the read coverage from the corresponding bam file.

If you click on the filter  icon on the top left corner this will take you to a new page, showing a long list of genes. You can filter for your gene of interest:



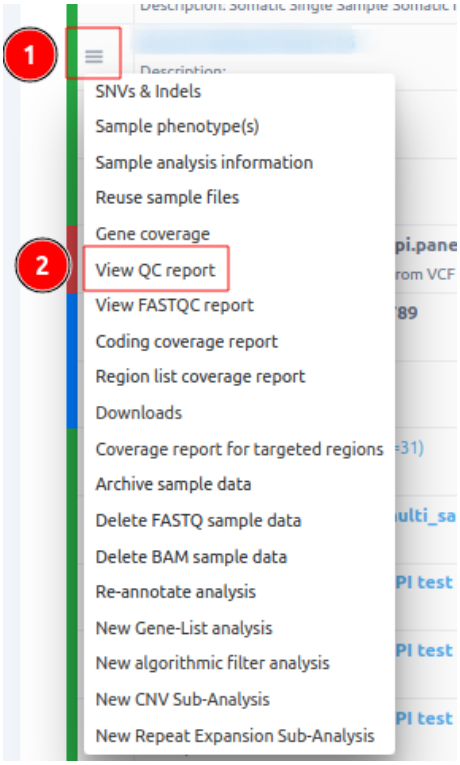
The following options concern the generation of different reports and can be useful when the user needs to check the quality of the sequencing in terms of coverage or retrieve information about the alignment.

View Quality Control report

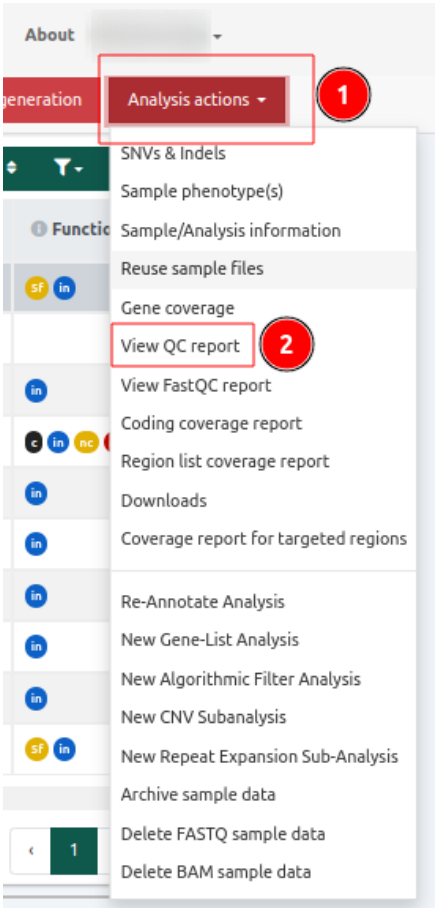
You can view a quality control report about your analysis and download it as a PDF file or export it as a docx file. The report includes a list of information such as sequence technology, read alignment results, regions reported, coverage, number of identified variants by class, summary for ACMG rules and number of SNV found in coding regions.
In order to find the Quality Control (QC) Report of an analysis in VarSome Clinical you can go

either:

- to the Analyses page and select from the three horizontal lines next to the analysis name you wish the "View QC Report" or



- in the Variant Table page you can click on the "Analysis actions" button and select the same option



A QC report includes all of the categories presented below:

General information

Simple information about the user, the reference genome, the sequencing technology, the file name and the type of pipeline used.

Software and Database versions

Information about Saphetor's and Sentieon's software versions used for alignment and variant calling and all available sources' versions that were used in the pipeline for annotation.

Read alignment results

This matrix reports information about the alignment of all reads and on the targeted regions of the selected assay.

Read alignment results for *short*

	All reads	On target
Total number of reads:	18 284	13 105 (71.67%)
Number of duplicates:	3 645 (19.94%)	2 789 (15.25%)
Number of reads singly mapped:	13 807 (75.51%)	10 103 (55.26%)
Number of reads unmapped:	18 (0.10%)	0 (0.00%)
Number of reads multiply mapped:	25 (0.14%)	2 (0.01%)
Number of reads with gapped alignments:	431 (2.36%)	212 (1.16%)
Number of reads below minimum quality:	383 (2.09%)	1 (0.01%)
Number of reads with insert size:	14 166 (77.48%)	10 295 (56.31%)
Average insert size:	138 bp	139 bp
Number with supplementary mapping:	0 (0.00%)	0 (0.00%)
Number out of mapping range:	0 (0.00%)	0 (0.00%)
Number of bases in reads:	2 211 519	1 611 185

Regions reported and Average region coverage

Display of information regarding the assay used for the analysis and the average coverage of the targeted regions.

Regions reported:	TruSight Cancer (1736 regions, 320536 bp)
Average region coverage:	76

Coverage

This matrix shows the coverage depths and percentages of all positions in total and of the targeted regions

Coverage	bp (All positions)	% (All positions)	regions (Target regions)	% (Target regions)
5x	309 140	96.4	1 483	85.4
8x	307 135	95.8	1 469	84.6
10x	306 320	95.6	1 467	84.5
20x	296 875	92.6	1 435	82.7
30x	275 785	86.0	1 346	77.5
50x	211 401	66.0	1 010	58.2
100x	82 567	25.8	397	22.9
300x	1 682	0.5	9	0.5

Variant type summary

Variant type summary

Type	Variants observed	Clinically reported		In dbSNP		In GnomAD		Novel*	
Total	**20 953**	**915**	**4.37%**	**15 767**	**75.25%**	**9 655**	**46.08%**	**5 051**	**24.11%**
SNV	12 894	200	1.55%	9 354	72.55%	3 809	29.54%	3 507	27.20%
Deletion	4 172	369	8.84%	3 565	85.45%	3 199	76.68%	530	12.70%
Insertion	3 135	345	11.00%	2 824	90.08%	2 647	84.43%	286	9.12%
Substitution	752	1	0.13%	24	3.19%	0	0%	728	96.81%

"Novel" refers to variants not found in dbSNP, gnomAD, or clinical databases.

Number of identified variants by ACMG class*:

The matrix presents the number of automatically classified variants in each of the 5 standard pathogenicity classes.

Number of identified variants by ACMG class*:

ACMG class	
Benign	23
Likely Benign	2
Uncertain Significance	0
Likely Pathogenic	0
Pathogenic	0
Total	**25**

* (NOTE this classification is of the variant itself, and doesn't say how it affects this patient. For example the variant can be Pathogenic but it may not be causative if it's recessive and the patient is heterozygous)

Summary for ACMG Rules

Display of the number of times (and its percentage) each Germline Variant Classification rule was triggered.

Summary for ACMG Rules

Total	**65**	
BA1	22	33.85%
BP4_Strong	21	32.31%
BP6_Moderate	6	9.23%
BP6_Very Strong	4	6.15%
BP4	3	4.62%
BP7	2	3.08%
PM2_Supporting	2	3.08%
BP3	1	1.54%
BP6	1	1.54%
BP6_Strong	1	1.54%
BS1	1	1.54%
BS2	1	1.54%

Variants found in Coding Regions

Variants found in Coding Regions :

Type	Total	In ClinVar		In dbSNP		In GnomAD	
SNV	4	3	75.00%	4	100.00%	4	100.00%
Insertion	1	0	0.00%	1	100.00%	1	100.00%

Type	Total	In ClinVar		In dbSNP		In GnomAD	
SNV	7	6	85.71%	7	100.00%	4	57.14%
Deletion	1	0	0.00%	1	100.00%	1	100.00%
Insertion	1	1	100.00%	1	100.00%	1	100.00%

Quality Control report for CNV analyses

The CNV quality control report is a useful tool to evaluate the performance of a CNV analysis and to check the correlation between the control samples chosen for each test sample in a given cohort.

It can be accessed from the CNV main analysis as shown above, and provides the following information:

- SampleId: VarSome Clinical unique ID assigned to the sample
- User sample name: the name given by the user for each test sample.
- Median fragment count: It is the median fragment count in each genomic interval of the assay .bed file. These are data generated by exomedepth and used to ensure that a sufficient number of reads is present in each sample. Samples should have a similar range of values (same order of magnitude).
- Number of reference samples: number of reference samples used as controls for this test sample.
- Reference sample names: the names of the samples used as a reference (control) set.
- Correlation: correlation coefficient between the test sample and its reference samples.
- Sex: sex of the samples in the cohort

SampleId	User sample name	Median fragment count	Number of reference samples	Reference sample names	Correlation	Sex
37127	NA12878_WGS_PIPE_262	1225	1	ERR174310_WGS_PIPE_262	0.9847	F
37141	ERR174310_WGS_PIPE_262	3438	2	NA24385_WGS_PIPE_262 NA12878_WGS_PIPE_262	0.9781	M
37142	NA24385_WGS_PIPE_262	8722	1	ERR174310_WGS_PIPE_262	0.9738	M

We highly recommend users check the CNV QC report first. This report provides useful information on how the reference set of samples was generated for each of the samples of the cohort. It also checks that the median coverage of each sample was sufficient.

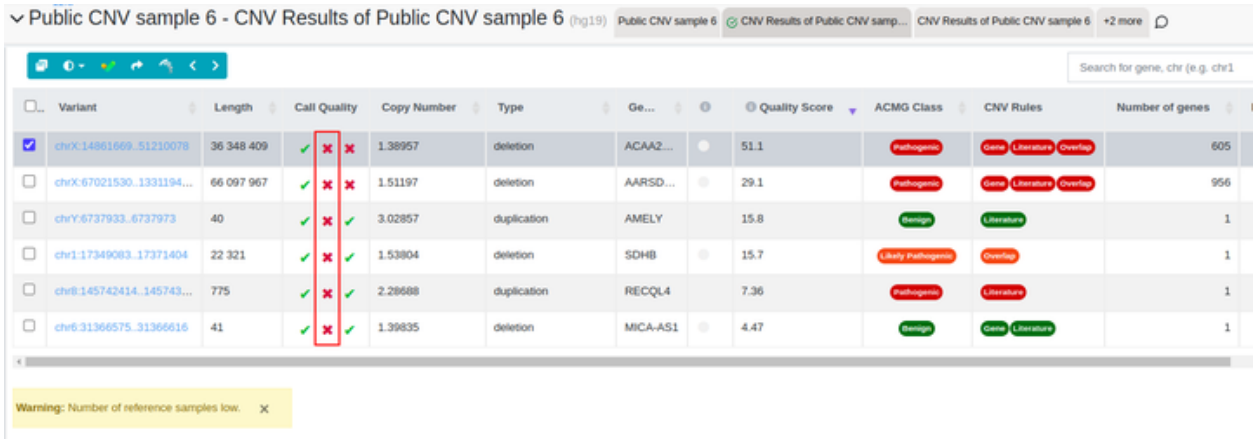
Each input sample is compared against an optimized set of reference samples chosen from among the other samples of the cohort. The set of reference samples is automatically generated by the algorithm for each sample of the cohort (test sample). ⚠️ Please bear in mind that the reference set of samples might not include all the other samples of the cohort. The reference set is chosen based on how well the coverage correlates with that of the test sample. The algorithm will choose the largest set of samples from the reference with the greatest correlation coefficient.
A sample will not be included in the reference set if it results in a decrease of the correlation coefficient.

In the CNV QC report, the user can check how many samples were used to construct the reference set for each test sample and their correlation coefficient. A high number of samples (at least 2) used for the reference set and a high correlation coefficient (> 0.97 for gene panels and > 0.98 for exomes) would be indicative of a good performance and a higher reliability of the CNVs found in that test sample. If the coefficient of correlation is too low for a test sample, the algorithm won't be able to generate meaningful results, and no CNVs will be called in that test

sample.

Calling CNVs on the sex chromosomes can create issues if the test sample and the reference samples it is being compared to are not gender matched. To make reliable calls on the X and Y chromosomes, you need to make sure that all samples are of the same sex.

The information from the three different metrics that can be found on the column call quality control can be used in combination with the CNV QC report.
For example, when the second call quality metric (number of reference samples) fails for all variants of the variant table:



we should go to the QC report and check the size of the reference set. In this example, the reference set used to call CNVs on "Public CNV sample 6" has only one sample and it is therefore below the threshold (2 samples), and this is why the second call quality metric fails for all variants. In this case, we suggest repeating the CNV analyzes with other samples that could be better correlated with the test sample.

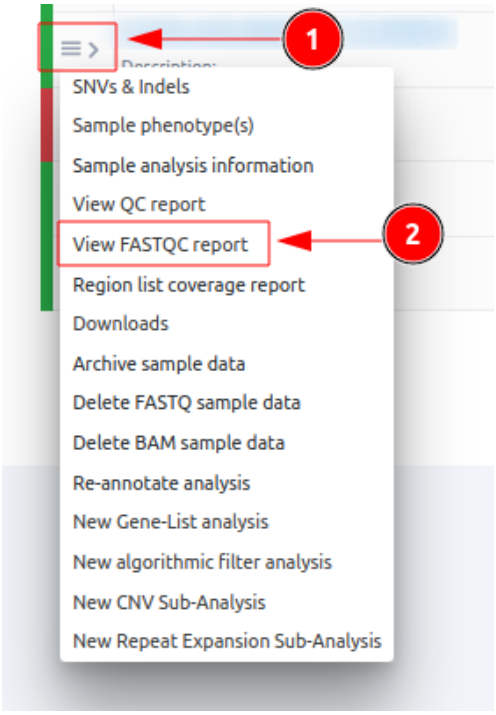
Sample name	Coverage (Median)	Number of reference samples	Reference sample names	Correlation	Gender (Predicted)	Sex (User defined)	wsex (Consensus)
Public CNV sample 1	604	2	Public CNV sample 3 Public CNV sample 2	0.9975	_	_	_
Public CNV sample 2	673	3	Public CNV sample 4 Public CNV sample 1 Public CNV sample 7	0.9988	_	_	_
Public CNV sample 3	739	2	Public CNV sample 1 Public CNV sample 4	0.9976	_	_	_
Public CNV sample 4	784	3	Public CNV sample 3 Public CNV sample 2 Public CNV sample 9	0.9987	_	_	_
Public CNV sample 5	1256	1	Public CNV sample 8	0.9954	_	_	_
Public CNV sample 6	1099	1	Public CNV sample 8	0.9882	_	_	_

View FASTQC report

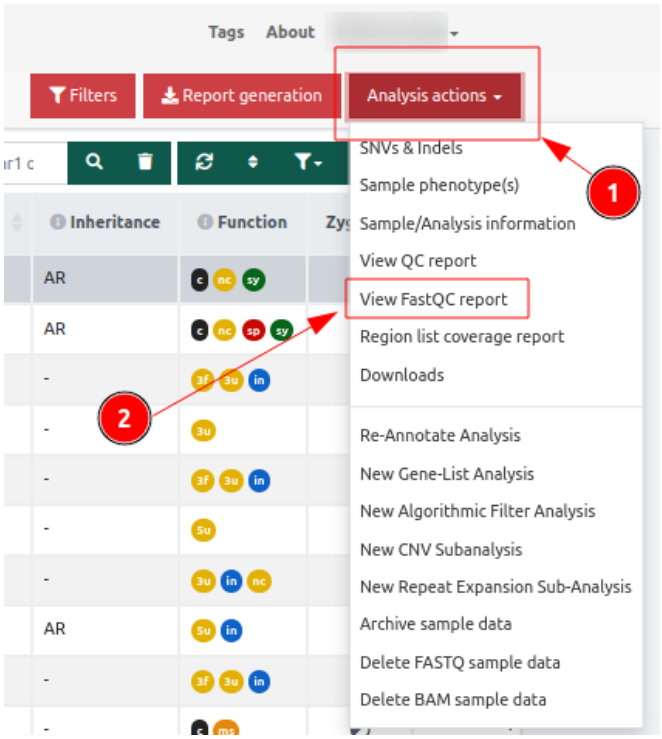
A quality control report for high throughput sequence data. For more information please see the documentation of the FASTQC tool.

In order to find the FastQC report of an analysis in VarSome Clinical you can go either:

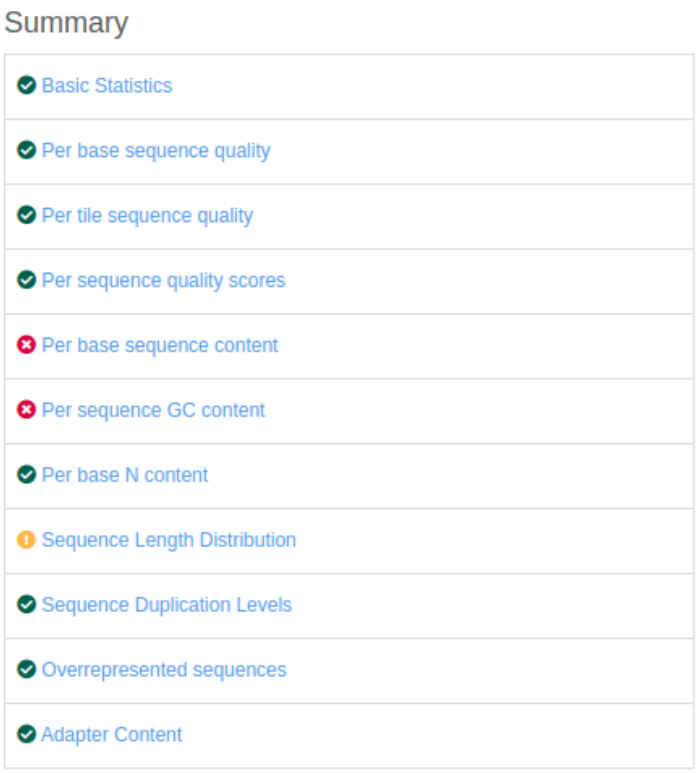
- to the Analyses page and select from the burger menu of the analysis you wish the "View FastQC Report" as shown below, or



- in the Variant Table page you can click on the "Analysis actions" button and select the same option.

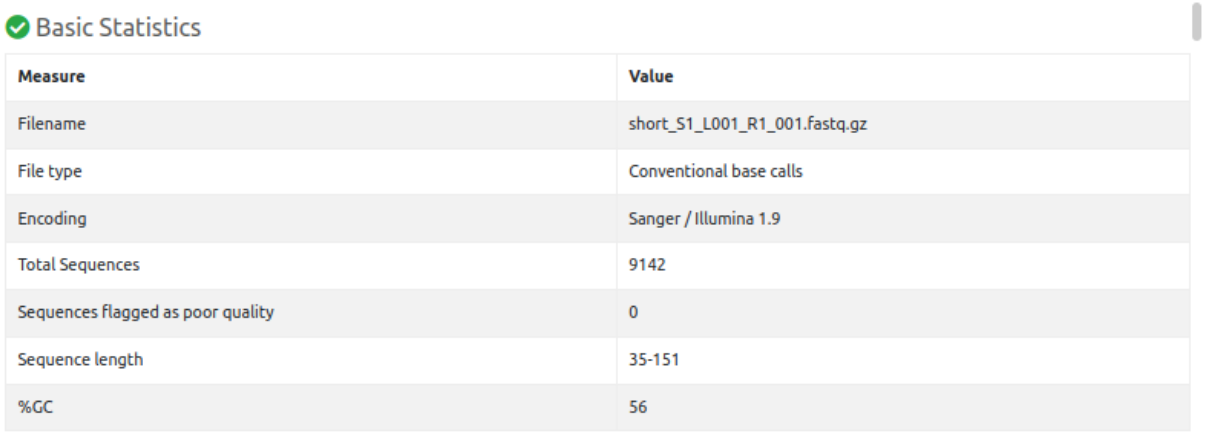


The different analyses (modules) performed for the FASTQC report are summarized on the left of the screen, for each FASTQ file used for the analysis.



The symbol on the left of each module corresponds to a flag of "Passed", "Warn" or "Fail".
Here, we include examples for each module with what should be expected as a result:

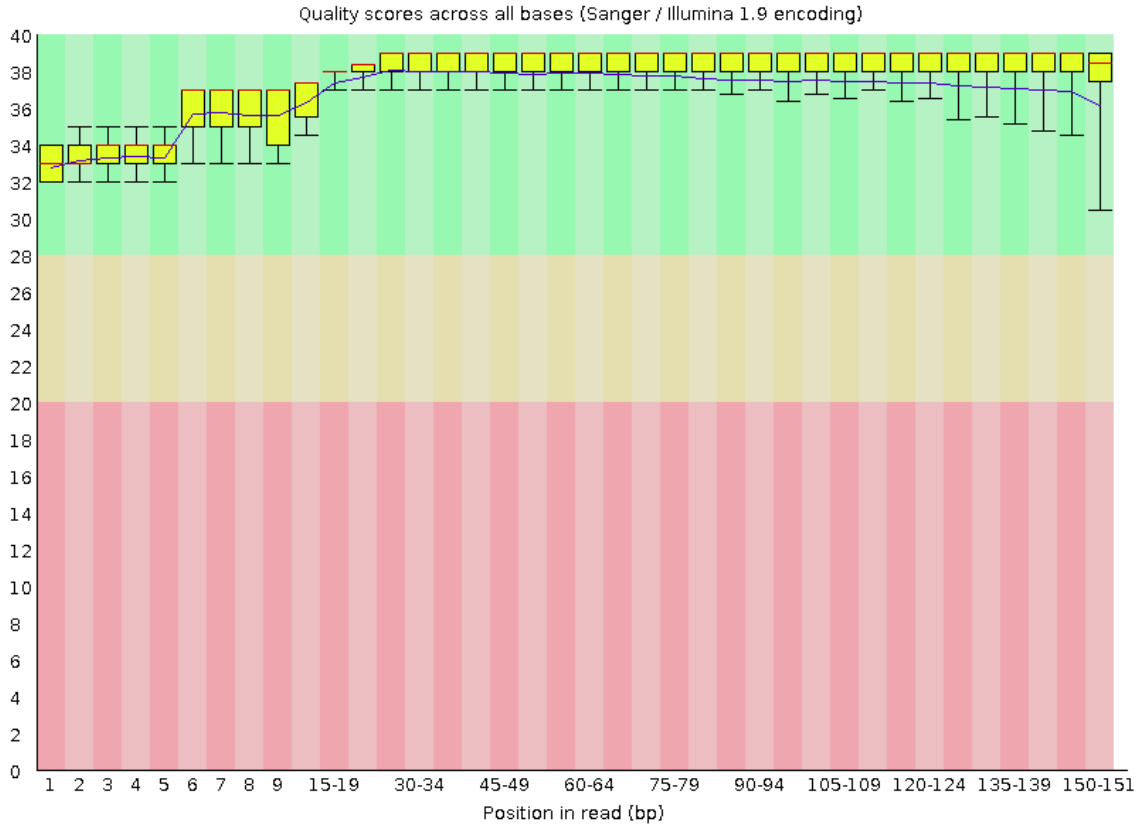
Basic Statistics
Simple information about input FASTQ file: its name, type of quality score encoding, total number of reads, read length and GC content.


Basic Statistics

Measure	Value
Filename	short_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	9142
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	56

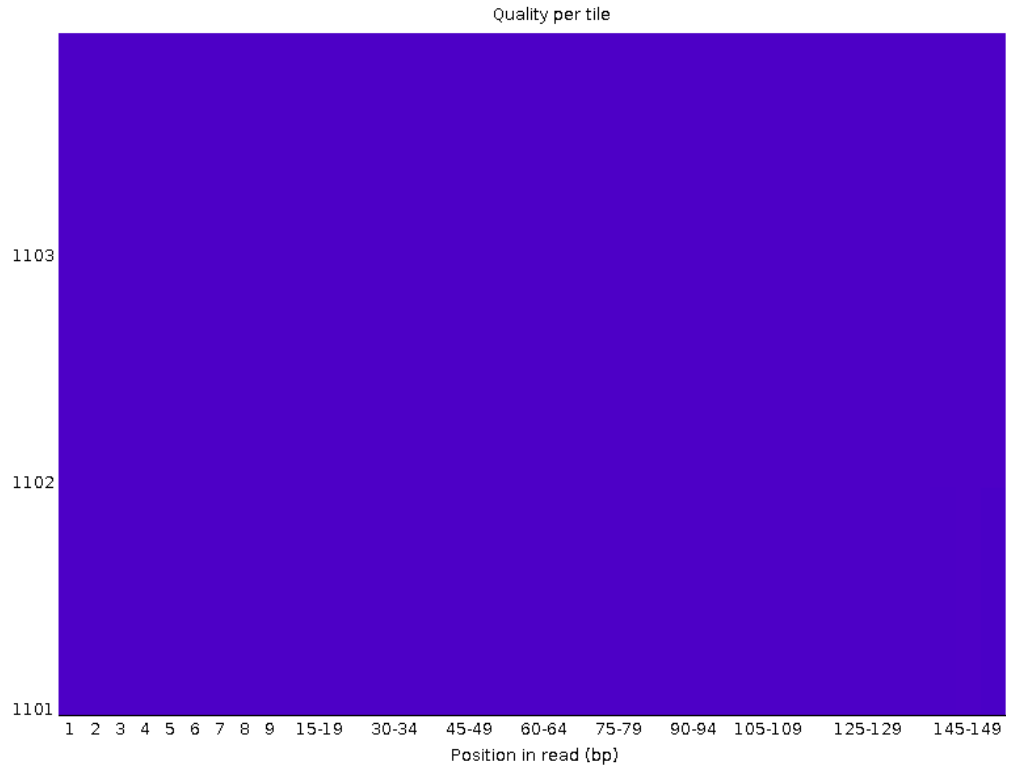
Per base sequence quality

A box plot showing aggregated quality score (Phred score) statistics at each position along all reads in the file.

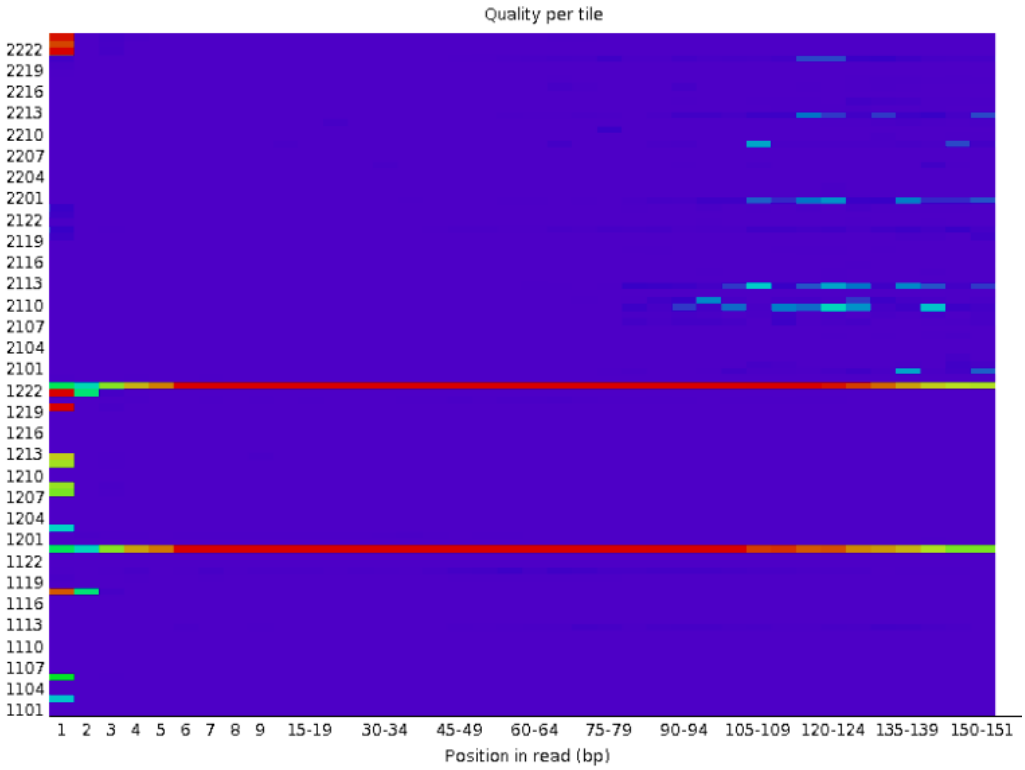


Per tile sequence quality

The graph allows you to look at the average quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flow cell.

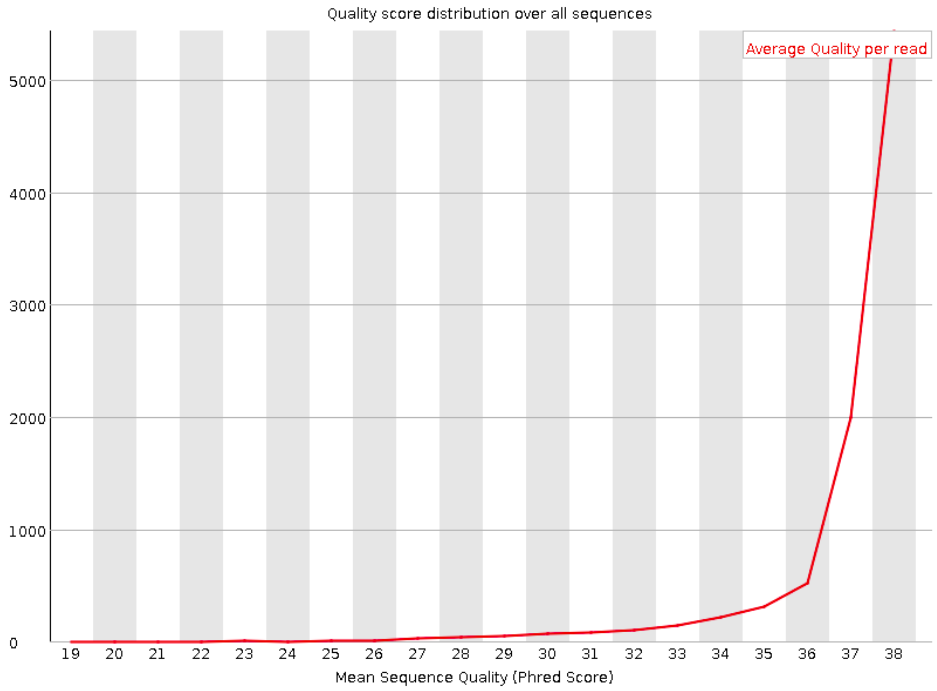


The picture above shows an ideal case, where no quality loss is reported in any tile of the flow cell. On the contrary, in the picture below you can see that certain tiles show consistently poor quality. The colors are on a cold to hot scale, with cold colors being positions where the quality was at or above the average for that base in the run, and hotter colors indicate that a tile had worse qualities than other tiles for that base.



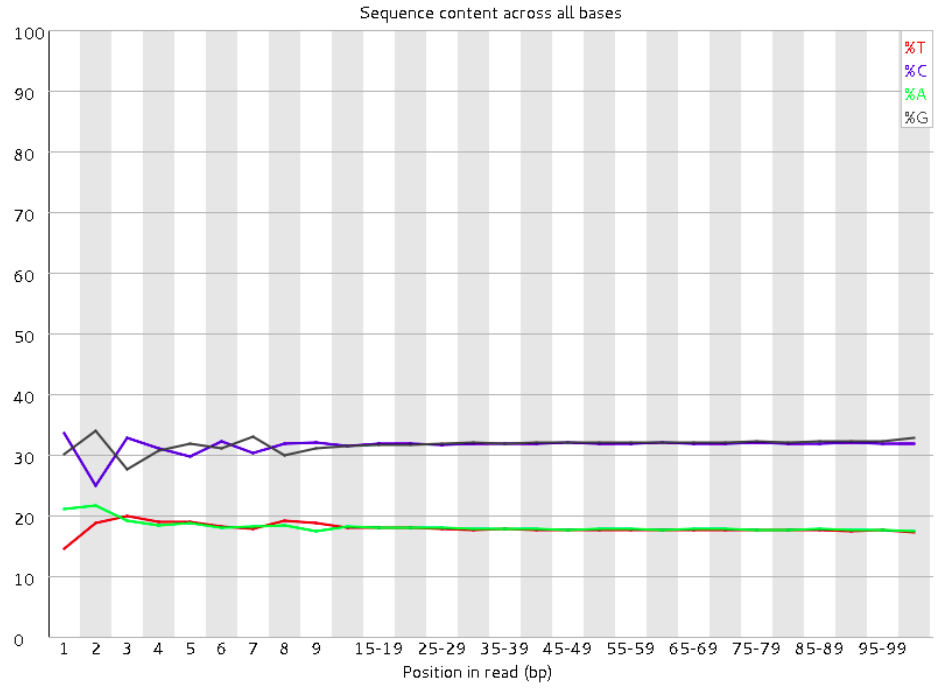
Per sequence quality scores

A plot of the total number of reads vs the average quality score (Phred score) over the full length of that read.



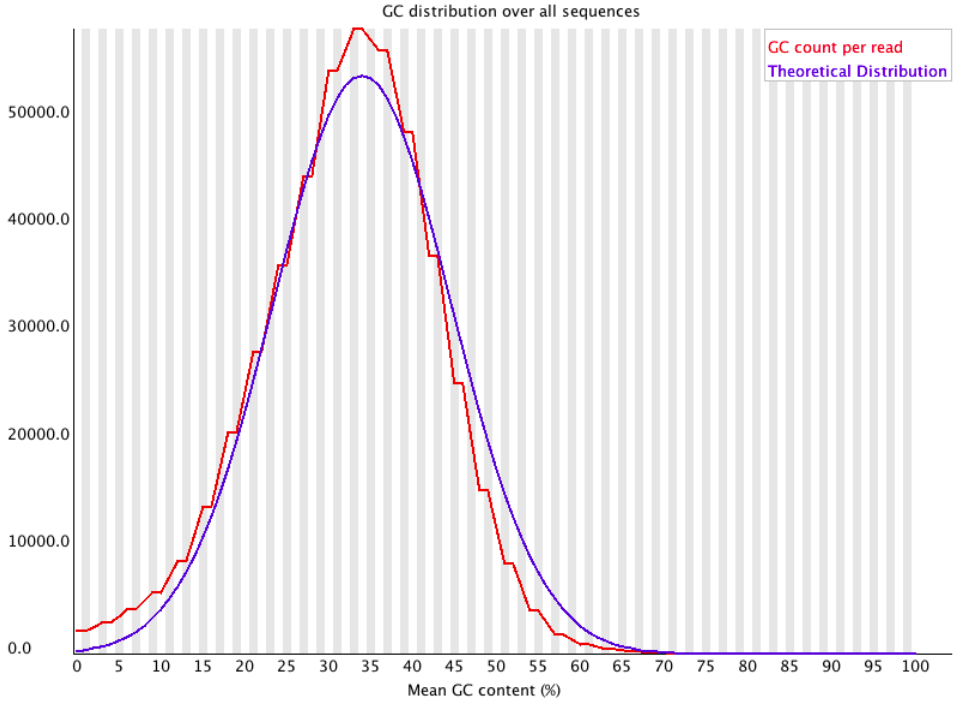
Per base sequence content

This plot reports the percent of bases called for each of the four nucleotides at each position across all reads in the file.



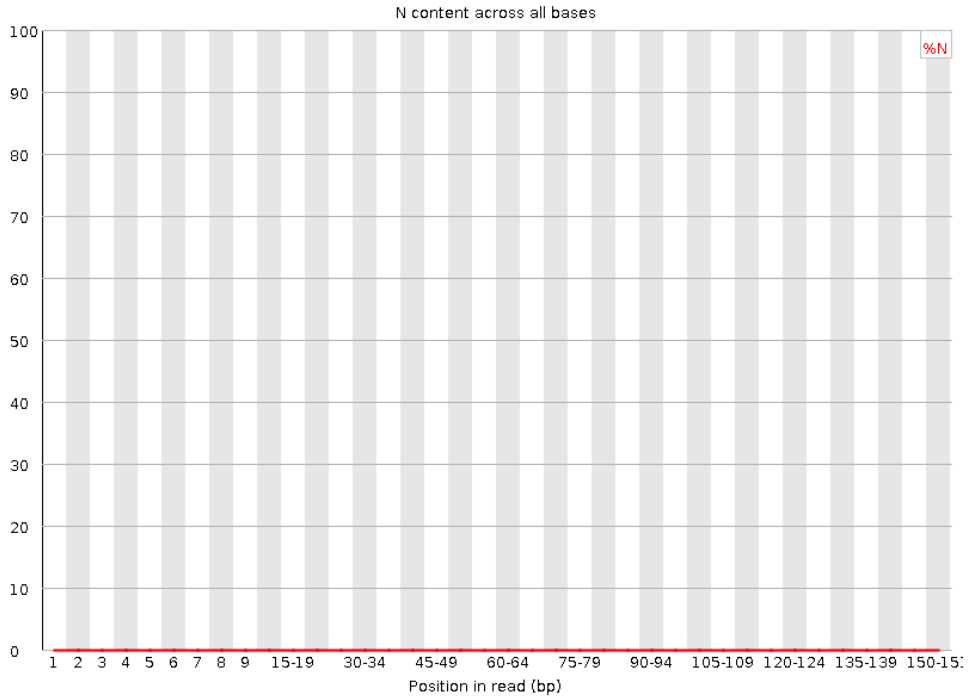
Per sequence GC content

Plot of the number of reads vs. GC% per read. The displayed Theoretical Distribution assumes a uniform GC content for all reads.



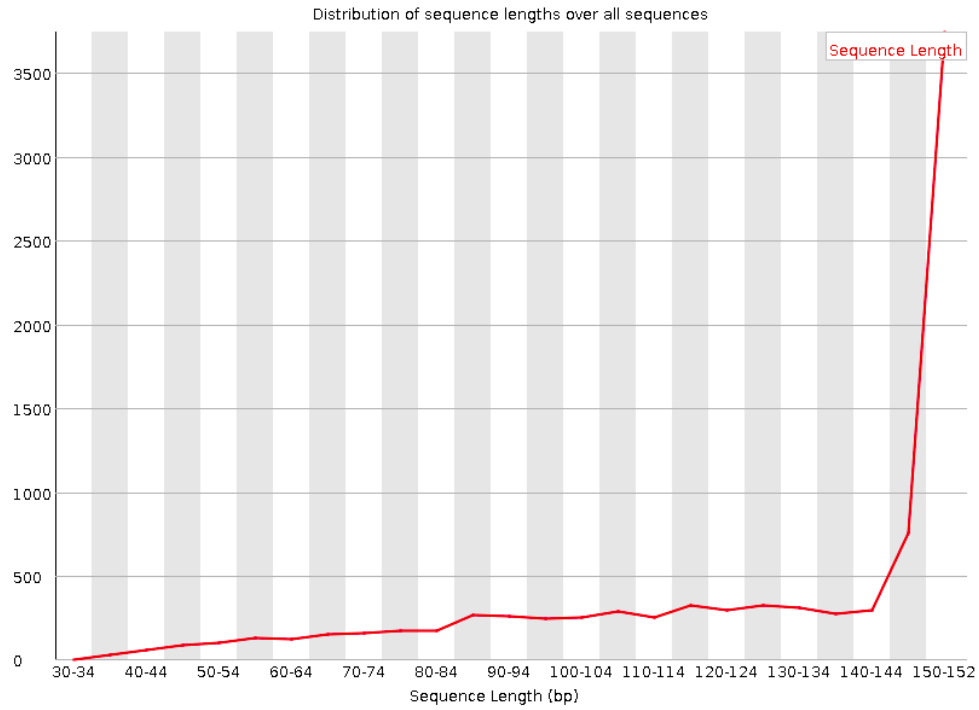
Per base N content

Percent of bases at each position or bin with no base call, i.e. 'N'



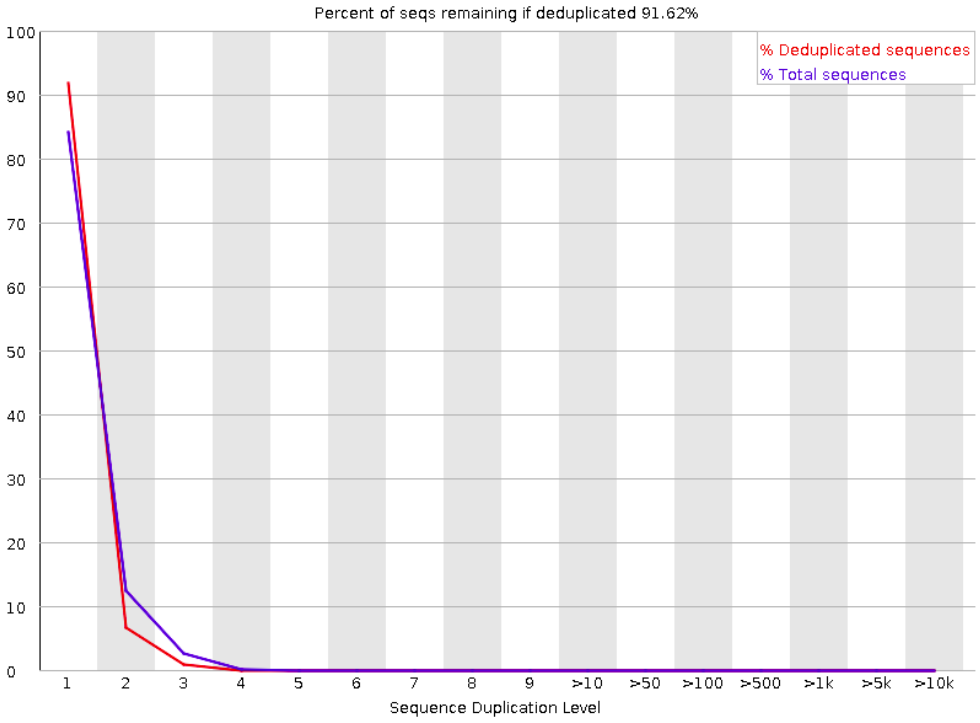
Sequence Length Distribution

Shows the distribution of reads lengths over all sequences.



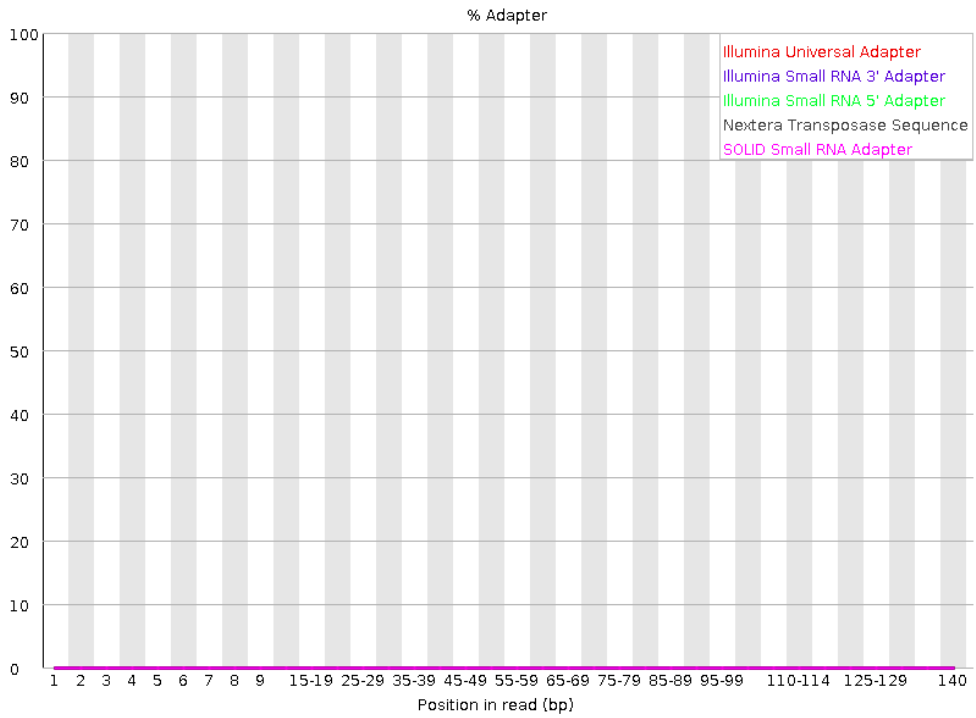
Sequence Duplication Levels

Percentage of reads of a given sequence in the file which are present a given number of times in the file.



Overrepresented sequences
List of sequences which appear more than expected in the file.

Adapter Content
Cumulative plot of the fraction of reads where the sequence library adapter sequence is identified at the indicated base position.



⚠️ Please note that if an error or warning is reported by FASTQC, VarSome Clinical will also provide a warning to the users. For more information see User messages.

Coding coverage report

You can use a previously created gene lists to produce an Excel document reporting coverage information of the coding regions included in the analyzed gene list. This report can be exported for all component samples of a multi-sample analysis.

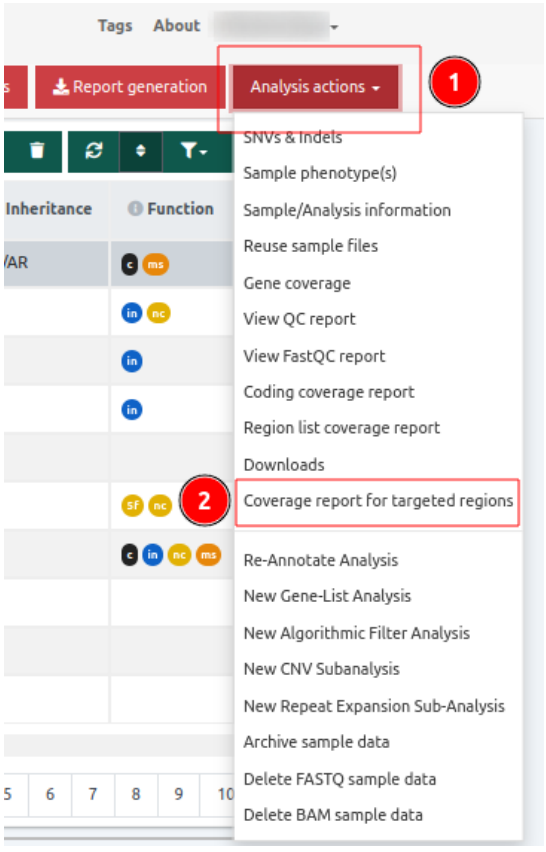
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Gene	Region	Coding Length	Mean Cover	#bp<5x	#bp≥5x	%Cover≥5x	#bp<10x	#bp≥10x	%Cover≥10x	#bp<20x	#bp≥20x	%Cover≥20x	#bp<30x	#bp≥30x	%Cover≥30x
2	ACTC1	chr15:35086881..35087009	129	100	0	129	100	0	129	100	0	129	100	0	129	100
3	ACTC1	chr15:35085446..35085770	325	321	0	325	100	0	325	100	0	325	100	0	325	100
4	ACTC1	chr15:35084609..35084770	162	269	0	162	100	0	162	100	0	162	100	0	162	100
5	ACTC1	chr15:35084291..35084482	192	294	0	192	100	0	192	100	0	192	100	0	192	100
6	ACTC1	chr15:35083315..35083496	182	181	0	182	100	0	182	100	0	182	100	0	182	100
7	ACTC1	chr15:35082613..35082756	144	223	0	144	100	0	144	100	0	144	100	0	144	100
8	ACTC1	total	1134	249	0	1134	100	0	1134	100	0	1134	100	0	1134	100
9	ACTN2	chr1:236849974..236850099	126	207	0	126	100	0	126	100	0	126	100	0	126	100
10	ACTN2	chr1:236881158..236881272	115	148	0	115	100	0	115	100	0	115	100	0	115	100
11	ACTN2	chr1:236882194..236882313	120	344	0	120	100	0	120	100	0	120	100	0	120	100
12	ACTN2	chr1:236883405..236883491	87	248	0	87	100	0	87	100	0	87	100	0	87	100

Coverage report for targeted regions

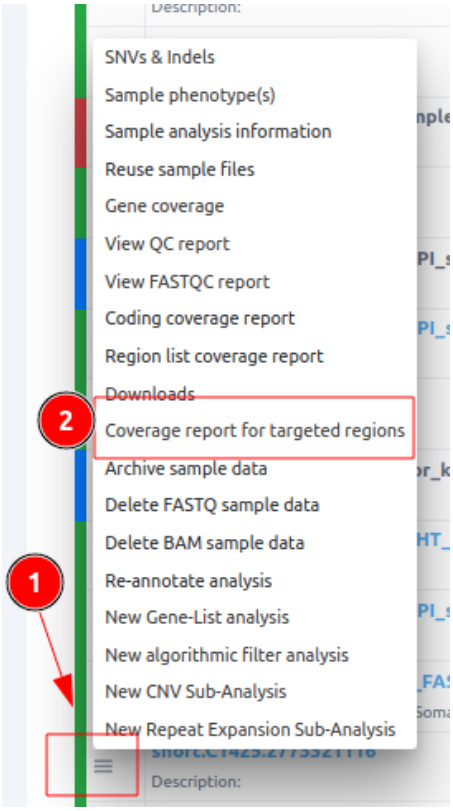
Generates an Excel document for the coverage of the regions captured by the specific assay that you selected when launching the analysis.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Coverage report on targeted regions																				
2	Region	Coding Le	Mean Cove	Minimum	#bp<5x	#bp≥5x	%Cover≥5	#bp<8x	#bp≥8x	%Cover≥8	#bp<10x	#bp≥10x	%Cover≥1	#bp<20x	#bp≥20x	%Cover≥2	#bp<30x	#bp≥30x	%Cover≥3	#bp<50x	#bp≥50x
3	chr1:69091	918	0	0	918	0	0	918	0	0	918	0	0	918	0	0	918	0	0	918	0
4	chr1:62109	939	0	0	939	0	0	939	0	0	939	0	0	939	0	0	939	0	0	939	0
5	chr1:86132	72	0	0	72	0	0	72	0	0	72	0	0	72	0	0	72	0	0	72	0
6	chr1:86553	182	0	0	182	0	0	182	0	0	182	0	0	182	0	0	182	0	0	182	0
7	chr1:86641	51	0	0	51	0	0	51	0	0	51	0	0	51	0	0	51	0	0	51	0
8	chr1:87115	125	0	0	125	0	0	125	0	0	125	0	0	125	0	0	125	0	0	125	0
9	chr1:87442	90	0	0	90	0	0	90	0	0	90	0	0	90	0	0	90	0	0	90	0
10	chr1:87469	186	0	0	186	0	0	186	0	0	186	0	0	186	0	0	186	0	0	186	0
11	chr1:87652	163	0	0	163	0	0	163	0	0	163	0	0	163	0	0	163	0	0	163	0
12	chr1:87751	116	0	0	116	0	0	116	0	0	116	0	0	116	0	0	116	0	0	116	0
13	chr1:87779	79	0	0	79	0	0	79	0	0	79	0	0	79	0	0	79	0	0	79	0
14	chr1:87793	500	0	0	500	0	0	500	0	0	500	0	0	500	0	0	500	0	0	500	0
15	chr1:87863	125	0	0	125	0	0	125	0	0	125	0	0	125	0	0	125	0	0	125	0

You can find it under the “Analysis actions” tab, as shown below,



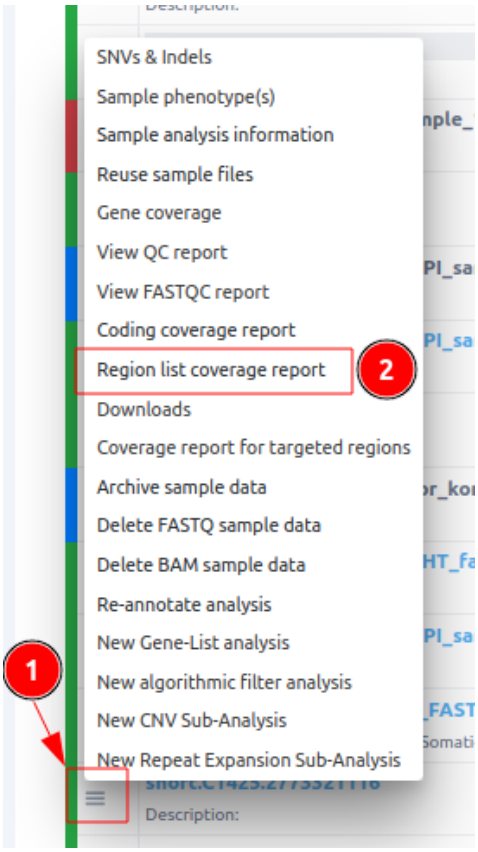
or on the Analyses page if you follow the steps shown below:



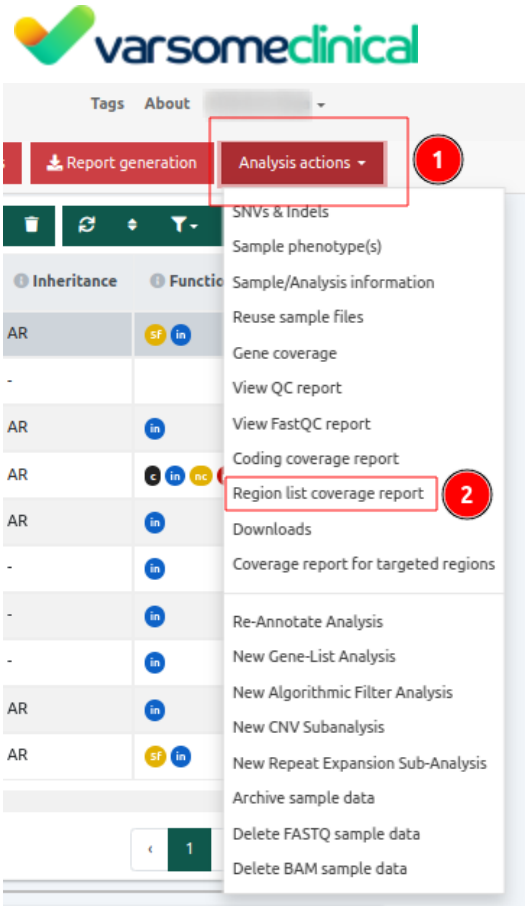
Region list coverage report

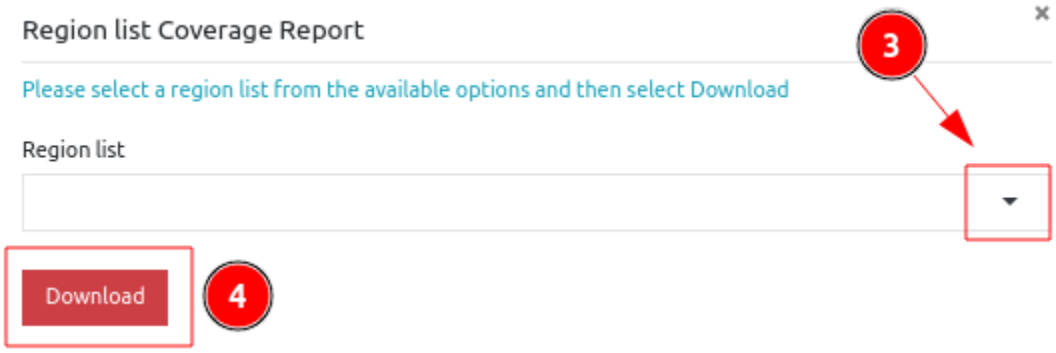
You can use a list of public/custom regions of interest to produce an Excel document reporting coverage information about these regions. Contact us in case you want to add regions to your analysis. In order to find the Region list coverage report of an analysis in VarSome Clinical you can go either:

- to the Analyses page and select from the burger menu of the analysis you wish the "*Region list coverage report*" or



- in the Variant Table page you can click on the "Analysis actions" button and select the same option





Downloads

This option directs you to a new screen where you can see any files associated with this analysis that are ready to be downloaded.

Sample Data Files:

- **VCF file**: A compressed (*.vcf.gz) VCF file will be downloaded with the results of the variant calling.
  - For sub-analysis (gene list analysis and algorithmic filters) there is the new option to “Generate VCF” that contains **only** the filtered variants.
- **BAM file:** Download the bam file (the sample's reads aligned against the reference genome) used in the analysis. For multi sample analyzes you will find the BAM (and BAI) files for each component sample.
- **BAI file:** The BAI file format is the index file of a BAM file. This is a companion file for

your previous BAM file, which doesn't contain any sequence data but acts as an external table of contents allowing a computational tool to navigate in the BAM file and locate specific parts.

Quality control File:

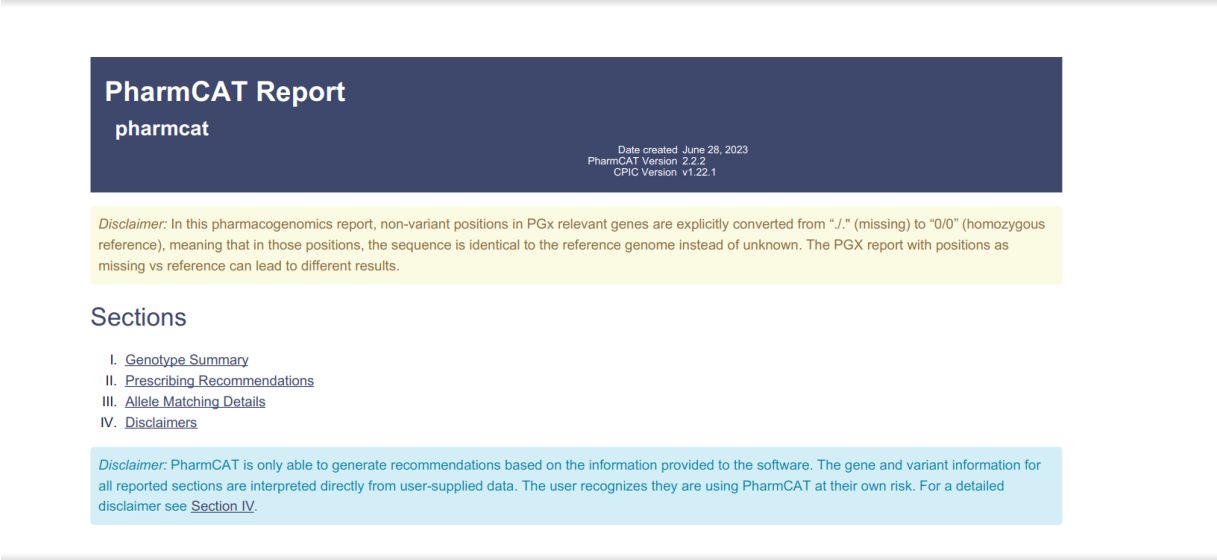
- **QC report (PDF)**
- **QC report (docx)**

Download PGX report

PharmCAT is a pharmacogenomics clinical annotation tool that generates a report containing genotype-based prescribing recommendation. This option is available for WGS samples, run against hg38 and using one of the WGS capture methods: 'WGS+PCR' or 'WGS-PCR'.

- The report comprises 4 sections:
  - **Genotype Summary:** This section has a summarized table of the matched genotypes and the following associated clinical annotations: Drug, Gene, Allele Functionality and Phenotype
  - **Prescribing Recommendations:** This section includes a list of recommendations per associated drug, along with bibliographical reference material
  - **Allele Matching Details:** Detailed information about how data in the sample VCF matches up with haplotype definitions
  - **Disclaimers:** PharmCAT disclaimer (in blue) as stated in their website and the disclaimer about the default running parameters in Varsome Clinical (in yellow), as shown below.

.


PharmCAT Report

pharmcat

Date created June 28, 2023
PharmCAT Version 2.2.2
CPIC Version v1.22.1

Disclaimer: In this pharmacogenomics report, non-variant positions in PGx relevant genes are explicitly converted from "./." (missing) to "0/0" (homozygous reference), meaning that in those positions, the sequence is identical to the reference genome instead of unknown. The PGX report with positions as missing vs reference can lead to different results.

Sections

I. Genotype Summary
II. Prescribing Recommendations
III. Allele Matching Details
IV. Disclaimers

Disclaimer: PharmCAT is only able to generate recommendations based on the information provided to the software. The gene and variant information for all reported sections are interpreted directly from user-supplied data. The user recognizes they are using PharmCAT at their own risk. For a detailed disclaimer see Section IV.

Typically in WGS samples, a large fraction of the PGX relevant positions that are considered by pharmCAT are absent from the input vcf. If those missing positions are assumed to be "no-call", the resulting PGx report results in the output of multiple possible genotypes, and hence less specific treatment recommendations.

In VarSome Clinical, we assume that all missing PGx relevant positions are homozygous

reference. However, this may not reflect reality, because such positions may in fact be unreadable or uncallable. Running PharmCAT with positions as missing vs reference can lead to different results.

Re-annotate Analysis

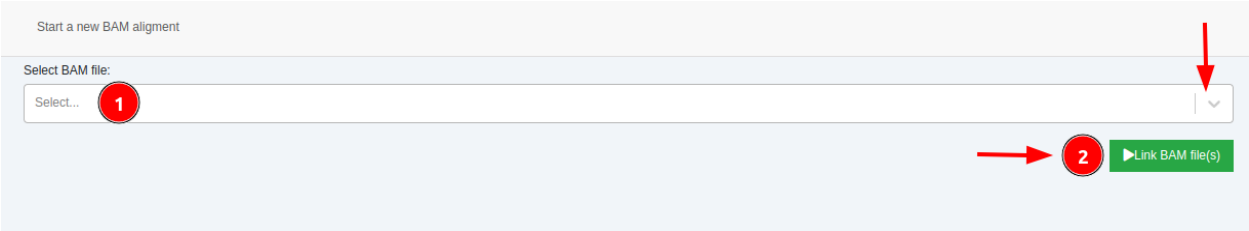
Reannotating will not change the list of identified variants (we do not perform the calling again), it will simply update the annotations for the same list of variants using the data available on the day of re-annotation. This means that some Germline Variant Classification might change if they are affected by newer data. Re-annotating an analysis is charged at 50% of the price of the original analysis.

New Gene-List Analysis

This option will re-start the analysis with the same sample using a previously generated gene list of selected genes or a gene list created from selected phenotype(s) or disease(s). It will appear as a sub-analysis of the initial analysis.

Add BAM file(s)

Select a BAM file for alignment visualization. If your sample type is VCF, you can upload and add a BAM file to visualize the alignment.



Once the BAM file is added to a sample, IGV and JBrowse links will be displayed on the "Coverage" column of the Variant Table. These links will allow you to visualize the alignments in each variant position.

⚠️ Please note that the BAM file will be taken into account when calculating storage fees. For more details please read the Storage Management section.

New Algorithmic Filter Analysis

It opens a new screen with the available algorithmic filters, as seen in Algorithmic filters section. Please click on the info icon, to get a description of each filter. Algorithmic filters create a snapshot of the parent analysis, with a filtered subset of variants, according to each Algorithmic filter. You can run a sub-analysis choosing a filter from the list or you can ask for a custom made filter.

New CNV Sub-analysis

Launching a CNV analysis from VCF as seen in CNV annotation (from VCF).

New Repeat Expansions Sub-analysis

Launching a CNV analysis from VCF as seen in Repeat expansion annotation from VCF file**.**

Archive sample data

The analysis will be archived and a VCF file will be stored.

Delete FASTQ sample data

Allows to delete the fastq files used for the analysis. Once this option has been used, this will be disabled meaning that the files have been deleted from the server.

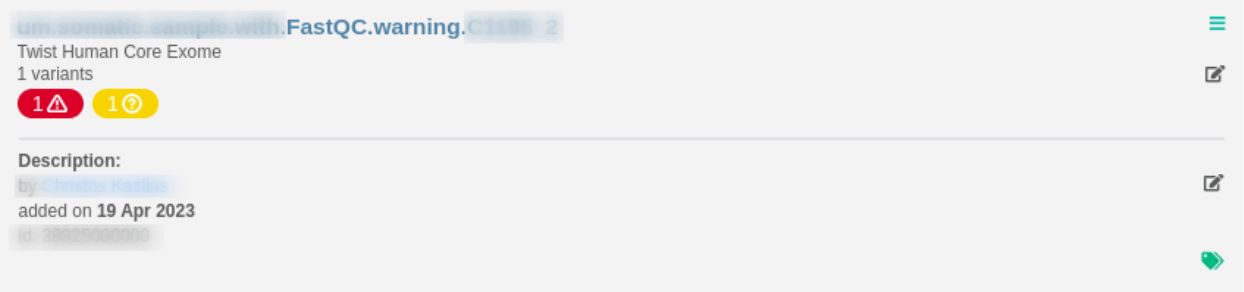
Delete BAM sample data

Allows to delete the bam file produced by the alignment during the analysis. Once this option has been used, this will be disabled meaning that the files have been deleted from the server.
For more information on storage fees please refer to this article: Data Deletion and Data Storage Fee explained.

6.2 User messages

VarSome Clinical introduces a monitor to highlight quality issues related to user input data and ensure that they won't be missed. The easily accessible display consists of badges of different colors based on their severity (errors-red, warnings-amber and information-blue). These messages are visible both on "Dashboard"/"Analyses" and "Variant table" menus, per sample and provide useful information about the status of each sample during the quality control step.

Quality control is performed at the FASTQ (FASTQC tool) and the BAM level, to ensure that the raw data is good enough to reliably analyze the results. For each analysis, warning or error icons may appear on the "Analyses" / "Dashboard" menu to inform the users that the quality control of the FASTQ file produced warnings and/or errors.



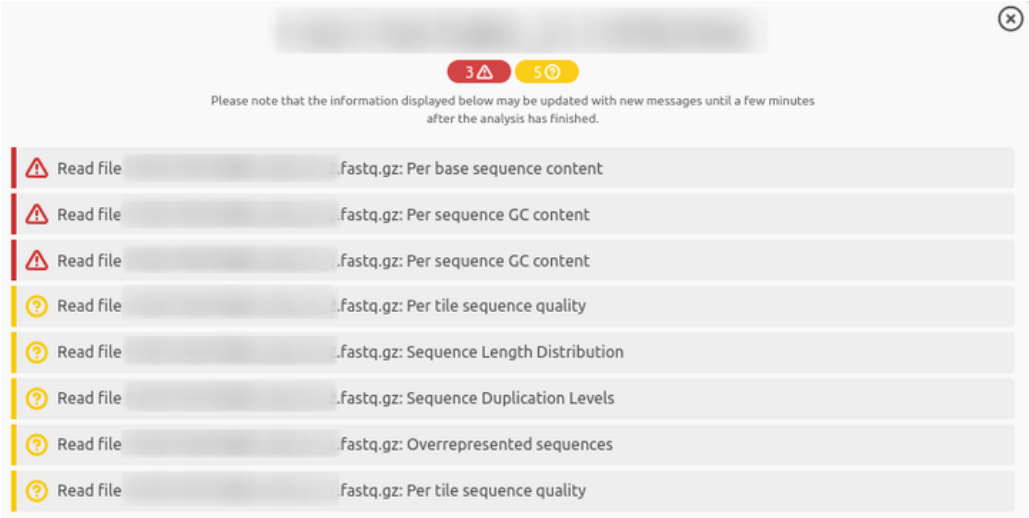
To reduce the number of error/warning messages displayed on the “Analyses”/ “Dashboard” menu and improve their practicality, FASTQC errors have been divided into three groups according to their frequency: Frequent, Standard and Rare. The Frequent messages include Per base sequence content as well as Per sequence GC content. Frequent errors will not display any warning or error message, therefore their messages will not appear as warnings or errors on the "Analyses" / "Dashboard" menu.

On the other hand, Standard errors include:

- Per tile sequence quality
- Overrepresented sequences
- Adapter Content
- Sequence Duplication Levels

and will display a single warning when there are more than two present in the report.

Note: Sequence Duplication Levels will be ignored when the analysis uses an amplicon kit.



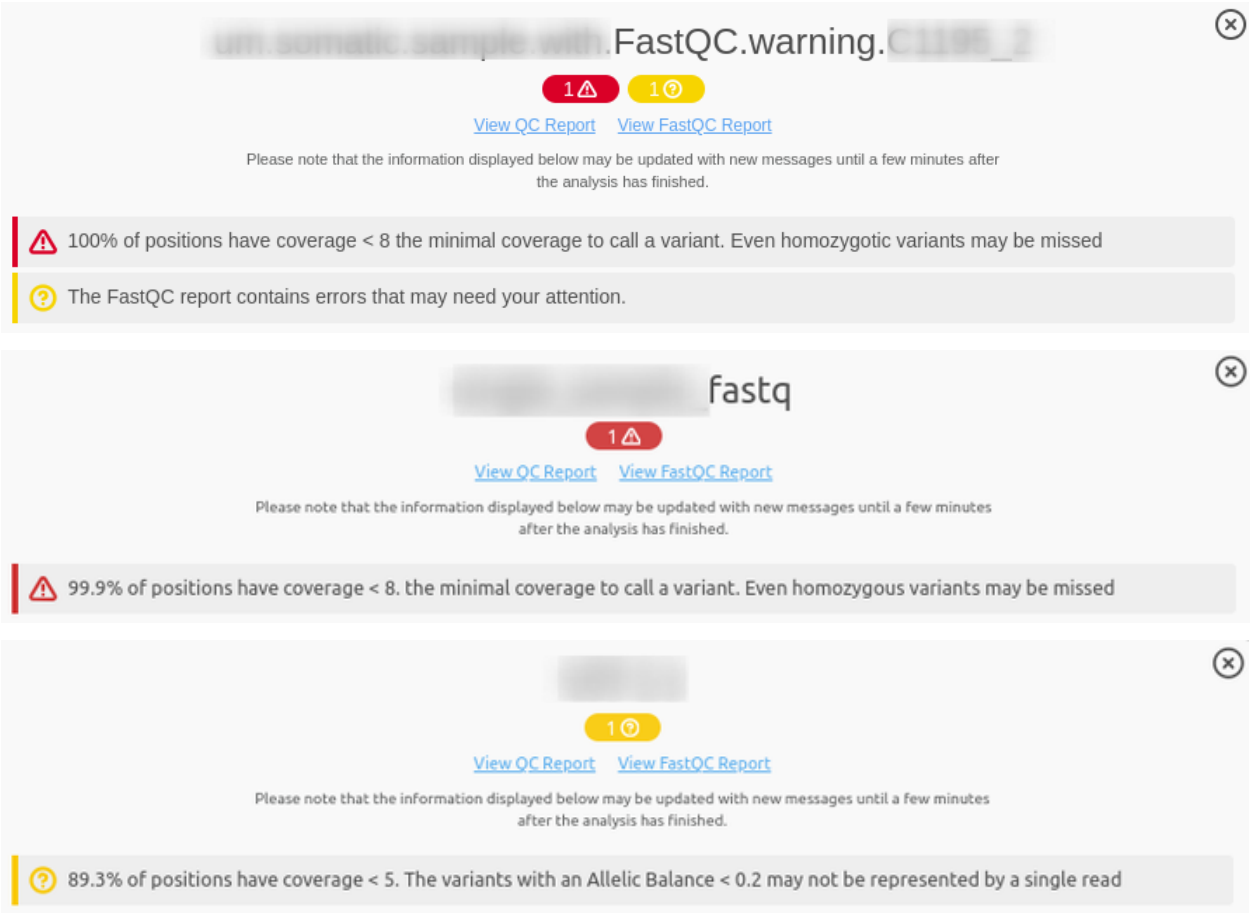
Finally, Rare errors include:

- Per base sequence quality
- Per base N content
- Per sequence quality scores
- Sequence Length Distribution

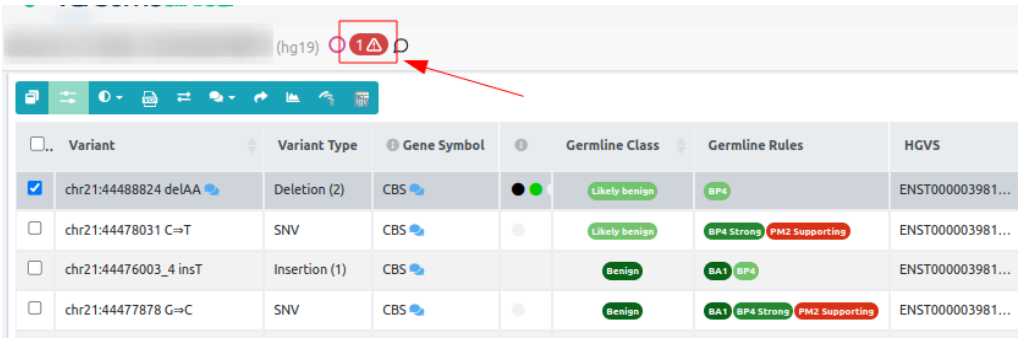
and will display an error message when there is one or more of them present in the report.

To know more about what parameters of FASTQC have raised the flag, you can click on the

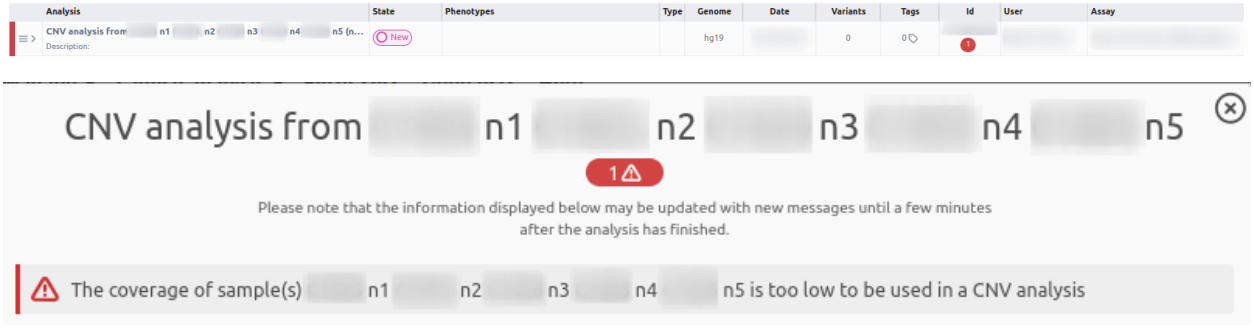
icons, and a pop-up window will appear showing you a detailed description.



The same information is also available in the variants table for an individual analysis:



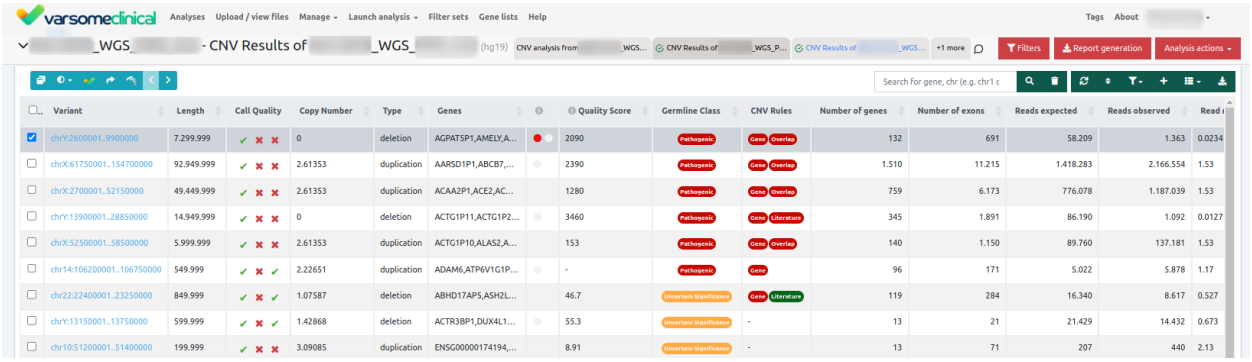
In the same way, there are some quality checks run at the alignment level that will produce warnings and/or error messages in case there are issues with the coverage of the sample. Additionally, if one sample included in a CNV cohort analysis does not have enough coverage, the whole CNV analysis will fail and an error message will be displayed informing about the problematic sample.

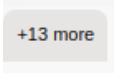


6.3 Tabs

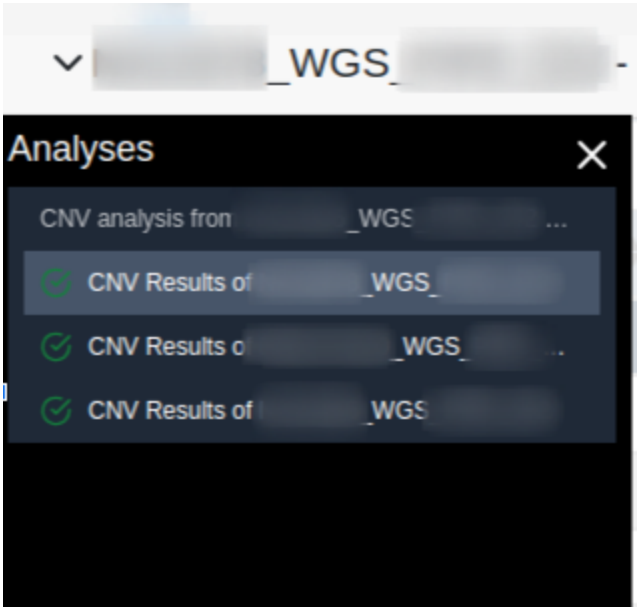
If multiple analyses have been performed using the same sample, they will be organized as tabs adjacent to the current analysis's title, with the first tab being the main analysis.

When there are more than three analyses, a total of four tabs will be displayed. These include the main analysis, the two most recent sub analyses, and a tab indicating the remaining number of subanalyses.



To view and access all the sub analyses associated with the main analysis, users can click on the arrow located to the left of the analysis title  or on the last tab with the title “+ n more”, where n is equal to the remaining number of sub analyses (for example 

). This action will reveal the analysis tree, presenting all the analyses in chronological order.



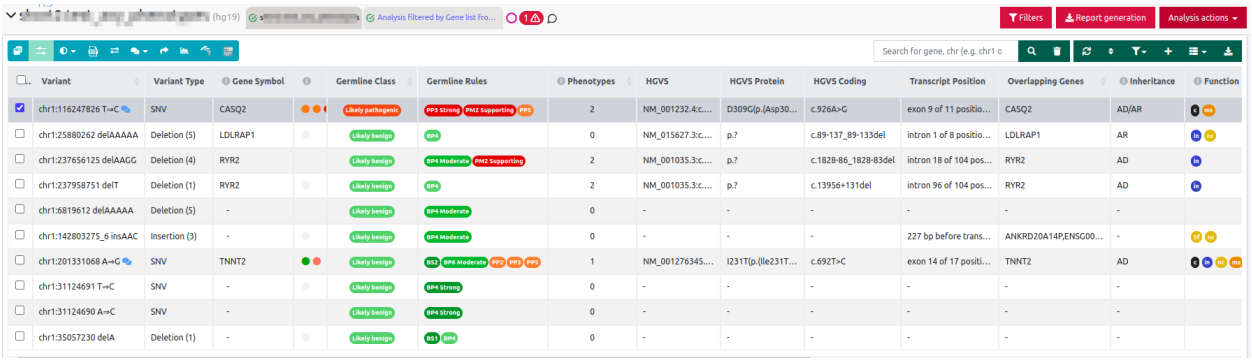
7. Variant table

The results are displayed in the variant table. Rows contain the identified variants, and columns contain core annotations for each variant (Variant, Variant type, Class, Genes, Function, Zygosity, Allelic balance, Coverage). However, none of the columns is mandatory, you can choose the ones to be displayed using the "Columns" icon . The length of each column of the variant table can be adjusted by dragging the sides of the column headers.

Tip: The column order in the variant table is user-specific, meaning each user can set up a custom order and visibility of columns.

You can hover over the info icon next to the column names to display the column information. The variant table is sorted by the Germline Variant Classification by default showing the most pathogenic variants first on the list. You can use the icon to sort the variant table by different columns (e.g. phenotypes, variant position, coverage) in ascending or descending order. Use the “Reset/refresh” icon to return the variant table to its original state.

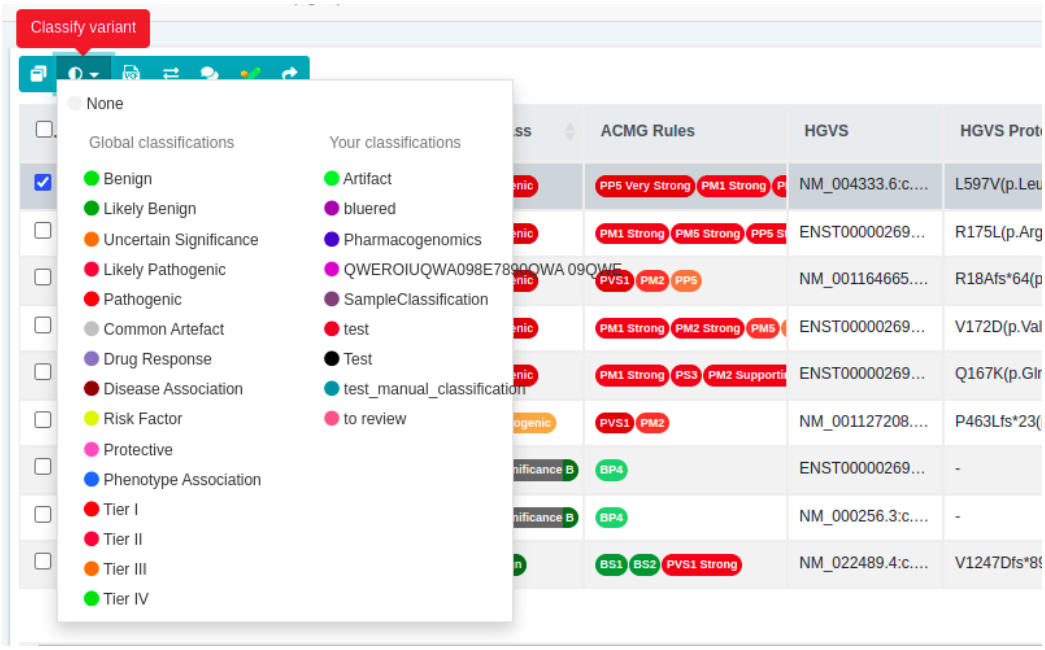
The variant table can be accessed by the user who requested the analysis or by other people belonging to the same group.



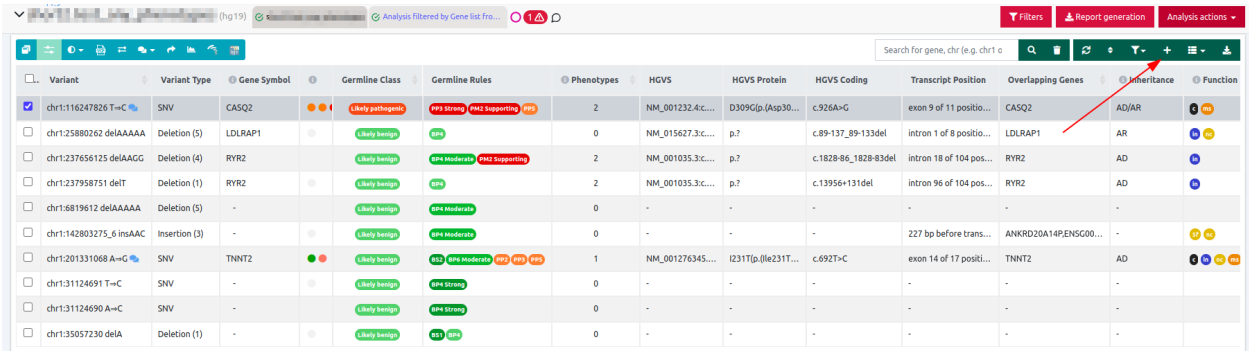
7.1 Description of results page functionalities

Columns for Germline/Somatic samples:

- **Variant:** The variant's sequence and genomic location.
- **Variant type**: SNV (single nucleotide variant); for INDELs and substitutions, the number of nucleotides affected are shown.
- **Gene Symbol:** Gene used for annotation and classification of the variant for ACMG (& AMP for somatic samples)
- **User variant classification:** custom classification for variants marked by the user. User classifications are also available for AMP and for ACMG rules when the user clicks on “Save as manual classification” below the ACMG/AMP verdict. The custom classifications are linked to the variant and will be displayed in other analyses of a user's group if the same variant is found.
  More specifically, you can classify a variant using either any of the five standard classes of pathogenicity ("Benign", "Likely benign", "Uncertain significance", "Likely pathogenic" and "Pathogenic") or any of the pre-set classifications as shown below, by clicking on the  icon:

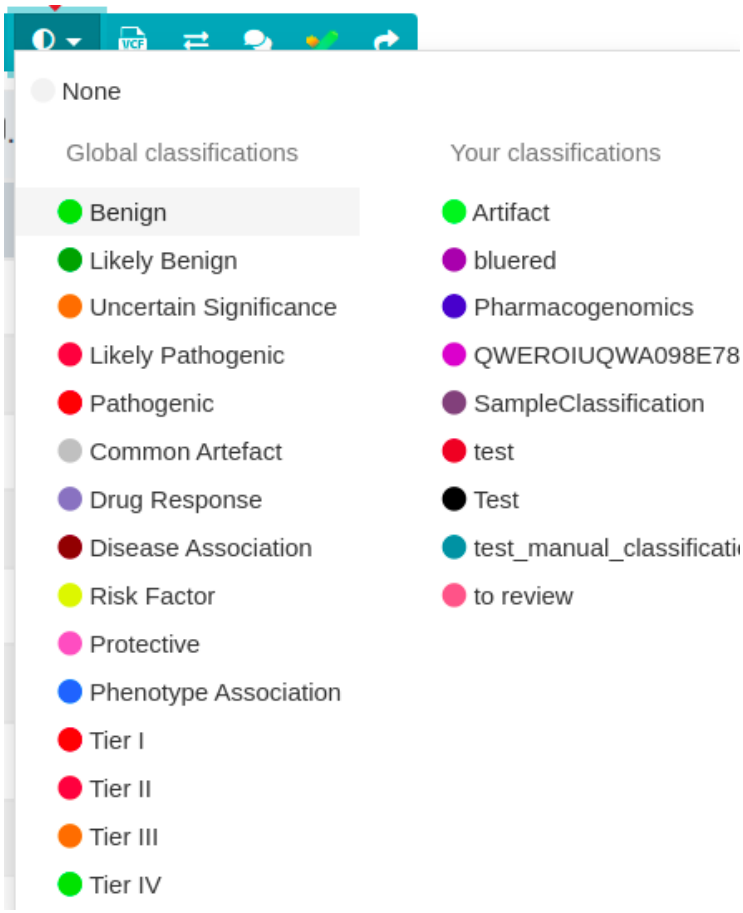


You can also create your own classification tags. To do this, click on the + icon located above the variant table:



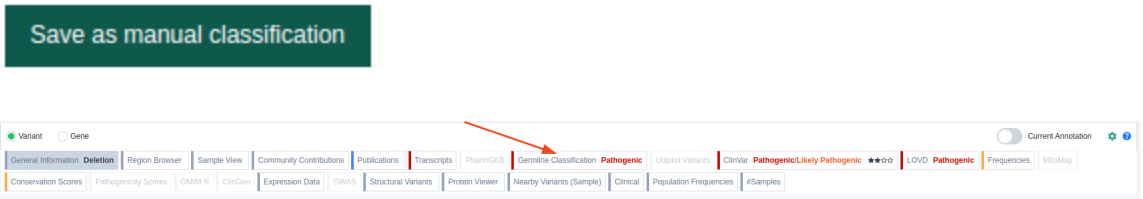
A pop-up window will be shown where you can add your custom tags. You need to give a name to the tag and a unique classification code. Finally, you can choose a color for your tag and then just click "Save".

The new custom classification will be available together with rest of the options:

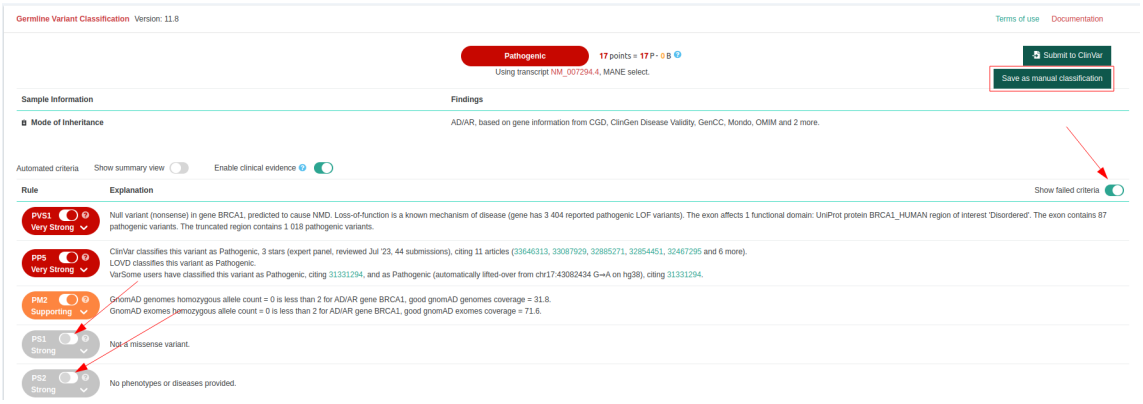


Custom Germline & Somatic Variant Classifications

The Germline and Somatic Variant Classifications allow users to modify the set of triggered rules if they do not agree with the verdict. Once you have modified the set of rules, you click on the "Save as manual classification" icon.



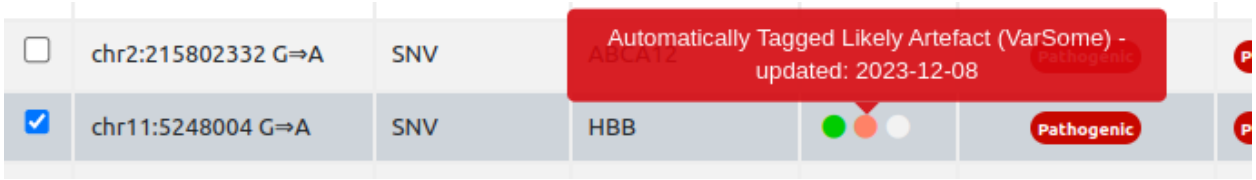
This will be saved as a custom classification attached to the variant. If you hover over with the mouse you will be able to see the manual classification with the set of rules used and the user who made the classification (in brackets).



Potential artifacts

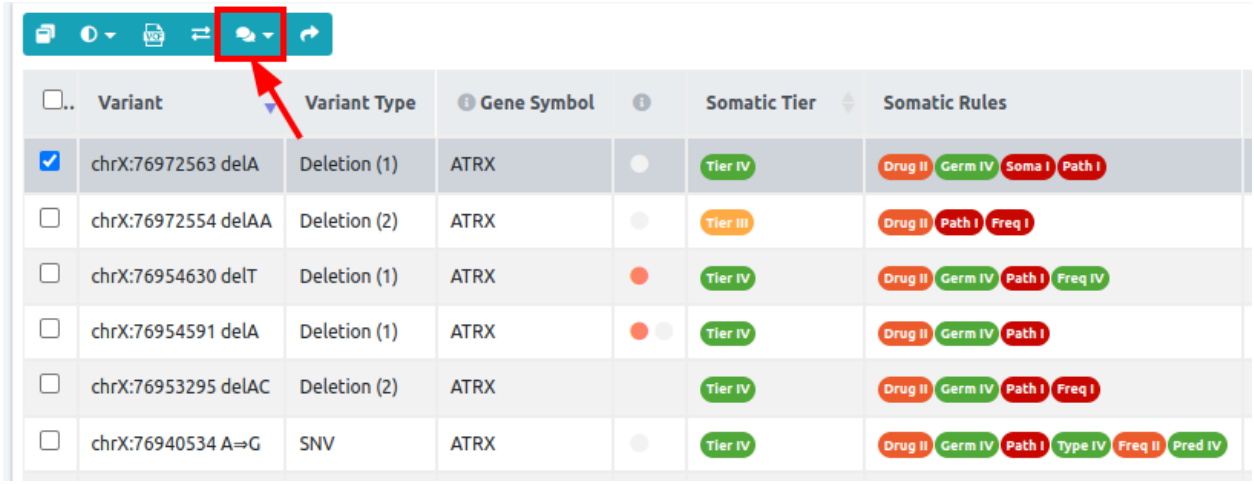
This can help you to identify potential artifacts by comparing the number of times a variant is observed within the samples of the group. The system will label the variants as “Automatically Tagged Likely Artefact (VarSome)” if:

- The group has more than 20 samples analyzed
- The variant has a population frequency lower than 1% (according to gnomAD)
- The variant is found in more than 10% of the samples.

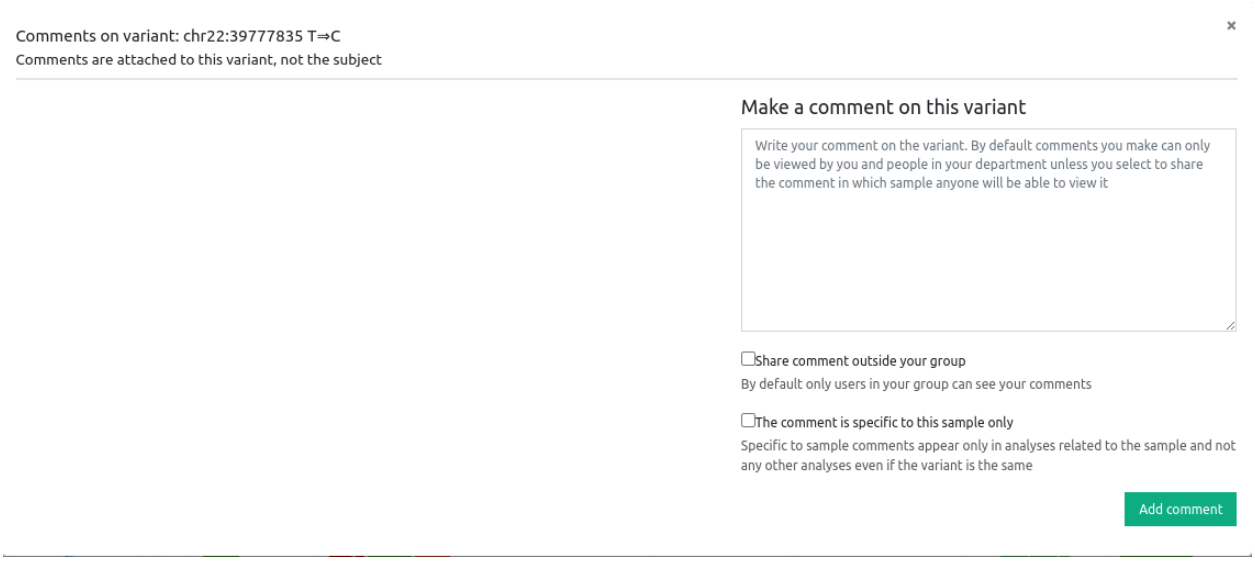


Comment on variants

You can attach comments to variants by clicking on the "Comments" icon:

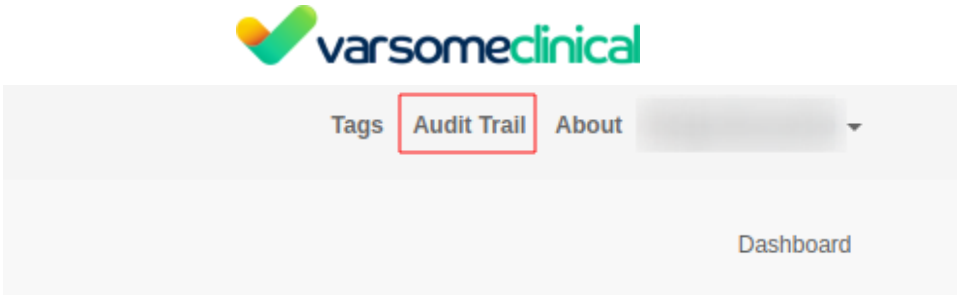


A pop-up window will be shown where you can add your comment. The comment will be attached to the variant and it will be shown in other samples of your group where the same variant is found. The comments are also private for your group by default.



- Share comment outside your group: if you click this option the comment will be shared with the VarSome community of users.
- The comment is specific to this sample only: if you choose this comment, the comment will be attached only to that sample and it won't be shown in other samples even if the same variant is found.

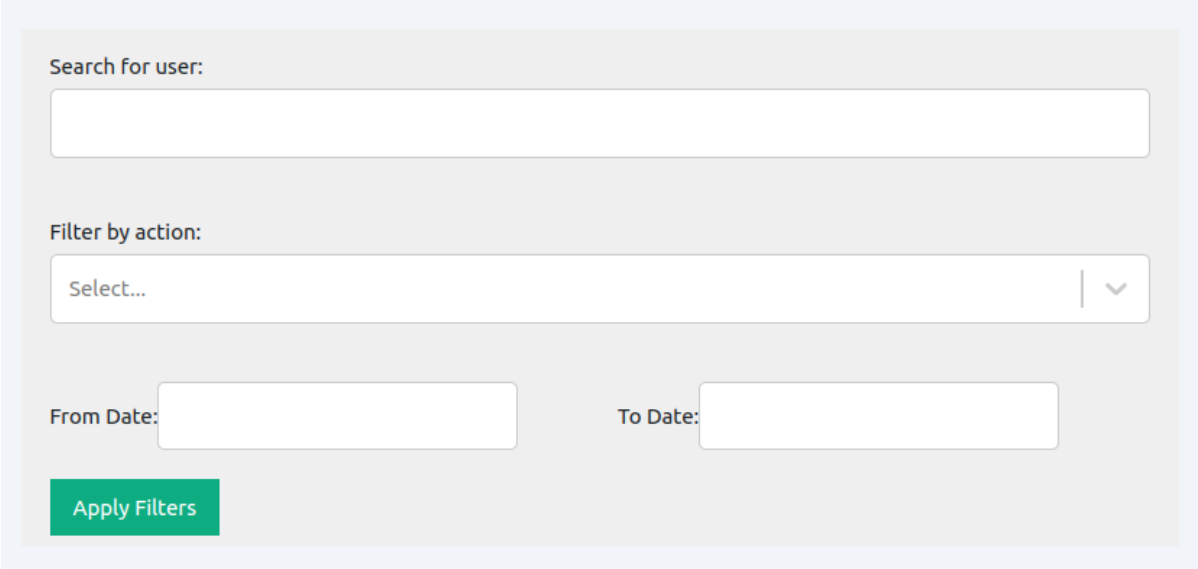
All custom classifications, along with other users' actions can be monitored by the group's supervisor via the audit trail tab.



If your group does not have a group supervisor or you would like to transfer this role to another user please contact our support team.

The Audit Trail shows the record of different actions that have been made from all users of a group on the samples analyzed. This log is only visible to the group supervisor.

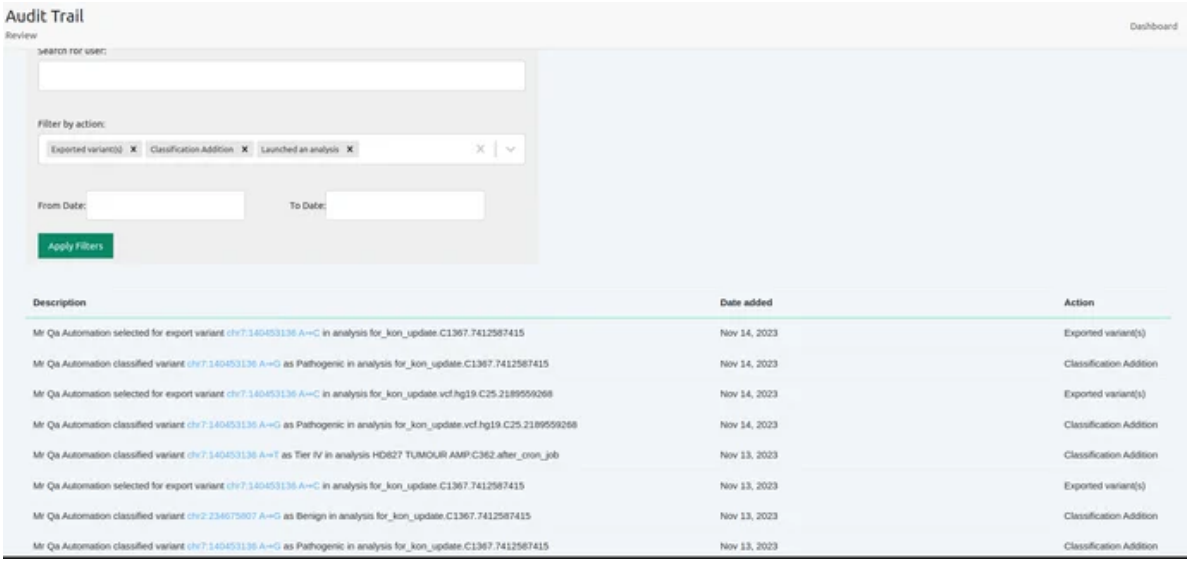
The group supervisor can retrieve specific actions recorded in the Audit Trail by using the filters which allow to filter by user, action type and date.



The Audit Trail shows the following action types performed by users of the same group:

- **Sign in / sign out:** who has signed in and out and the IP address.
- **View / access a URL:** who has accessed which analysis.
- **Accepted analysis terms:** users who have accepted the analysis terms before accessing the results. Note that the users need to accept the analysis terms the first time they access the results. This is not required for analyses for which they have already accepted the terms.
- **View analysis messages:** sometimes there are messages associated with the analyses. These messages are related to the quality of the input samples and are aimed to warn users of potential quality issues that may require the user's attention. If a user checks these messages, this action is recorded in the Audit Trail.
- **Analysis launch:** users who have launched an analysis.
- **Variant classification:** when users add, modify or remove manual variant classifications, this is also stored in the Audit Trail.
- **Analysis filtering:** users who have applied filters in analyses.

- **Variant export:** variants selected for export and which user has selected them.



Import list of custom classifications and variant comments

VarSome Clinical can also take your list of custom classifications (i.e. manually curated list of variants) and upload it privately to your account, along with variant comments.

If you have your own, custom pathogenicity classifications for certain variants or wish to add information in the form of a comment, you can send us a file with this information. We will then include this information and display it in any of your analyses where the relevant variants appear. However, the provided file needs to fulfill certain requirements in order to be processed correctly:

- The file provided by the user should be in CSV (comma-separated values) format with content in the following order: chr,position,ref,alt,classification,comment(s). The classification should be one of 'P' (pathogenic), 'LP' (likely pathogenic), 'VUS' (variant of unknown significance), 'LB' (likely benign) or 'B' (benign).
- Multiple comments should be separated by semicolon (;) and the size of each comment is limited to 2000 characters.
- One classification per variant is expected.
- Each CSV file corresponds to one specific user. Multiple files need to be provided for multiple users.
- It is essential that the corresponding reference genome (hg19 or hg38) is specified by the user along with the provided file(s).

This is an example of a valid CSV file:

```
chr5,125562,C,T,P,found in a young adult with melanoma;

chr6,34229972,T,G,B,Common artifact
```

Classification Nomenclature

Standard classifications must be declared with the following names to comply with the nomenclature used by VarSome Clinical. They can be referred to either by the full name or its code:

- B or Benign
- LB or Likely Benign
- VUS or Uncertain Significance
- LP or Likely Pathogenic
- P or Pathogenic
- CA or Common Artefact
- DR or Drug Response
- DA or Disease Association
- RF or Risk Factor
- PR or Protective

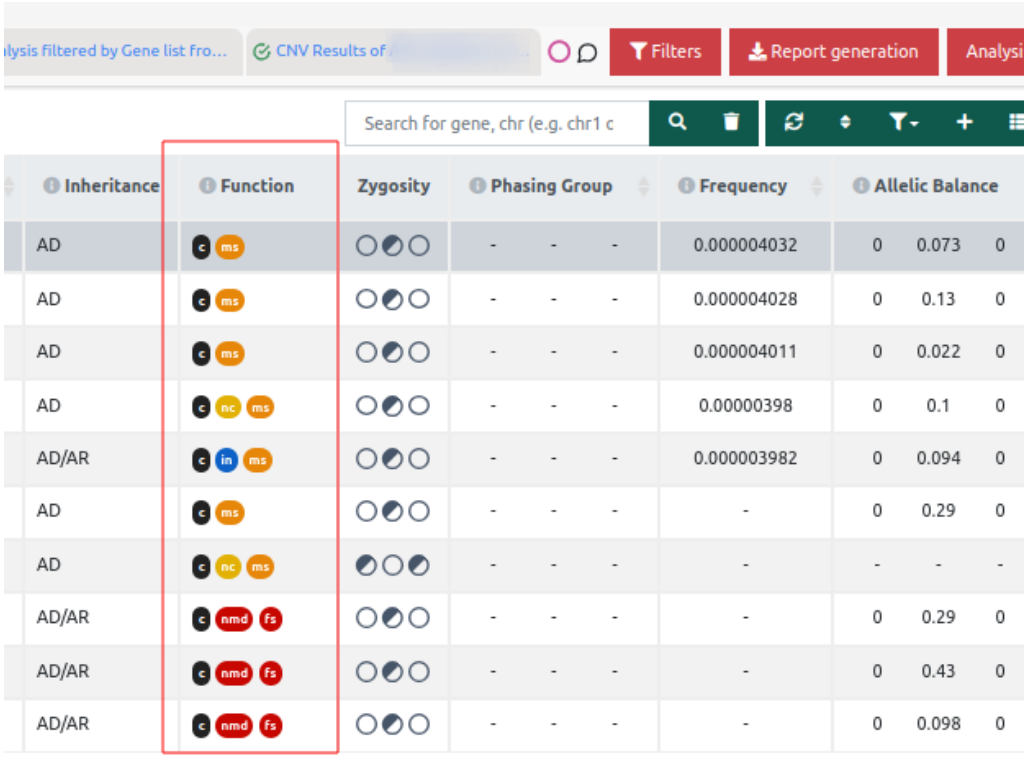
- **CNV:** this column is shown when the user runs a CNV analysis with the sample. It will have green ticks if any of the following conditions meet:
    - there is a CNV that overlaps with the position of the current variant
    - there is a CNV in the same gene where the current variant was found (not necessarily in the same position of the variant)

- **Class:** Variants are ordered by our pathogenicity classification**:**
  5 = Pathogenic,
  4 = Likely pathogenic,
  3 = Uncertain significance,
  2 = Likely benign,
  1 = Benign.
- **ACMG Rules:** The set of triggered ACMG rules are displayed in clickable bubble icons that include the rule's description and explanation for triggering.
- **HGVS:** HGVS nomenclature for the variant.
- **HGVS Protein:** HGVS nomenclature for the protein sequence change compared to the reference.
- **HGVS Coding:** HGVS nomenclature for the variant.
- **Transcript position:** Variant described on the DNA level in relation to a specific **gene** based on the coding DNA reference sequence.
- **Overlapping Genes:** The name of any gene(s) the variant falls within.
- **Inheritance**: Mode of inheritance of the gene from the CGD, OMIM, ClinGen Disease Validity, gene2Phenotype, GenCC and Domino databases:
    - AD: autosomal dominant
    - AR: autosomal recessive
    - XL: X-linked
    - BG: blood group
- **Function**: The position of the variant with respect to the gene it falls within, and its coding effect (if any).

5f	5' Flank	**nc**	non-coding exon
5u	5' UTR	**fs**	Frameshift
3f	3' Flank	**ms**	Missense
3u	3' UTR	**ns**	Nonsense
C	Coding	**sl**	Stop Loss
In	Intronic	**sy**	Synonymous
Sp	Splicing	**ed**	exon deletion
If	in frame	**nmd**	nonsense-mediated decay

VarSome Clinical allows the use of custom transcripts for annotation.

In terms of variant function, VarSome Clinical annotates variants against all the transcripts available (Ensembl and Refseq), and shows the results in the Function column when browsing results:



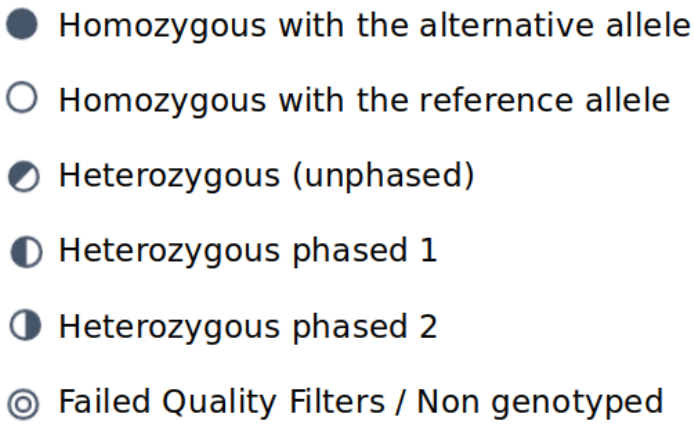
However, for the Germline Variant Classification, as a part of the annotation of variants, VarSome Clinical considers only one transcript - by default the transcript with:

1. the most severe coding impact,
2. otherwise the MANE Select transcript
3. if the above is not available, the longest canonical transcript
4. otherwise, the MANE Plus transcript

5. failing that, the longest transcript
6. and finally, the RefSeq transcript

For Ensembl transcripts, all must have TSL [1] = 1 or null.

- **Gene Symbol:** Gene used for annotation and classification of the variant for ACMG (& AMP for somatic samples).
- **Zygosity:**



If the variant did not pass the variant caller quality filter, , the zygosity is shown in the table as ◎ (failed quality/non- genotyped).

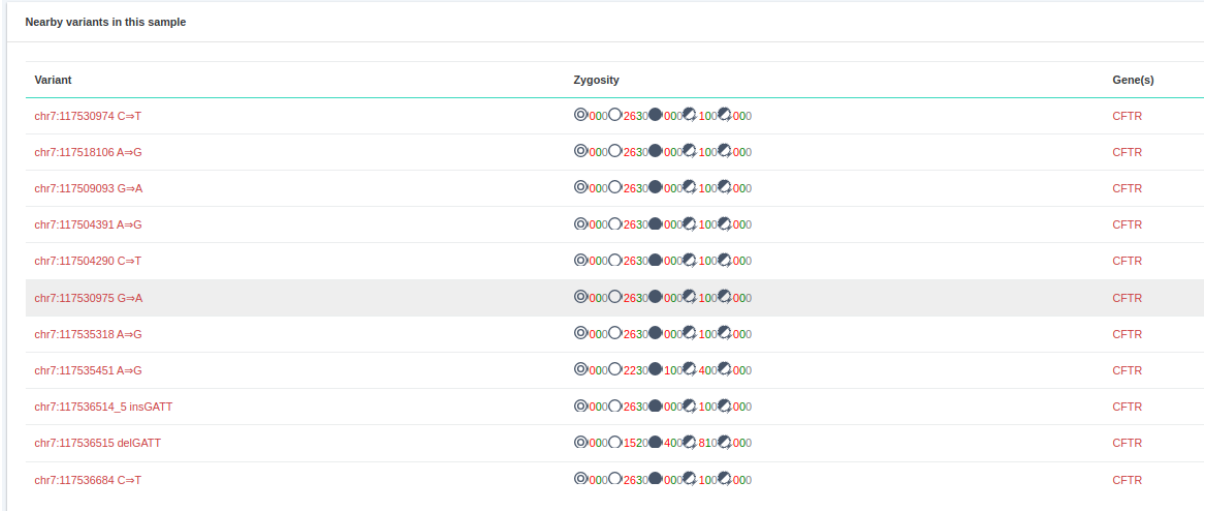
Representation of zygosity for large cohort analyzes

The zygosity representation shown below applies for multi-sample analyses larger than 10 samples and brings together information about both the zygosity and the status (affected - unaffected) of the samples at first glance.

In this layout, each type of zygosity found for a variant is shown only once with the respective box. Next to each box reside two numbers: the number in green font corresponds to the unaffected samples having this specific type of zygosity for this variant. The number in red font corresponds to the affected samples having this type of zygosity in the dataset.

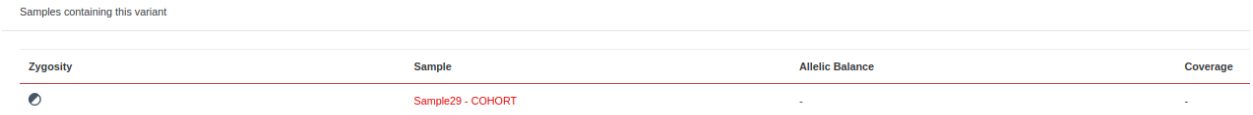
Each box is associated with this pair of color-coded numbers which provides information about both the number of occurrences in the dataset and the status of these samples in a compact manner.

The "Nearby variants in this sample" tab is modified, as well to fit more efficiently large cohort analyses of more than 10 samples using the aforementioned layout described for the "Zygosity" column.



Finally, the tab "Cohort Zygosity", on the variant table, reports exclusively the names of

the samples carrying the variant, also color-coded in red or green font to inform about the sample status and their respective type of zygosity. Note that this tab appears in all multi-sample analyses, even those with fewer than 10 samples.



- **Phasing group**: this column indicates the phasing group of the variant if it exists. Note that this column will be available only for those variants that are phased. If you would like to filter the variant table and visualize only the variants that share the shame phasing group with your variant of interest, you can click on the following icon  . If you would like to know more about phasing please go here.
- **Allelic balance:** Proportion of reads that support the variant. For example, if a variant's location is covered by 100 reads, of which 25 support the variant and 75 do not, then the variant would have an allelic balance of 25/100 = 0.25.
- **Frequency:** Frequency of the variant in the general population or (if applied) the specified ethnicity.
- **Coverage**: Number of reads that align to the variant's position. For analyses of FASTQ samples, the blue numbers are links to JBrowse, showing the read alignments at the variant's position.
- **Filters**: filters that have been applied to the data. Filters are associated with the variant calling quality filters that have been applied to the variant to decide whether it has a call status of PASS or FAIL.
- **rsID**: the reference SNP cluster ID of this variant

Extra columns for *somatic* samples:

- **AMP Tier:** Variants are ordered by an aggregate AMP score (see AMP Implementation documentation for more details), from most pathogenic to benign. Tier I = Cancer with approved drug therapies, Tier II = Cancer but no approved therapy, Tier III = Uncertain Significance, Tier IV = Benign or not related to cancer.
- **AMP Rules:** The set of triggered AMP rules are displayed in clickable bubble icons that include each rule's description and explanation for triggering
- **Sample Metrics:** Each icon represents the sample information introduced by the user. They light up when there is data in one of the cancer-related databases matching the variant to the relevant sample characteristic:
    - Cancer type: this highlights any variants for which evidence is found linking to the same cancer type as the sample.
    - Tissue: This will highlight any evidence associating the variant or gene to the sample tissue.
    - Age: This will display the patient's age relative to an age histogram

in reported somatic samples.

- Ethnicity: It will report the variant's frequency in the relevant ethnic group.
- Sex: more than 50% of reported cases across somatic sample databases match the sample's sex**.**
- Variant Allele Frequency: variants with a low VAF are most likely tumor variants, whilst VAFs of 50% and 100% indicate germline variants.

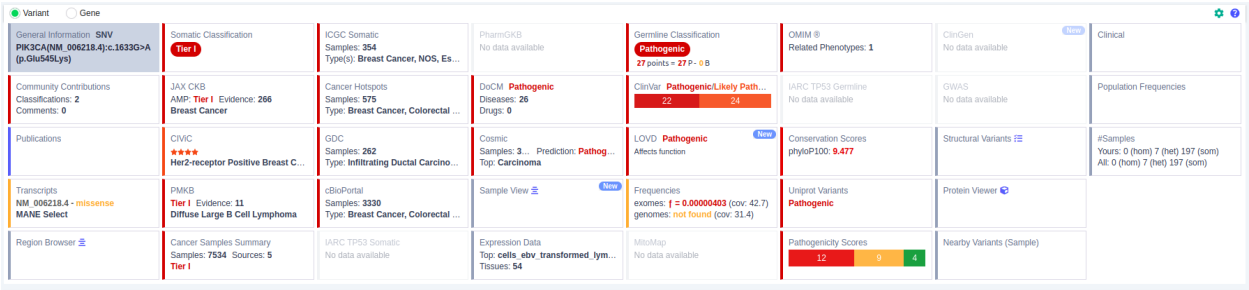
- **Somatic Samples:** Sum of available affected samples from databases included in Cancer Sample Summary (ICGC Somatic, COSMIC, CBioPortal, Cancer HotSpots, GDC).

Clinical Cards

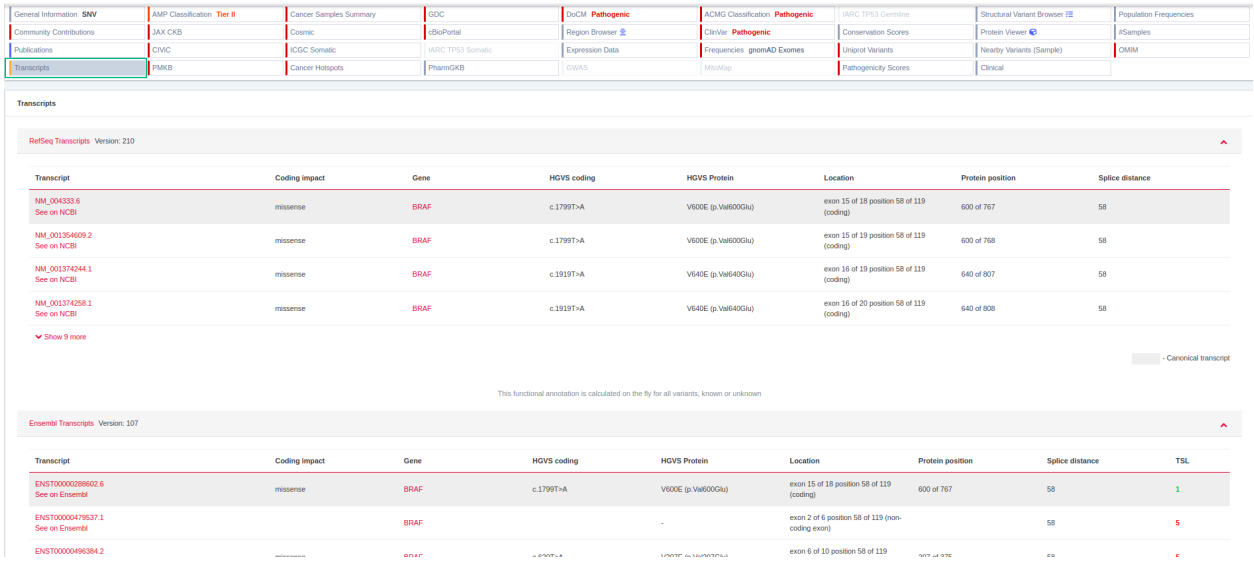
On the bottom of the variant table the clinical cards display complementary information of the *variant* or the *gene*.

⚠️ Please note that, when selecting a variant through the check box on the Variant table in VarSome Clinical, the Clinical Cards do not change because the selection of the variant is different from clicking on it to view the variant information. Therefore, to view the clinical cards and the information on a variant, simply click anywhere on the row containing the variant and not the check box to select it for any action.

Variant clinical cards:



- **General Information:** Information about the genomic location of the variant, its type, cytoband, HGVS notation, sequence.
- **Community Contributions:** VarSome's community public contributions for this variant.
- **Publications:** publications from PubMed related to the selected variant or gene where genes, variants, diseases, phenotypes, chemical compounds, drugs (if exist) are tagged by our internal AI tool.
- **Transcripts:** Chromosomal location, link to UCSC genome browser, dbSNP (rs number), Refseq transcripts containing the variant, HGVS notation, etc. Available transcripts for the selected genes are displayed (information from RefSeq and Ensembl, along with the MANE transcript). The transcript used in classification is highlighted in the Transcripts summary card.



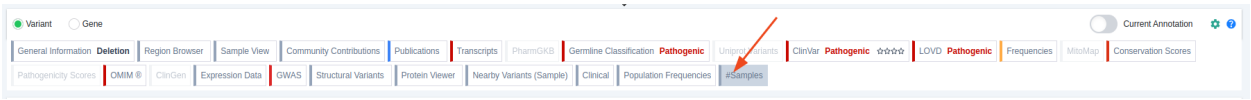
- **Region Browser:** Genomic region browser, lollipop graph of the pathogenicity of each variant, frequencies from gnomAD and Bravo, variant visualization with filtering according to databases and coding impact.
- **Sample View**: Sample's region browser presents information about SNVs, transcripts, ROH, conservation scores/position (and CNVs) of the sample. For further information please refer to section Sample View.
- **PharmGKB:** Information on the impact of genetic variation on drug response from PharmGKB database.
- **Expression Data:** Tissue-specific variant expression data from the Genotype-Tissue Expression (GTEx) project.
- **GWAS:** Associations of specific genetic variations with particular diseases from the genome-wide association study (GWAS) database.
- **ACMG Classification:** The Germline Variant Classification and its triggered rules for the selected variant.
- **ClinVar:** Information about clinical associated phenotypes connected to the selected variant. Description/Submitter column displays ClinVar Comments (View ClinVar Comments).
- **Frequencies:** If known, Gnomad frequencies for the selected variant and for any other known variants that overlap with it.
- **MitoMap:** Information from the Human Mitochondrial Genome Database.
- **Conservation Scores:** Conservation scores from different resources.
- **UniProt Variants:** Variant information from UniProt.
- **In-Silico Predictors: :** Variant pathogenicity predictions produced by *in silico* predictors. In this clinical card, the pathogenicity is displayed with a different color and there are two bar charts, one for the meta-score (predictor which combines multiple predictors into a single score) and one for the individual predictors.
- **ClinGen:** Curated data from ClinGen variants database. Information will be available when the queried variant has already been curated by any of the ClinGen expert panels.
- **Structural Variants:** Structural variant browser for the detected variants.
- **Protein Viewer:** The 3D protein viewer tool is available to map variants onto the protein

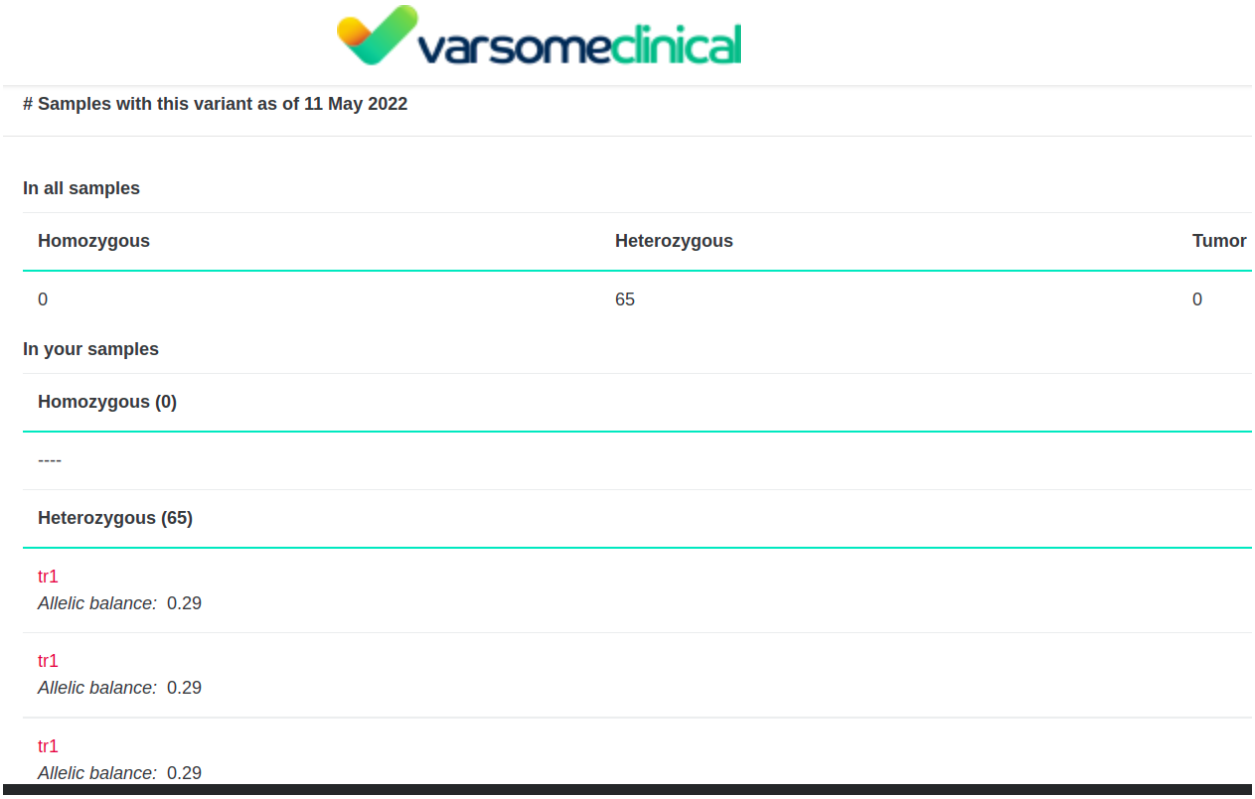
structure. By clicking on the “Protein Viewer” card, a new window will open with the 3D Protein Viewer tool. For more information please refer to 3D Protein Viewer.

- **Nearby Variants:** Variants in the genomic neighborhood of the selected variant. This variant list is not affected by the filters applied to the sample.
- **Clinical:** ClinVar and Cosmic annotations for the selected variant and for any other known variants that overlap with it.
- **Population Frequencies:** gnomAD frequencies and coverage, from Gnomad Exomes and Gnomad Genomes.
- **#Samples:** The number of samples in which a specific variant has been found. This column gets updated daily. The number of homozygotes and heterozygotes in Saphetor for the variants are shown, but only sample IDs of samples analyzed by you and your group are reported.

  VarSome Clinical comes with powerful sample cross-referencing, which links all your samples together on the variant level. It can be found in each #Samples tab and it reports two things:
  1. the number of samples of all VarSome Clinical users in which a specific variant has been found and
  2. in which other samples of you or your group this variant is present.

  This information is updated daily. The number of homozygotes and heterozygotes in VarSome Clinical for the variants are shown, but only sample IDs of samples analyzed by you and your group are reported.





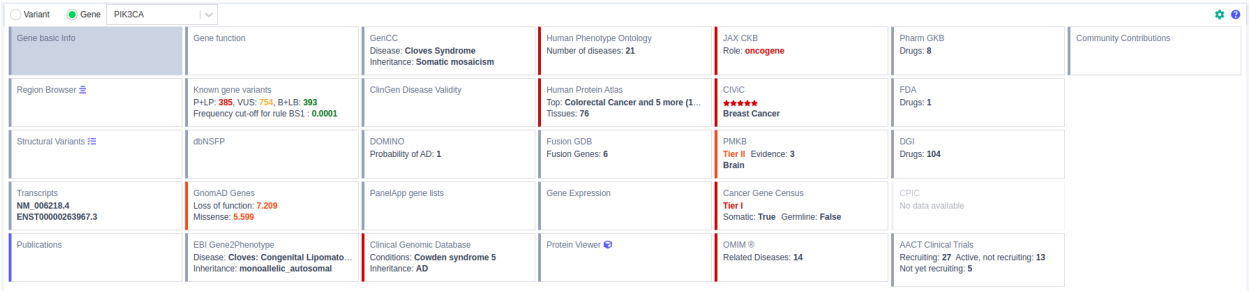
- **OMIM:** Information about phenotypes related to the selected variant, as retrieved from Online Mendelian Inheritance in Man®.
- **LOVD**: Presents the clinical evidence available from the Leiden Open Variation Database (LOVD) for a given gene and/or variant and it has also been incorporated into the Germline Classification as a new source of clinical evidence.
- **DVD:** Variant annotations from the Deafness Variation Database. DVD provides a comprehensive guide to genetic variation in genes known to be associated with deafness.
- **Multi:** Displays the components and their corresponding files used to run the multi-sample analysis, and if these have been selected as affected or not.
- **Audit trail:** Shows the record of the actions that have been made from all users of a group on the samples analyzed. Only the group supervisor has access to this information.

Additional cards for somatic samples:

- **AMP Classification**: The AMP tier and the set of triggered rules for the selected variant.
- **JAX CKB:** Somatic gene variant annotations and related content provided by The Jackson Laboratory Clinical Knowledgebase.
- **CiViC:** Somatic variant annotations retrieved from CiViC.
- **PMKB:** Clinical interpretations of somatic variants retrieved from PMKB.
- **Cancer Samples Summary:** Aggregated information across different data sources.
- **Cosmic:** Somatic variant annotations from COSMIC database.
- **ICGC Somatic:** Somatic variant annotations from ICGC database.
- **Cancer Hotspots:** Somatic annotations from Cancer Hotspots database.
- **GDC:** Somatic variant annotations from GDC database.

- **cBioPortal:** Summary of samples matched in cBioPortal with the selected somatic variant.
- **IARC TP53 Somatic & Germline:** Somatic annotations of TP53 gene mutations in human cancers.
- **DoCM:** Information retrieved from DoCM, about known, disease-causing mutations associated with the variant.

Gene clinical cards:



- **Gene basic info:** Description, synonyms, cytoband, links to clinical or other resources.
- **Transcripts:** Strand, chromosomal location, length, mRNA length, UniProt accession number, etc.
- **Publications:** Publications from PubMed related to the selected gene where genes, variants, diseases, phenotypes, chemical compounds, drugs (if exist) are tagged by our internal AI tool.
- **Gene function:** Functions related to the selected gene, as provided by Genetics Home Reference (GHR).
- **Known gene variants:** Variants in the selected gene with known pathogenicity.
- **dbSNFP:** Functional prediction and annotation of all potential non-synonymous single-nucleotide variants.
- **GnomAD Genes:** Data summary from a wide variety of large-scale sequencing projects associated with the selected gene.
- **EBI Gene2Phenotype:** Gene association with a disease entity based on an allelic requirement and a mutational consequence.
- **GenCC:** Curated information about the gene-disease relationship.
- **NHI ClinGen Disease Validity:** Gene-disease association validity information.
- **DOMINO:** Probability of the selected gene to cause dominant changes.
- **PanelApp gene lists:** Catalog of available gene panels including the selected gene.
- **Clinical Genomic Database:** Age affected, condition, inheritance, indicated intervention categories, publications as retrieved from Clinical Genomic Database.
- **Human Phenotype Ontology:** Disease and their phenotypic abnormalities associated with the selected gene.
- **Human Protein Atlas:** Protein expression information by cell and tissue type
- **Fusion GDB:** Functional annotation of fusion genes in cancer and their related drugs
- **Gene Expression:** Tissue-specific gene expression data from the Genotype-Tissue Expression (GTEx) project.

- **Protein Viewer:** The 3D protein viewer tool is available to map variants onto the selected protein structure. By clicking on the "Protein Viewer" card, a new window will open with the 3D Protein Viewer tool. For more information please refer to the section 3D Protein Viewer.
- **JAX CKB:** Evidence and clinical trials content related to the selected gene as provided by the section 3D Protein Viewer Jackson Laboratory Clinical Knowledgebase.
- **OMIM:** Information about phenotypes related to the selected gene, as retrieved from the Online Mendelian Inheritance in Man®.
- **PharmKGB:** Information on the impact of genetic variation on drug response from PharmKGB database.
- **FDA:** Approved drugs associated with the selected gene, from FDA.
- **DGI:** Information about drug-gene interactions interpreted by the "Drug Gene Interaction" Database.
- **CPIC:** CPIC levels to genes/drugs retrieved from "Gene Drugs Interactions and Levels"
- **AACT Clinical Trials:** Information about every clinical study registered in ClinicalTrials.gov associated with the selected gene as provided by AACT.
- **Community Contributions:** VarSome's community public contributions for this variant.

Additional cards for somatic samples:

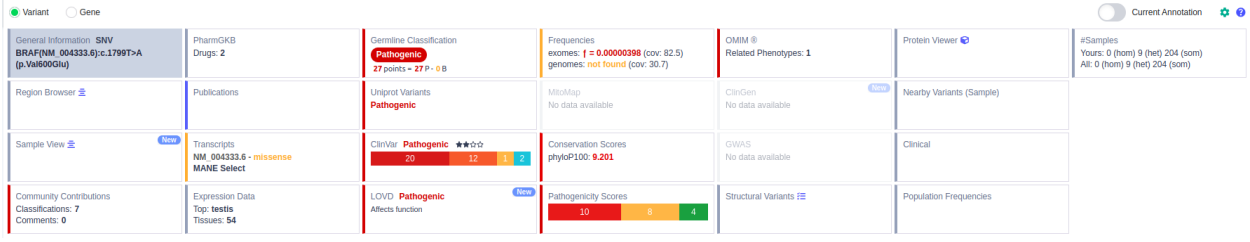
- **CiViC:** Cancer-related clinical evidence as derived from CiViC database.
- **PMKB:** Clinical interpretations of gene variants retrieved from PMKB.
- **Cancer Gene Census:** Information about gene's mutations that are causally implicated in cancer as retrieved by COSMIC database.

⚠️ Please note that grayed out tabs are disabled if no related information is available.

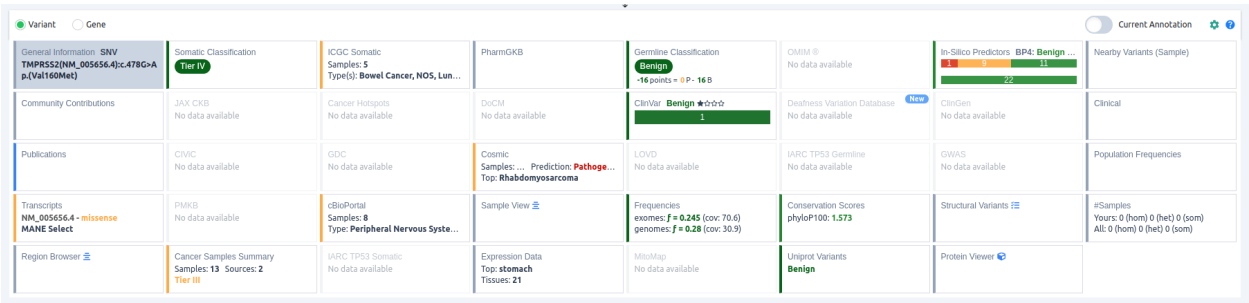
Current Annotation:

VarSome provides "Current Annotation" of **germline and somatic** variants using the latest data and the latest germline or somatic classifier. This feature allows you to see whether additional data are available for a variant, and whether the automated classification is altered by new evidence or refinements to the classifier. *This feature is provided for research purposes only in VarSome Clinical*.

Original view:
for germline variants:

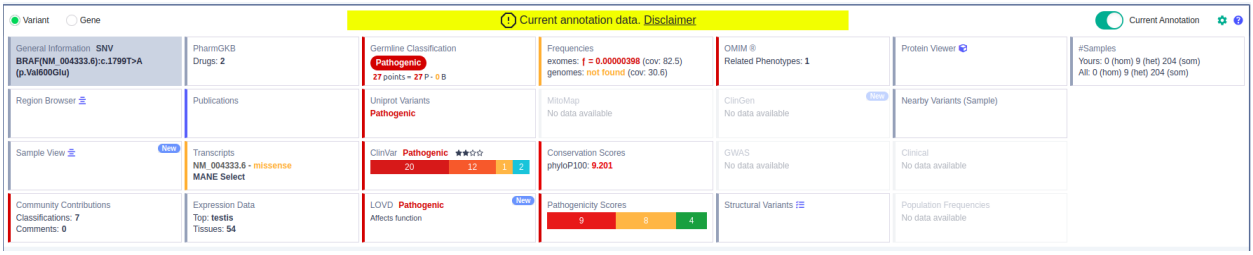


for somatic variants:

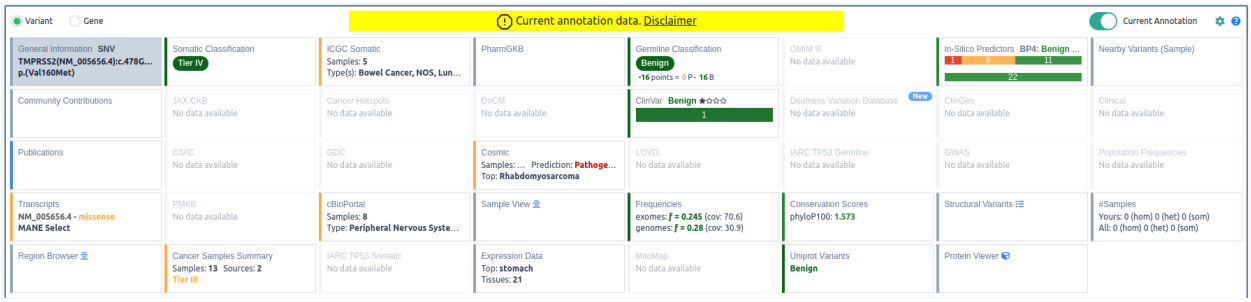


Current annotation view:

for germline variants:



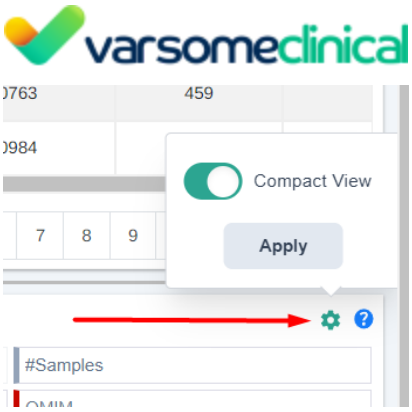
for somatic variants:

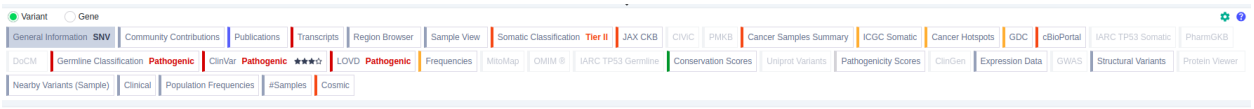


⚠️ Please note that the **current annotation view** is now available for germline and somatic analyses. The variant link to VarSome has been removed from these types of analyses.

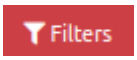
Compact View:

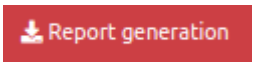
It is possible to make the clinical cards much smaller by clicking on the “Display Options” wheel on the top right hand side of the cards panel, and selecting “Compact View”. This removes the summary data from the cards and reduces their size.



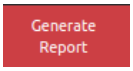


Some useful icons:

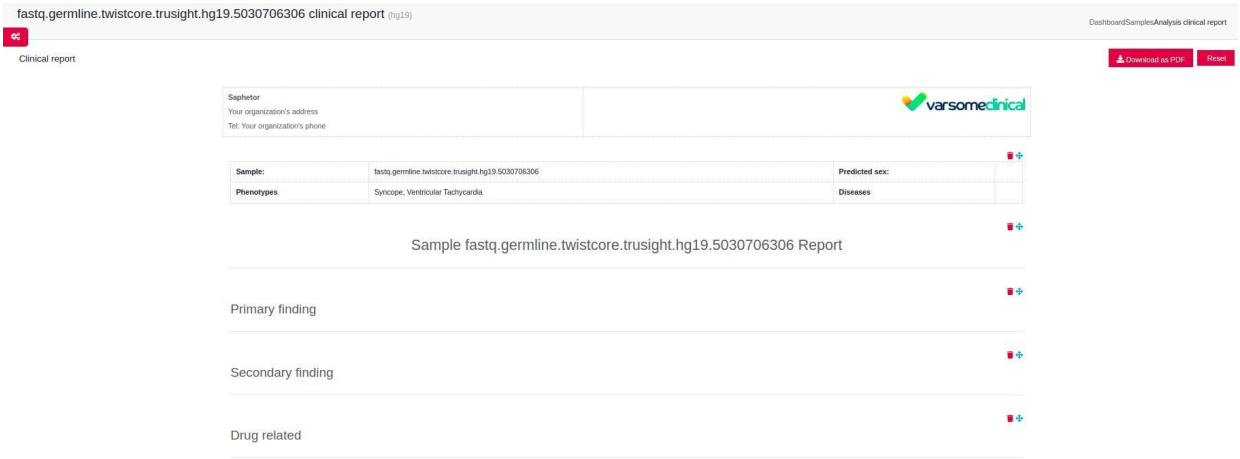
- Click on this box 

to get access to your saved filter sets
- Click on the **Report generation** icon 

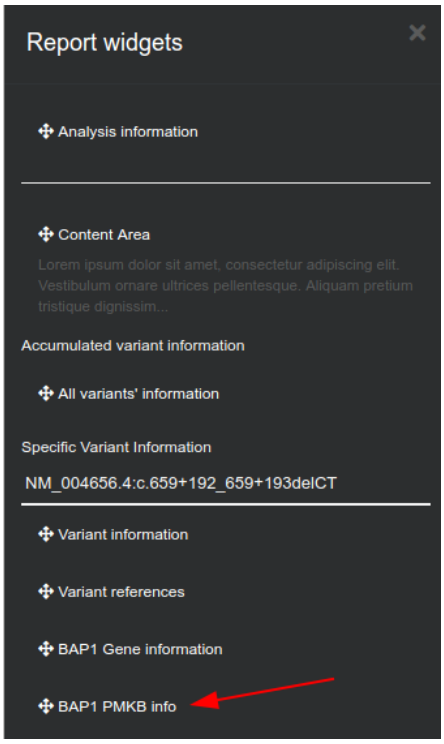
to see the list of **variants selected for export**.

  You can download a **Report** of the selected variants in PDF format. To do so, click on the above icon**,** then select 

and you will be directed to the following screen.



By clicking on the  icon, a Report widgets menu will be shown in order to customize the report.

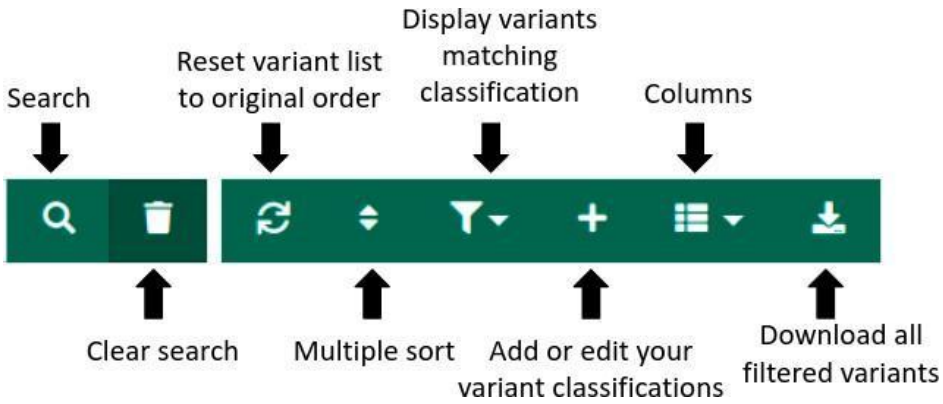


You can drag and drop the information you prefer to include in the report. At the last option you can find PMKB information for the gene that includes the variant. See more in section 15 Final Report Generation.

- 

The options provided here are the same as described in "Analysis actions" options.

Green menu



-  **Search:** You can search through your results, by querying according to the VarSome search format. The query can include any of the following:
  - gene: e.g. PIK3CA,
  - chromosome: e.g. chr3 or 3
  - chromosome position: e.g. chr3:178947865, chr3-178947865, chr3 178947865 or 3

178947865.
    - genomic range: e.g. chr3:178936091:178942431, chr3:178936091..178942431, chr3-178936091-178942431, chr3 178936091 17894243 or 3 178936091178942431.
    - variant (DNA): e.g. chr3:178936091 G⇒A, chr3:178936091-G-A, chr3-178936091-G-A, chr3 178936091 G A, 3:178936091 G⇒A, 3:178936091-G-A…
    - variant (HGVS): e.g. NM_004448.4:c.1947-3C>A
    - variant (protein): e.g. BRAF:V600E or BRAF V600E.
    - rsIDs ("rs" followed by a number)
    - COSMIC IDs

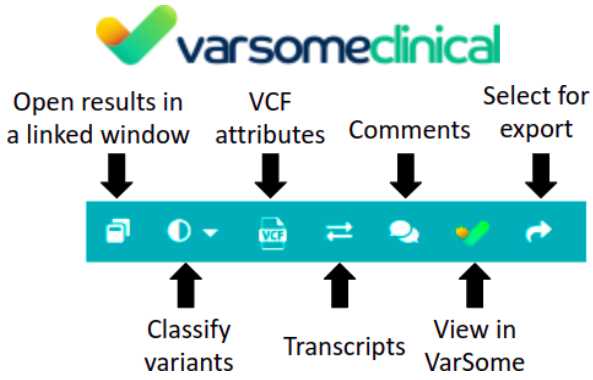
This will filter the table and show only the results for that query.

-  **Clear search:** This will empty the search box and show all variants again.
-  **Reset variant list to original order:** Clicking on this icon resets the sorting order of the columns to the default (the variants will be ordered by **Class**).
-  **Multiple sorting:** The list of variants can be sorted by multiple columns. A pop-up window will appear and multiple columns of interest can be selected in order to sort the variants in ascending or descending order.

  ⚠️ Please note that multiple column sorting will return informative results as long as the **first column,** which is selected to sort the variants, has **numeric** values (Frequency, number of samples, Phenotypes etc). For example, the user should *not* sort first by “AMP Tier” or “ACMG Class” and then sort by other values like allelic balance, frequency, etc. However “ACMG Class” and “AMP Tier” can be used as second or later in order of columns to sort by.

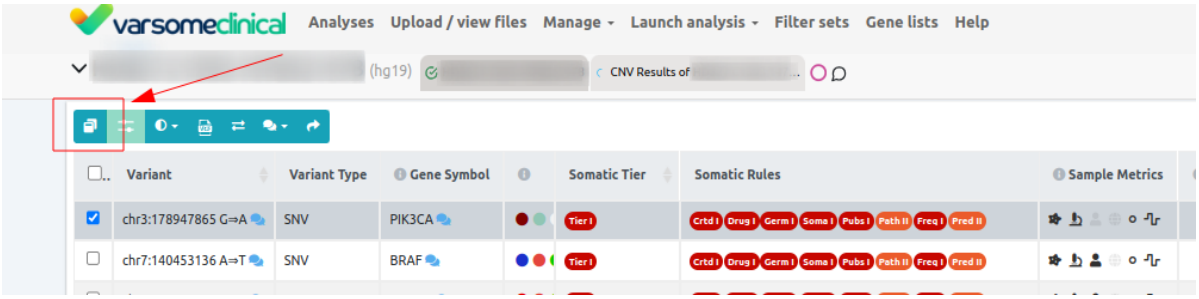
-  **Display variants matching classification:** Filters for custom variant classifications.
-  **Add or edit your variant classifications**: Open the Custom Tag creation menu. Custom tags allow you to classify variants using user-defined tags.
-  **Columns**: Remove or add columns to the table. This functionality can be used to remove columns that are not relevant for the analysis.
-  **Download all filtered variants from the table below (max 50000) in Excel format**: Download the list of variants (max. 50000) that pass any currently applied filters in Excel format. The Excel file also contains information about the filters used to obtain the exported table.

Blue menu

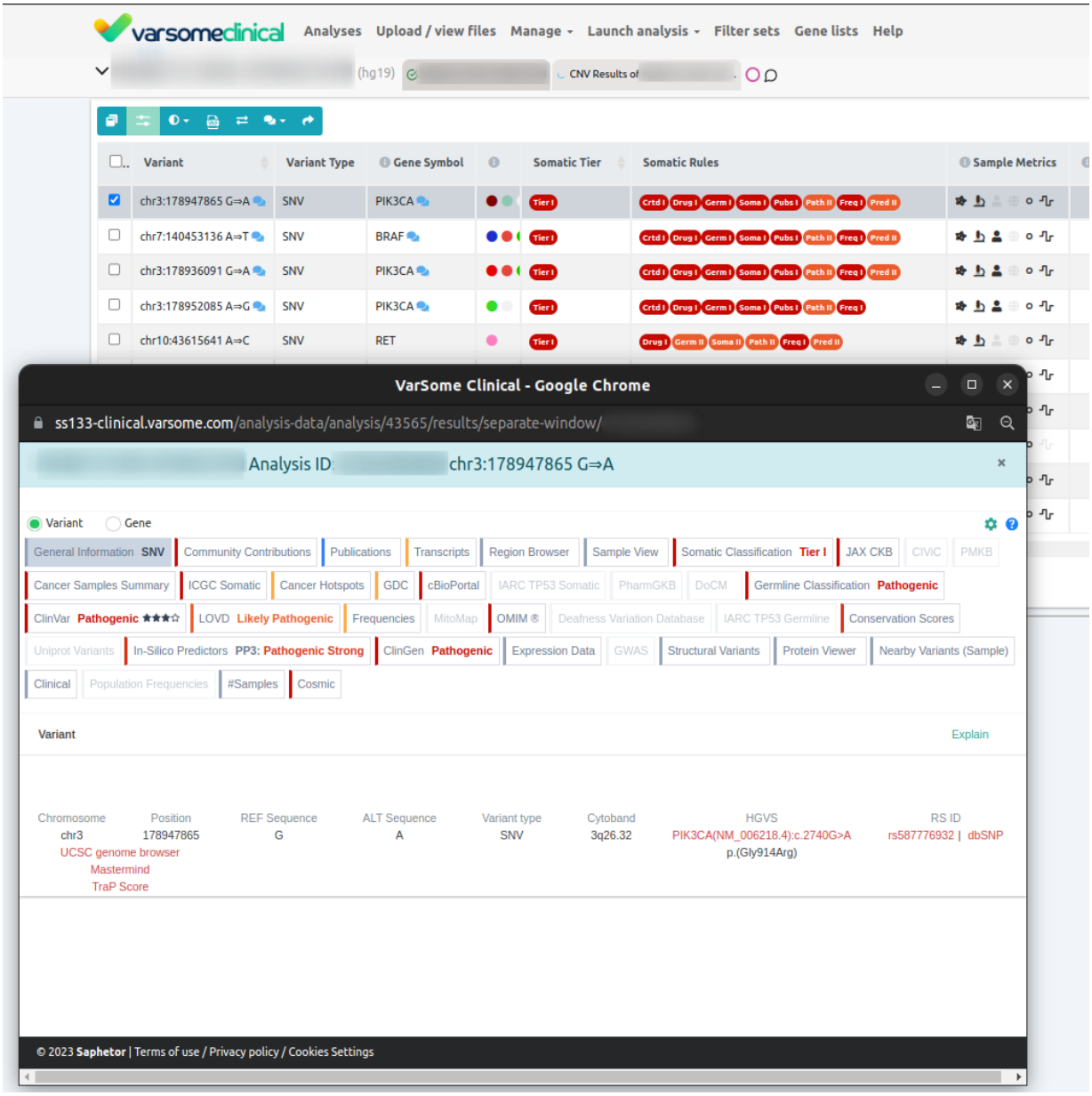


- **Open results in a linked window:** This functionality allows you to utilize multiple screens by generating "linked" sub windows that contain the results of an analysis.

  Clicking on the icon will split out the variant detail data into a separate window which you can reposition on a second monitor.

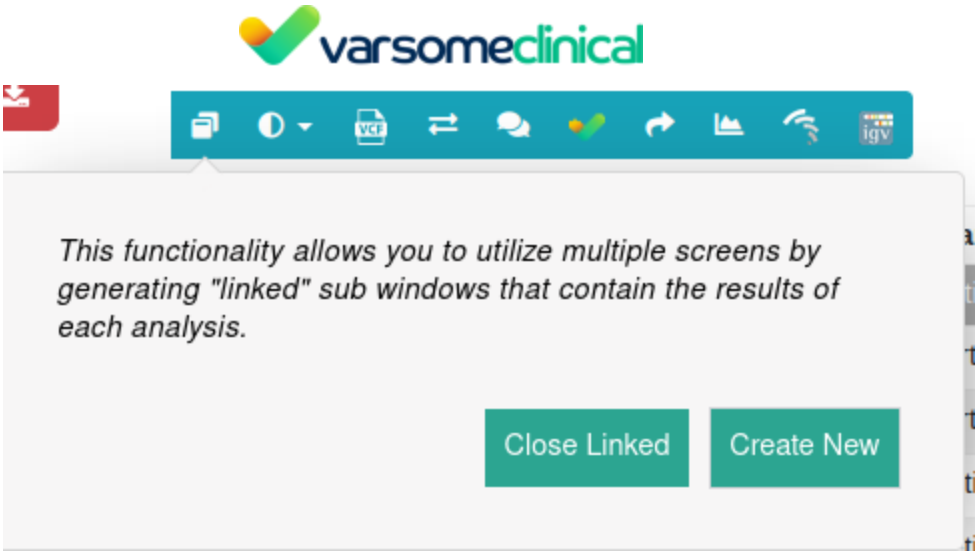


  A pop-up box will inform you of the functionality and by clicking on the "Open New" button, you will see a new window and the result data will be split between the two windows:

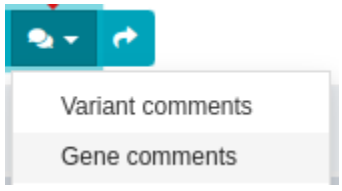


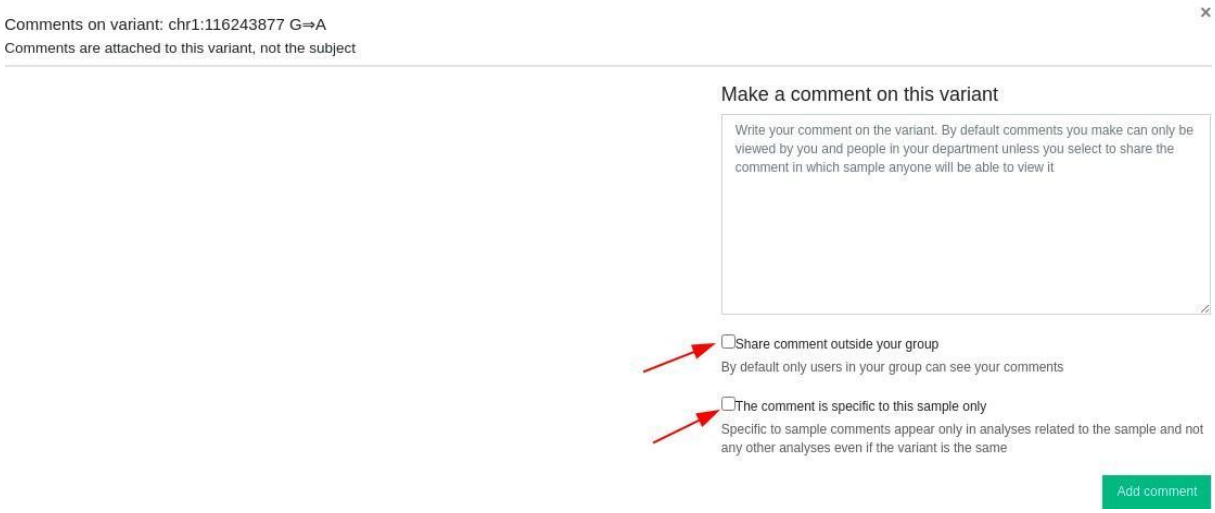
Note: Only one new window can be open simultaneously with the main analysis tab. If you are inspecting more than one analysis in different tabs of your browser, the "dual-monitor" window will sync with the tab of the analysis that is active each time.

You can close an existing pop-up window either by hitting the "Close Linked" button in the main screen, or by just hitting "X" on the window. Also, you can click on the "Create New" button to refresh the existing window with the synced data from the variant/ analysis you are inspecting.

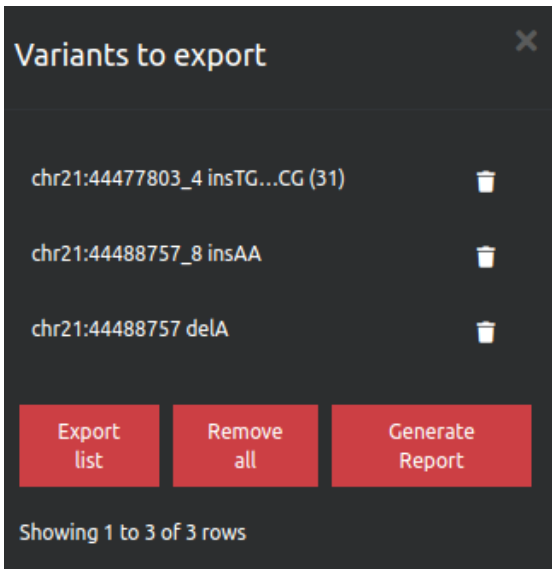


- **Classify variants:** add your own classification to a variant.
- **VCF attributes**: pop-up window describing the quality details for each software tool used to identify the variants. For more information see the section Explanation of VCF attributes.
- **Transcripts**: pop-up window with all the RefSeq transcripts containing the variant. It also shows the **location** of the variant (intron/exon, amino acid position), its **HGVS notation,** and genomic **function** (intronic, exonic, splicing, UTR ...). Canonical transcripts are shown in red.
- **Comments**: It is possible to attach a **short** comment to a selected variant (long comments will not be added and will return an error message). These comments will be linked to the variant or the gene and will be displayed in other analyses if the same variant is found. Variants with comments will have an icon in the **Variant or Gene** column. Comments are shared only within your group, unless you decide to make your comments public by selecting the "Share comment outside your group" option. You can also select the "The comment is specific to this sample only" option and the comment will be available only to this specific sample analysis. If however the variant is present in other analyses, the sample specific comment will not be shown.

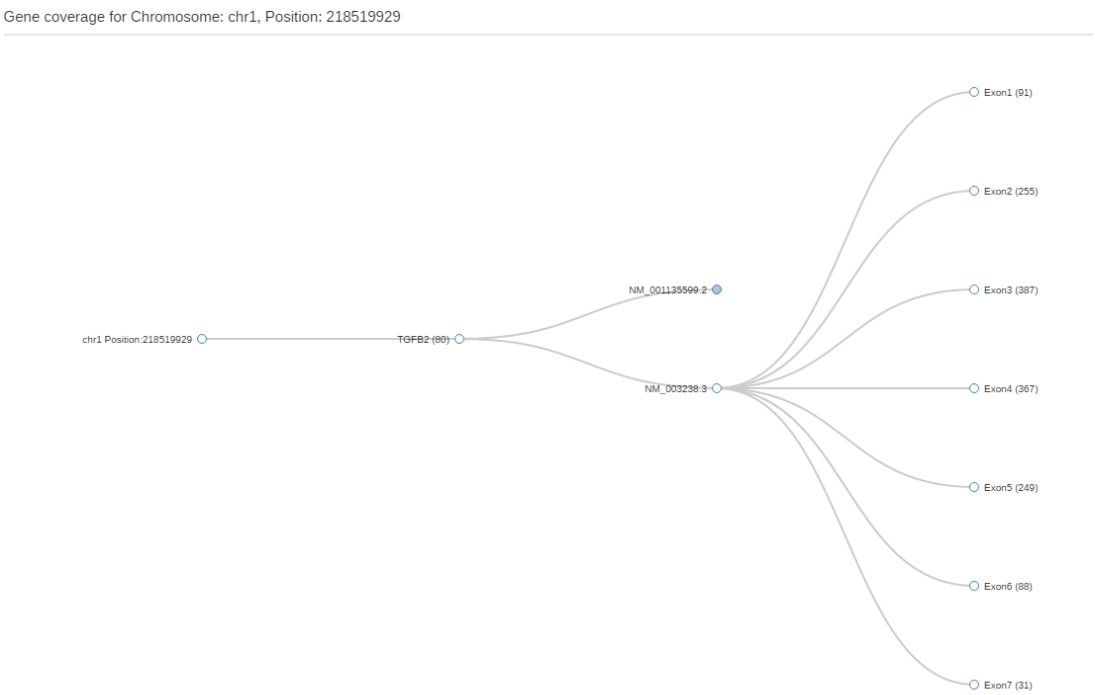




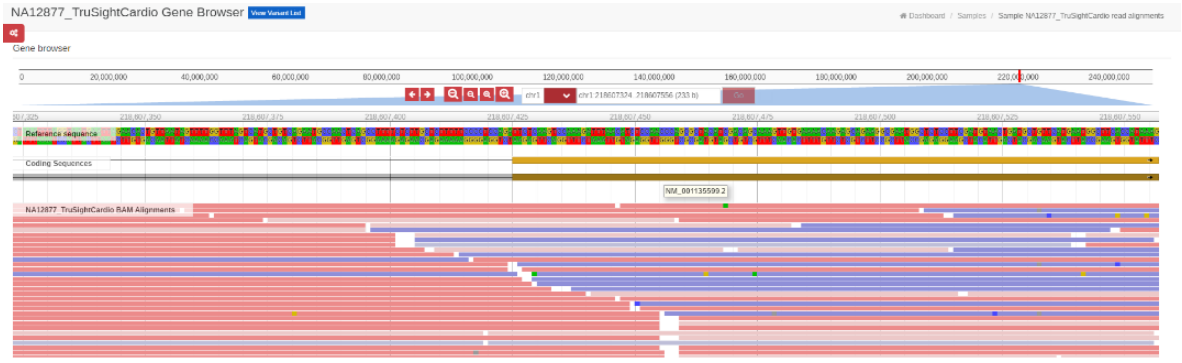
- **View in VarSome**: link to our free knowledge base and database aggregator, VarSome.
- **Select for export**: Clicking on this box selects the variant for export, and information about the variant can be exported in Word and Excel format ("Export variant list" box)



- **Gene coverage**: a pop-up window showing the average coverage for the selected gene and its different transcripts. Clicking on the nodes will expand or collapse the tree.



Also, by clicking on one of the Exons, a new tab will open with a JBrowse (jbrowse.org) window showing the alignment details from the analysis' bam files. JBrowse is a software tool installed on our secured servers.

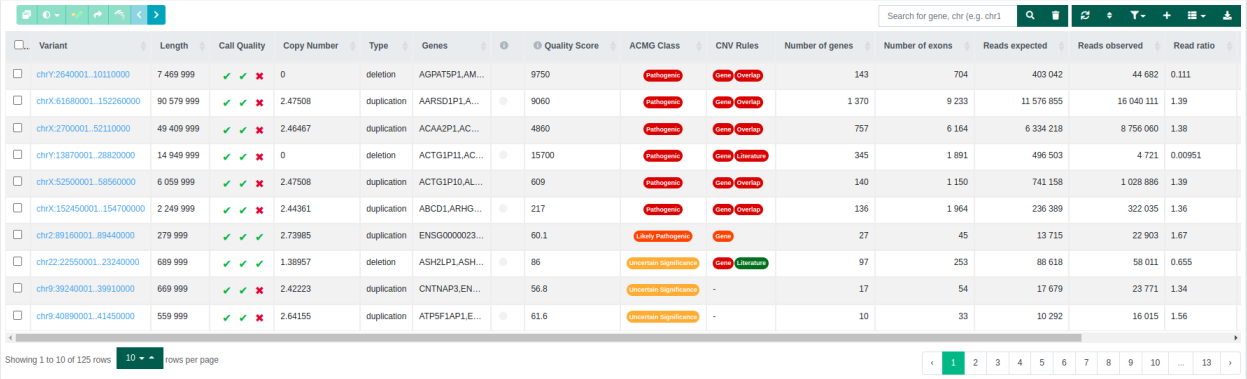


- **Read Alignment on JBrowse**: Opens a new tab with a JBrowse representation of the BAM files
- **Read Alignment on IGV**: Opens a new tab with an IGV representation of the BAM files

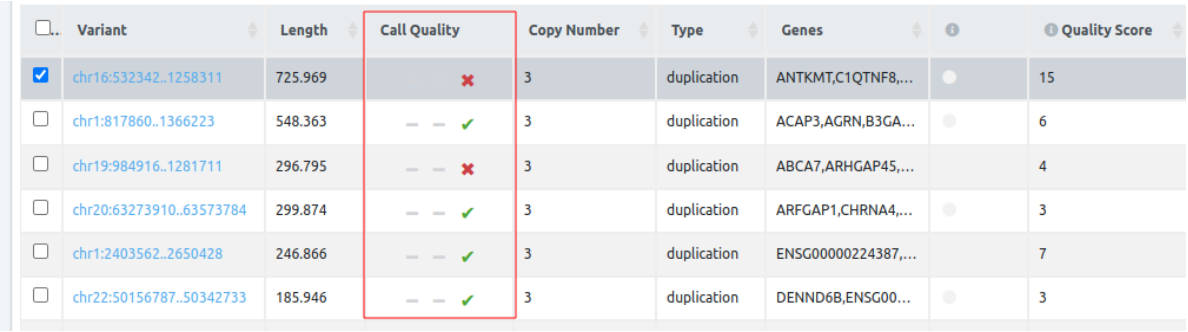
7.2 Description of CNV analyses' functionalities

CNV variant table

CNV/SV analysis variant table contains the following information:



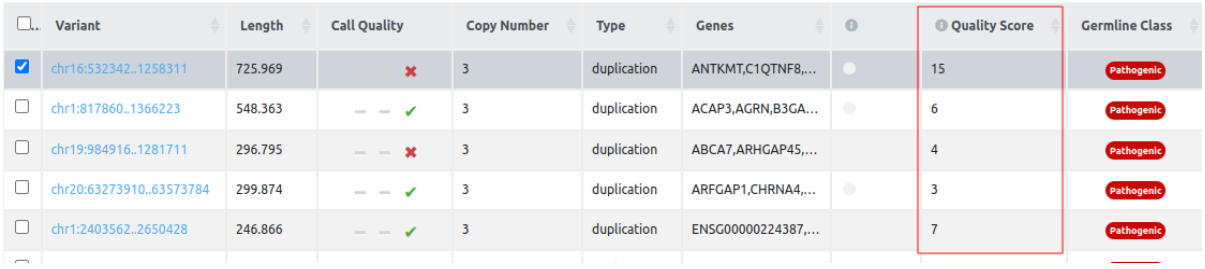
- **Length**: the length in bp of the region considered as a structural variation.
- **Call Quality:** Three quality control metrics collected for CNV analyses starting from FASTQ data (e.g. WES or gene panel). Each CNV call is assigned ticks, green, and "X"s, red, so at a glance you can see which has passed and failed the quality checks. The first and second metrics will have a grey dash for CNV results of analyses starting either from VCF or WGS data. From left to right, these are:
  - **Test sample coverage:** this quality control metric ensures a minimum coverage of the test sample at the CNV call region when calling duplications. Green: duplications with a coverage at least equal or higher than the minimum coverage threshold[1] (⚠️Please note that all deletions pass this filter and therefore will always have a green color). Red: duplications with a coverage lower than the minimum coverage threshold.
  - **Number of reference samples:** this is to ensure that a minimum number of samples from the reference set have a minimum coverage[1] in the CNV call region. Green: the reference sample set has at least two samples with coverage higher than the minimum coverage threshold, in the CNV region. Red: fewer than two reference samples with sufficient coverage in the CNV call region. [1]
  - **CNV call overlapping camouflaged region:** this is to check whether an overlap exists between the region of the CNV and the Camouflaged Regions. Camouflaged regions contain duplicated genomic sequences where confidently aligning short reads to a unique location is not possible. Green: no overlap. Red: overlap with a Camouflaged Region (Ebbert et al., 2019)



Variant	Length	Call Quality	Copy Number	Type	Genes		Quality Score
chr16:532342..1258311	725.969	✖	3	duplication	ANTKMT,C1QTNF8,...		15
chr1:817860..1366223	548.363	✔	3	duplication	ACAP3,AGRN,B3GA...		6
chr19:984916..1281711	296.795	✖	3	duplication	ABCA7,ARHGAP45,...		4
chr20:63273910..63573784	299.874	✔	3	duplication	ARFGAP1,CHRNA4,...		3
chr1:2403562..2650428	246.866	✔	3	duplication	ENSG00000224387,...		7
chr22:50156787..50342733	185.946	✔	3	duplication	DENND6B,ENSG00...		3

[1] Minimum coverage threshold (number of reads): the lowest value between 10 or sample median coverage/10.

- **User CNV classification:** custom classification for CNV variants for ACMG and AMP rules. For user-submitted VCFs with CNVs, only variants with a copy number value can be manually classified.
- **Copy Number:** estimated copy number of the CNV call calculated from the reads expected vs reads observed ratio assuming a diploid state.
- **Type**: type of CNV, can be either deletion duplication
- **Genes:** genes overlapping the CNV call region.
- **Number of genes:** number of genes overlapping the CNV call region.
- **Quality Score**: A measure of statistical support for each CNV call. Specifically, it is the log10 of the likelihood ratio of data for the CNV call divided by the null (normal copy number). The higher the Quality Score the more confident one can be about the presence of a CNV. While it is difficult to give an ideal threshold, and for short calls the scores may be unconvincing, the most obvious large calls should be easily flagged by ranking them according to this score.
  - For WGS CNV analyses, the quality score is given by delly, if it is a single sample, or by ExomeDepth for multiple samples .
  - For CNVs from VCF, provided for annotation only, the Quality Score displays the QUAL value from the VCF (if included).



Variant	Length	Call Quality	Copy Number	Type	Genes		Quality Score	Germline Class
chr16:532342..1258311	725.969	✖	3	duplication	ANTKMT,C1QTNF8,...		15	Pathogenic
chr1:817860..1366223	548.363	✔	3	duplication	ACAP3,AGRN,B3GA...		6	Pathogenic
chr19:984916..1281711	296.795	✖	3	duplication	ABCA7,ARHGAP45,...		4	Pathogenic
chr20:63273910..63573784	299.874	✔	3	duplication	ARFGAP1,CHRNA4,...		3	Pathogenic
chr1:2403562..2650428	246.866	✔	3	duplication	ENSG00000224387,...		7	Pathogenic

- **ACMG CNV class and CNV rules:** the ACMG CNV classification and the set of triggered ACMG rules. These rules are displayed in clickable bubble icons that include the rule's description and explanation for triggering.
- **Number of exons:** number of exons overlapping the CNV region.
- **Reads expected, reads observed and reads ratio:** these columns contain the values for

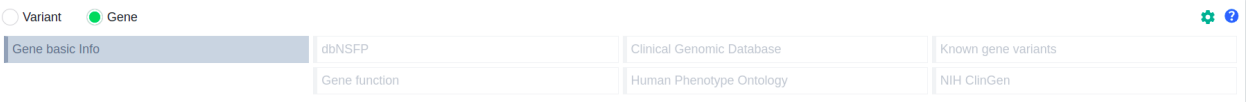
each CNV call of the reads expected, the reads observed, and the read ratio The column read ratio is calculated by dividing the number of observed reads by the number of expected reads. Since the number of expected reads is calculated from the reference set of samples, we highlight the importance of having an appropriate reference set, as we mentioned in the QC report section. Given a good reference set of samples with a good correlation between them, and all three call quality metrics passing the filters, the read ratio value can be used to rank the variants according to the strength of the signal.

CNV Rules	Number of genes	Number of exons	Reads expected	Reads observed	Read ratio	Freq
Gene Literature Overlap	66	576	0	0	-	-
Gene Literature Overlap	61	470	0	0	-	-
Gene Literature Overlap	45	561	0	0	-	-
Gene Literature Overlap	80	1.003	0	0	-	-
Gene Literature Overlap	61	884	0	0	-	-
Gene Literature Overlap	63	742	0	0	-	-
Gene Overlap Literature	2.533	24.516	0	0	-	-

- **Frequency:** frequency of overlapping CNVs in the same genomic region. The gnomAD database is used to get the general population frequencies for a given structural variant. Depending on the type of variant, the frequencies are calculated as follows:
  - Deletions: we use gnomAD variants if they fully overlap with the given variant.
  - Duplications in coding regions: we compare at the gene level and we use those gnomAD variants that encompass the same coding genes as the given variant.
  - Duplications in non-coding regions: we use gnomAD variants if they are at least covering 85% of the variant region.
- **Cytoband:** The cytoband of each CNV is displayed. In case of long CNVs spanning more than one cytoband, then they are displayed as a range.

CNV tabs

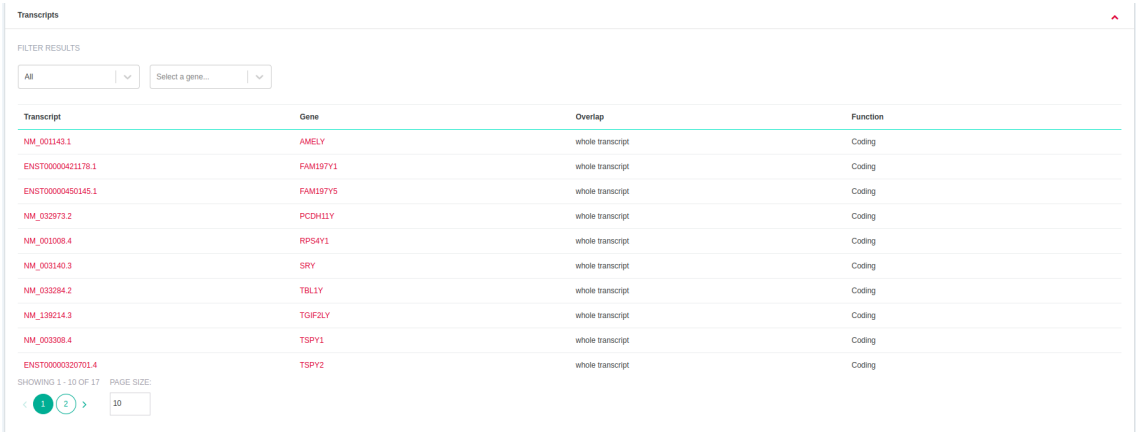
- **Genes:** the gene information for all the genes overlapping the CNV region is available at the right side of the window under the “Gene” option.



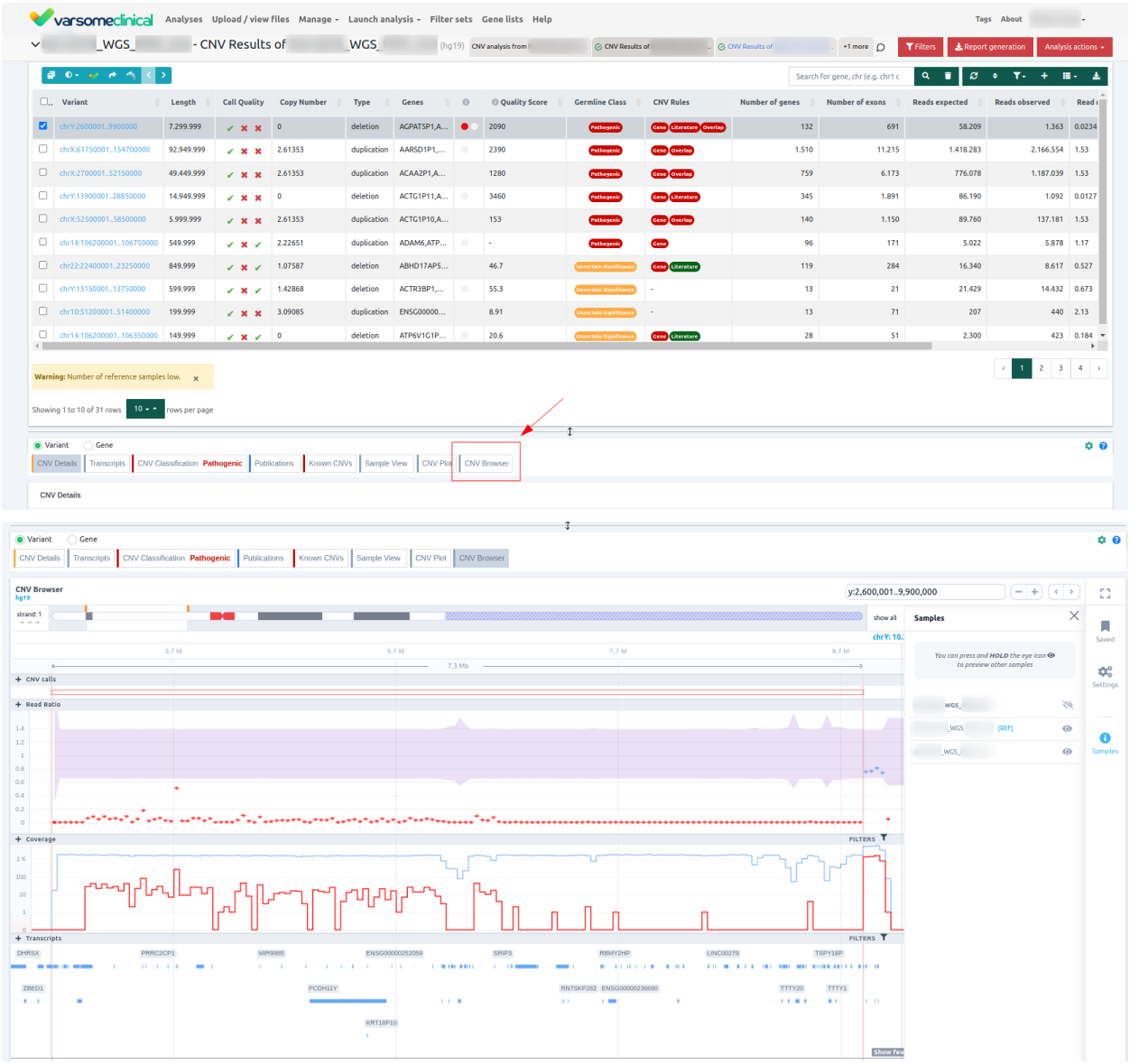
- **CNV Details**: Summary information about the selected variant (position, type, overlapping genes etc)
- **Sample View**: Sample's region browser which presents information about the

overlapping transcripts in the CNV region, conservation scores per position and SNVs of the sample. For further information please refer to section Sample View.

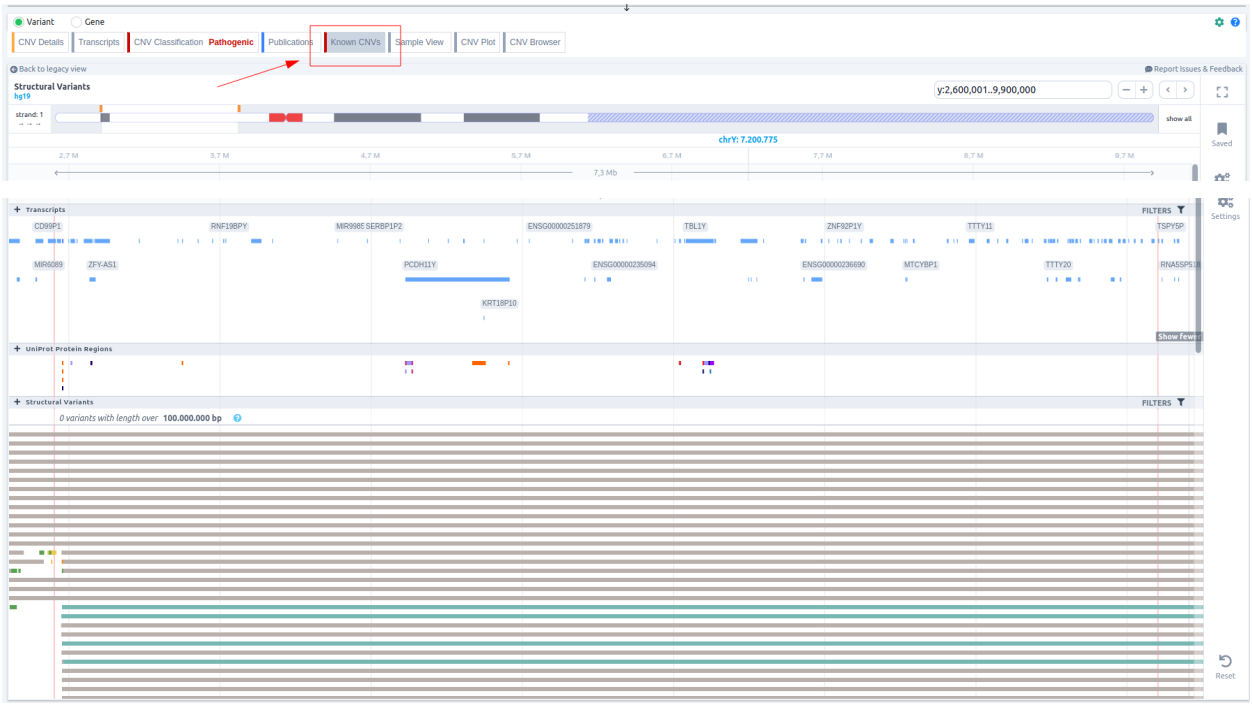
- **Transcripts**: A list of all the affected transcripts that overlap with each CNV is displayed on the bottom of the Variant Table, under the "Transcripts" tab. Transcripts can be filtered based on coding status and/or gene name.



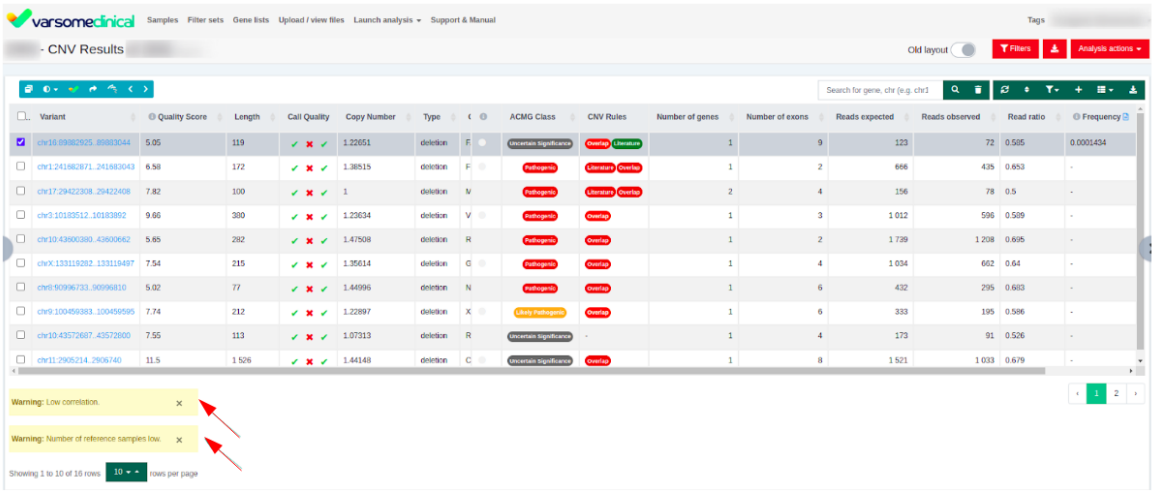
- **CNV Classification:** In this tab we show the ACMG classification for each CNV and the set of triggered ACMG rules. Click on “Show full detail” to find out the criteria not met.
- **Publications**: publications from PubMed related to the selected CNV or genes where variants, diseases, phenotypes, chemical compounds, drugs (if exist) are tagged by our internal AI tool as well as from the VarSome community users and our curation team verifies them .
- **CNV Browser:** An interactive browser showing a wider region around the position of the CNV call as well as its location on the chromosome level. The user can zoom in and out using the mouse scroll and select among different chromosomes, genomic positions, samples and CNV calls. Data points represent read ratios (observed/expected read counts). These are colored blue or red, depending if they fall within the gray shaded area 95% confidence interval or not, respectively. Call genomic location is indicated by coordinates and annotated for overlapping gene structures (exons/introns). The coverage track, at the bottom of the interactive plot, shows the trend of the coverage on a logarithmic or linear scale across all cohort samples. By hovering the browser, there is useful CNV call information including genomic location and span, as well as links to the same region in other analyses of the same cohort. You can find further information in the CNV visualization article.



- **CNV plot:** We provide a CNV plot (static), showing how the observed read depth in the area of the CNV differs from the expected. The CNV plots are generated using a modified version of the ExomeDepth tool. You can find further information regarding this in CNV Visualizations.
- **Known CNVs:** We display only the relevant CNVs for the classification according to the following criteria:
  - CNV deletions: we retain those that fully overlap with the given CNV for gnomAD variants. For CNVs coming from clinical sources (Decipher, DBVar, ClinVar CNVs) we use the overlapping CNVs if they are benign and the contained CNVs if they are pathogenic.
  - CNV duplications: we keep only the CNVs encompassing the same coding genes. If the CNV is non-coding, then we retain the CNVs that have at least 85% of overlap.



Warnings are being displayed under the variant table to inform the user of the reliability of CNV calls of the sample: if (1) the correlation of the sample to its reference samples is low and (2) the number of reference samples is low.

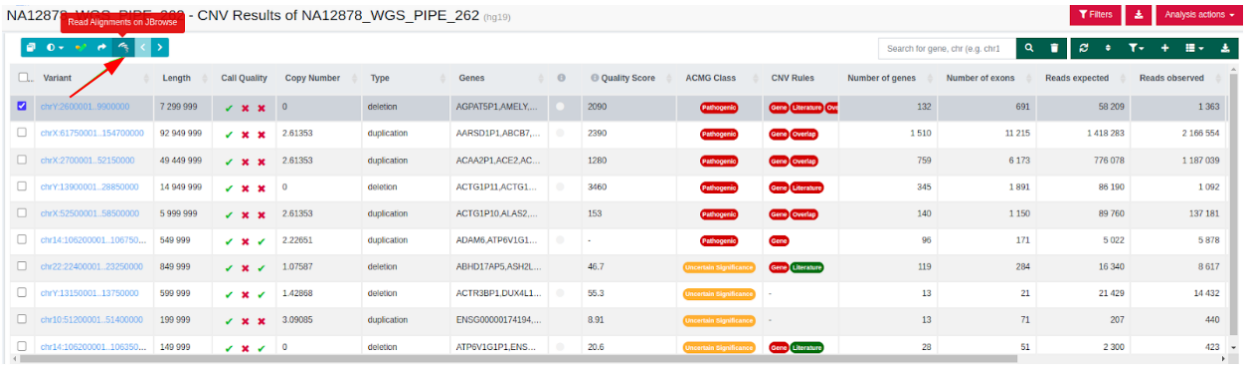


Searching through CNV results

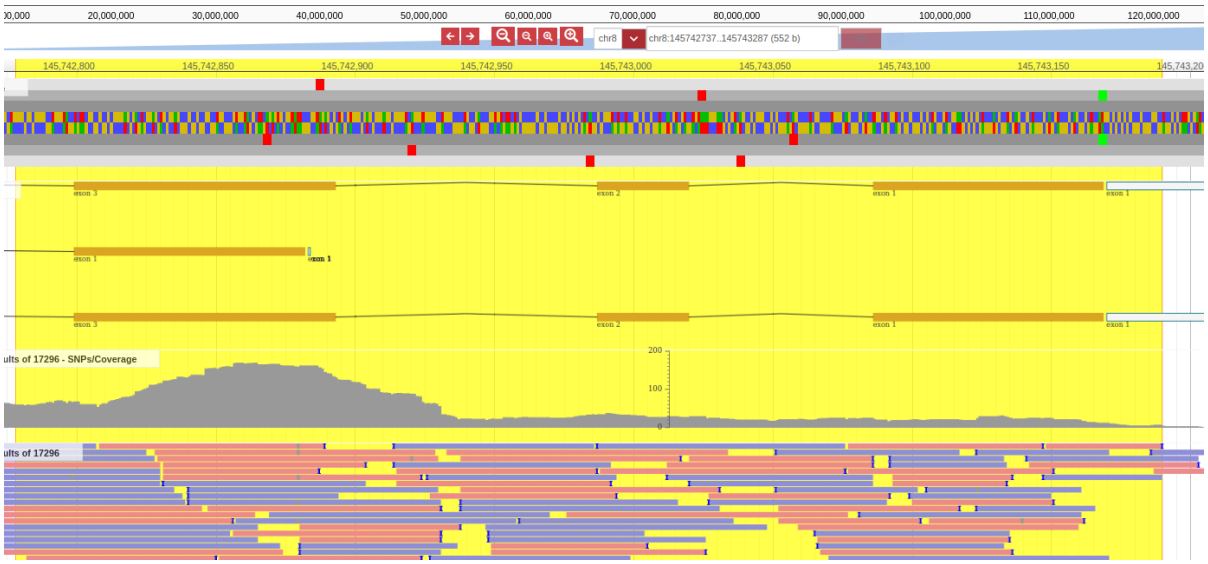
As you inspect the CNV results of your sample, you can search by a known, or previously detected from the main analysis of the sample, SNV or small INDEL and see if it overlaps with any detected CNV.

Reads alignment visualization for CNVs

You can view the alignment of the reads in the regions of the detected CNVs on JBrowse.

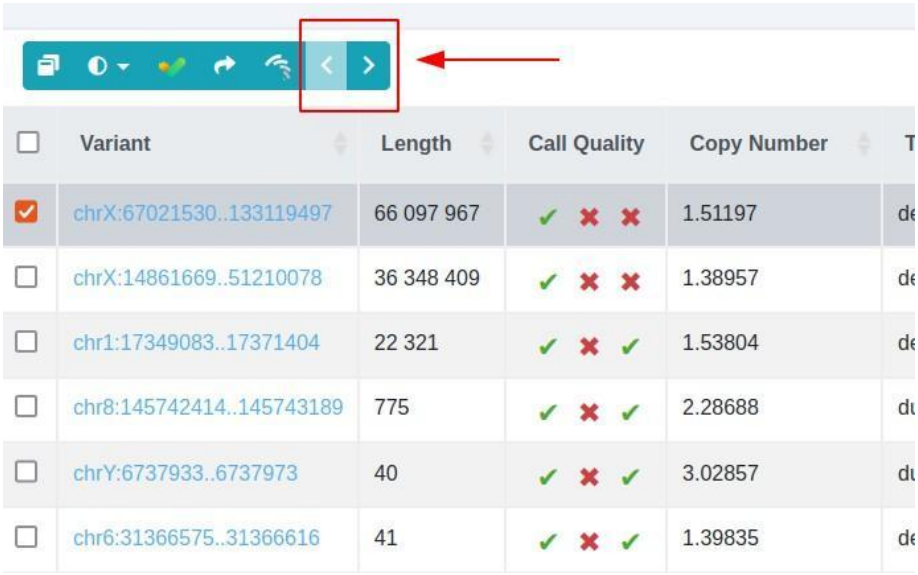


Once you have selected a variant on the Variant Table you can see the alignment of the reads by clicking on the Jbrowse icon on the top left of the screen. The CNV call region is highlighted in yellow.

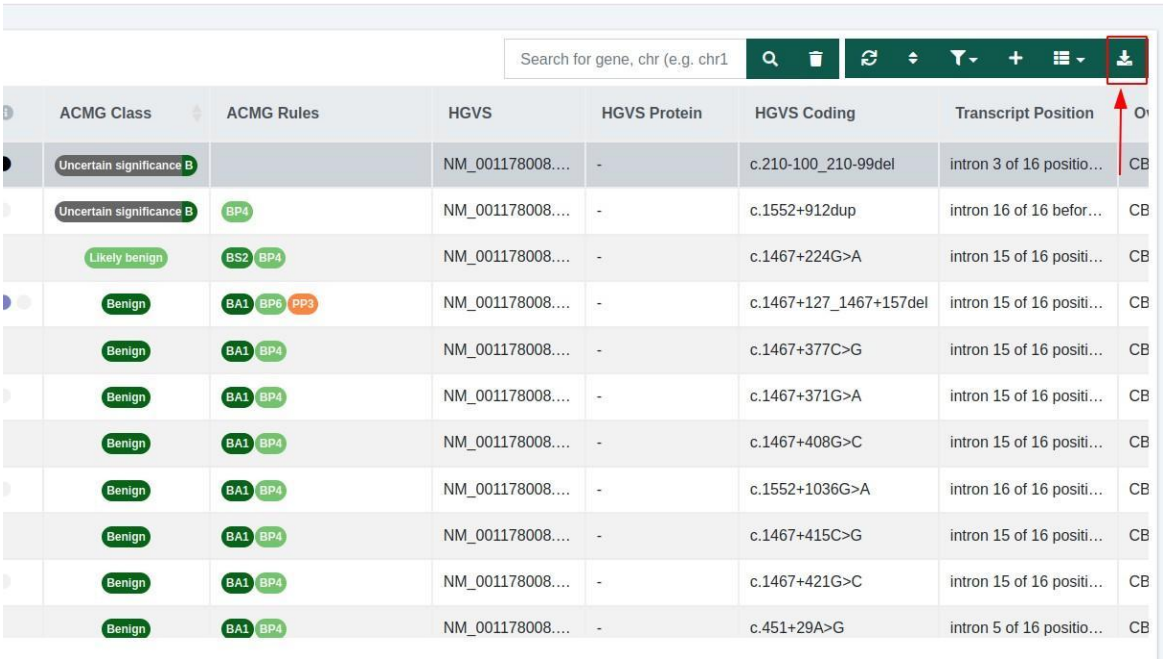


Browsing through the samples of a CNV analysis

You can browse through the samples analyzed under the same CNV/SV analysis by visiting the results page of one and using the red arrows you can be directed to the next or previous sample:

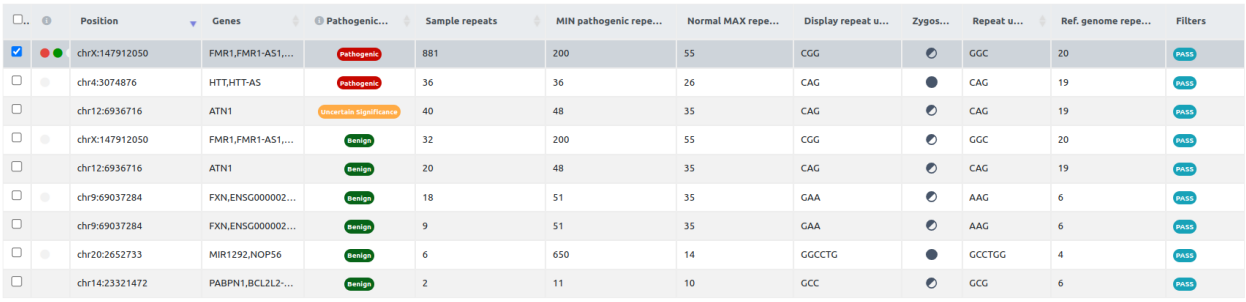


You also have the option to download you filtered CNV results, as it has been possible for SNP/small INDEL analyses, from the upper right corner of the Variant Table:



Repeat Expansion variant table

Repeat expansion analysis variant table contains the following information:



Position	Genes	Pathogenic...	Sample repeats	MIN pathogenic repe...	Normal MAX repe...	Display repeat u...	Zygos...	Repeat u...	Ref. genome repe...	Filters
chrX:147912050	FMR1,FMR1-AS1,...	Pathogenic	881	200	55	CGG		GGC	20	PASS
chr4:3074876	HTT,HTT-AS	Pathogenic	36	36	26	CAG		CAG	19	PASS
chr12:6936716	ATN1	Uncertain Significance	40	48	35	CAG		CAG	19	PASS
chrX:147912050	FMR1,FMR1-AS1,...	Benign	32	200	55	CGG		GGC	20	PASS
chr12:6936716	ATN1	Benign	20	48	35	CAG		CAG	19	PASS
chr9:69037284	FXN,ENSG000002...	Benign	18	51	35	GAA		AAG	6	PASS
chr9:69037284	FXN,ENSG000002...	Benign	9	51	35	GAA		AAG	6	PASS
chr20:2652733	MIR1292,NOP56	Benign	6	650	14	GGCCTG		GCCTGG	4	PASS
chr14:23321472	PABPN1,BCL2L2-...	Benign	2	11	10	GCC		GCG	6	PASS

- **Position**: chromosome and position of the repeat expansion.
- **Display repeat unit**: display repeat unit familiar to the clinician.
- **Repeat unit**: repeat unit in the reference orientation.
- **Pathogenicity**: repeat expansions can be classified as:
    - Pathogenic: the number of repeats detected in the sample is equal to or greater than the minimum number of pathogenic repeats.
    - Uncertain Significance: the number of repeats detected in the sample is greater than the maximum normal number of repeats but less than the minimum number of pathogenic repeats.
    - Benign: the number of repeats detected in the sample is smaller than or equal to the normal number of repeats.
- **Genes**: in which gene the repeat expansion has been found.
- **Sample repeats**: number of repeats observed in the allele.
- **MIN pathogenic repeats**: minimum number of repeats needed to consider the variant pathogenic.
- **Normal MAX repeats**: maximum number of repeats allowed to call the variant as normal.
- **Ref. genome repeats**: number of repeats in the reference genome.
- **Zygosity**
- **Filters**: VCF filters.

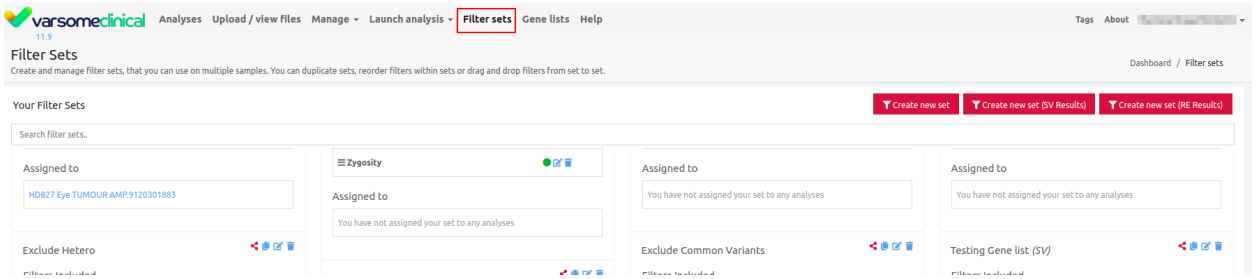
Repeat expansion cards:

- **Variant:** general variant information and region browser.
- **Gene:** it contains the same gene cards as the ones displayed under the gene level of the main analysis (small variant table).

8. Filter Sets

VarSome Clinical allows you to create 3 different filter sets:

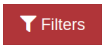
1. Create new set
2. Create new set (SV Results): specific filter sets for CNV analysis' results.
3. Create new set (RE Results): specific filter sets for Repeat expansion analysis' results.



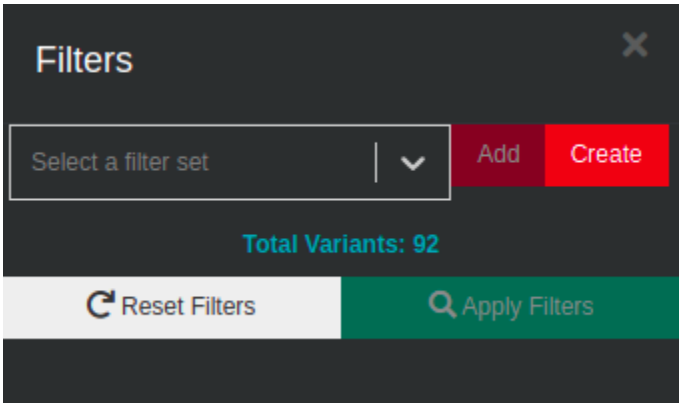
8.1 Create a new filter set

It is possible to create a filter set directly from the analysis results or from the “Filter sets” menu.

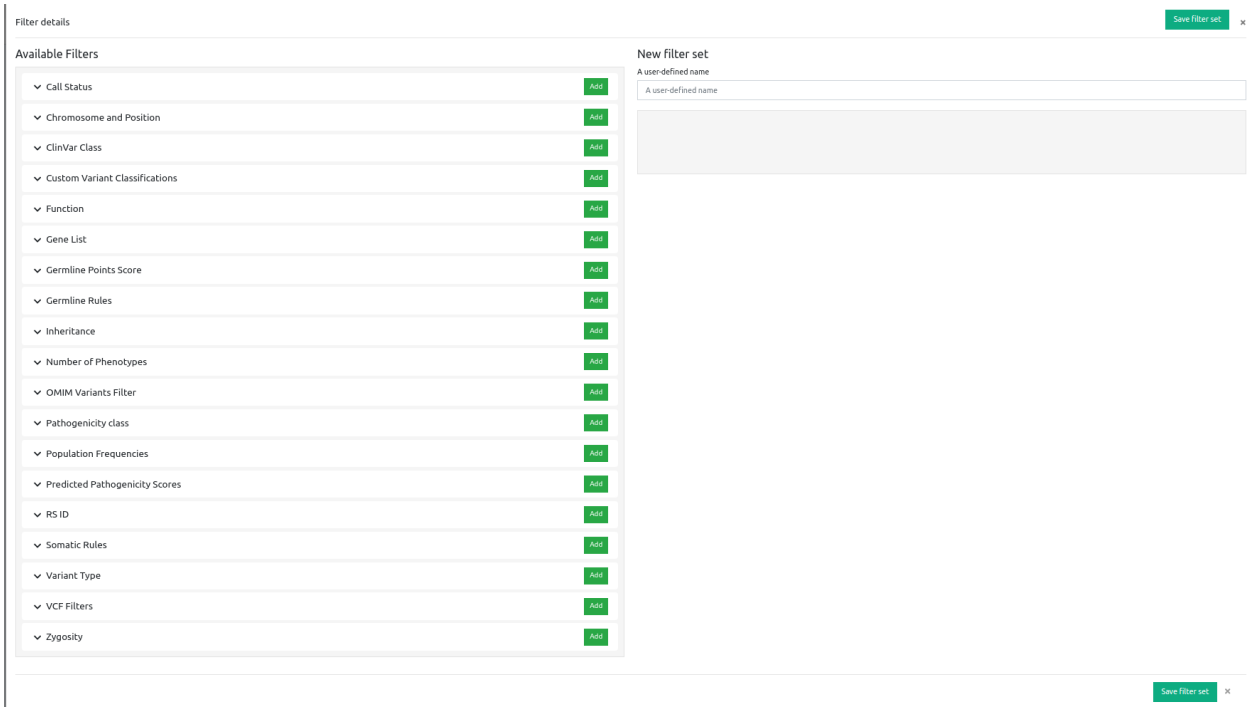
From the analysis results:

- Click on the **Filters** icon  to create and manage filter sets, which you can use on multiple samples. You can duplicate sets, reorder filters within sets or drag and drop filters from set to set.

- Click on the **Create** icon  to create a new filter set.

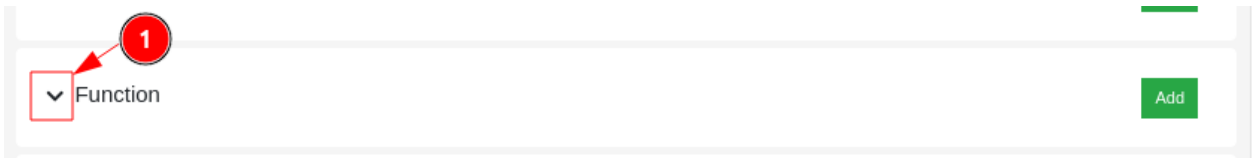


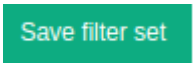
A pop-up window called **Filter set** will appear. This window is split into two columns: **Available Filter Sets** and **New Filter set**. The available filter sets can be selected by clicking on the green box ("Add") next to each filter's name. If you select a filter, the filter will move to the other column. By clicking on the "filter name" or on the arrow, a form is displayed where the filter can be edited. Once you have finished editing your new filter set, give it a **Name** and click on **Save filter set**.

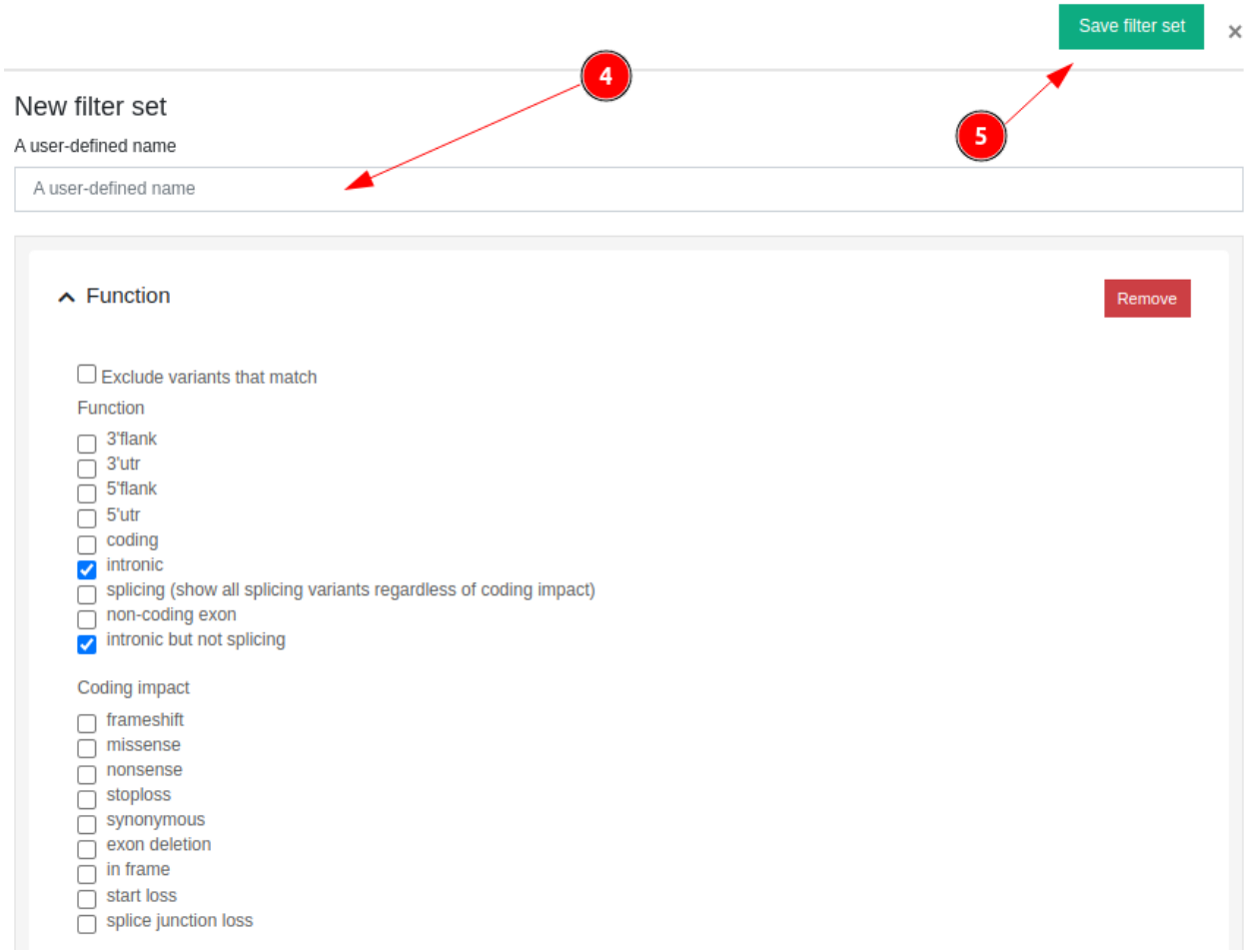


For example, if you want to create a filter that will give you all the variants except from the variants that are only intronic you can click on the Functions filter and then to intronic and intronic but not splicing functions as shown in the pictures below.

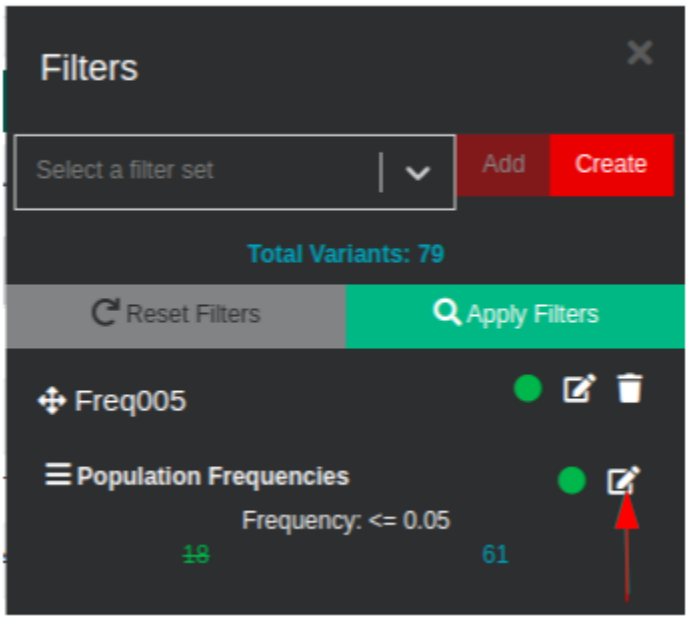
OBJ

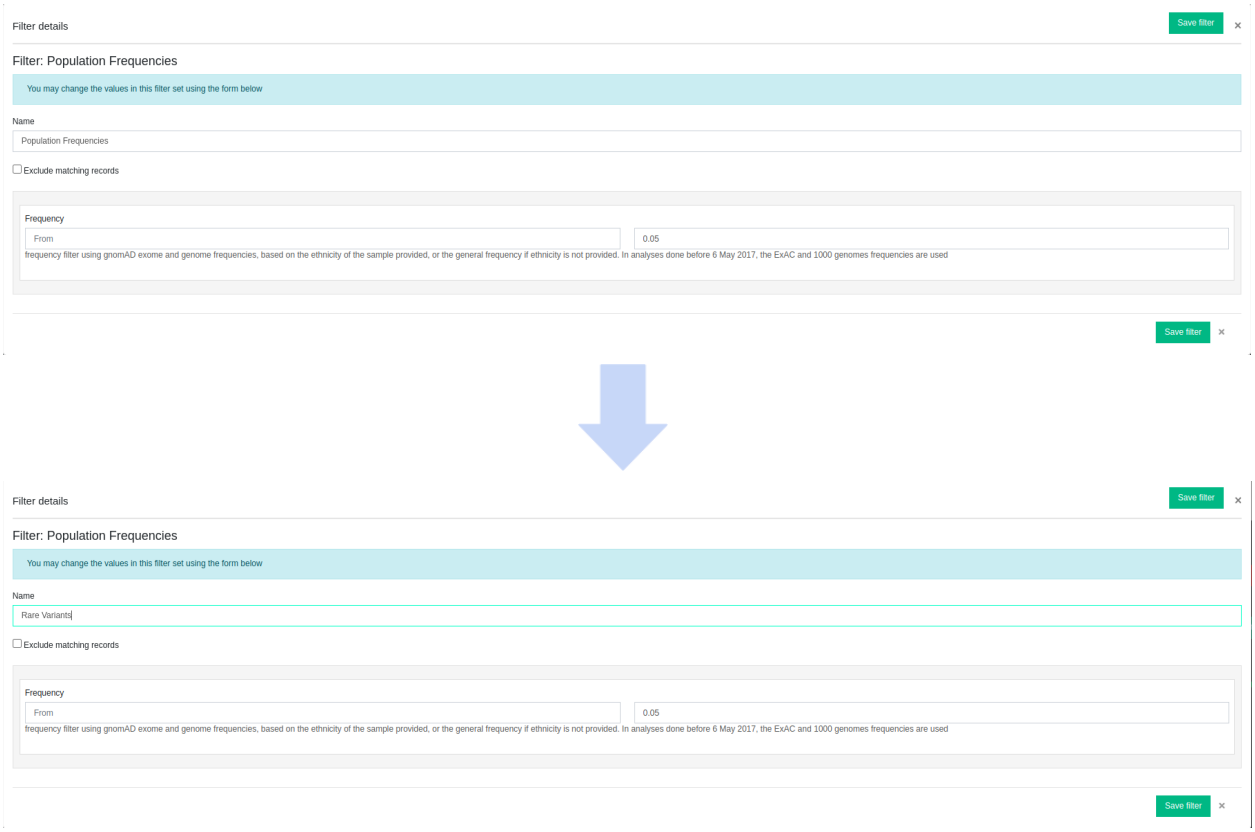


By clicking on the Add button  the filter will move to column “New filter set” in which you can give to your filter the name you want, you can edit your filter and finally you can save it by clicking on the Save filter set button  .

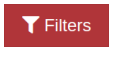


You can modify the name of the filter set by clicking on  . In the following example, the filter set "Population Frequencies" will be renamed "Rare Variants":





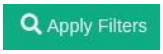
8.2 Applying an existing filter to a variant table

Clicking on the filter icon 

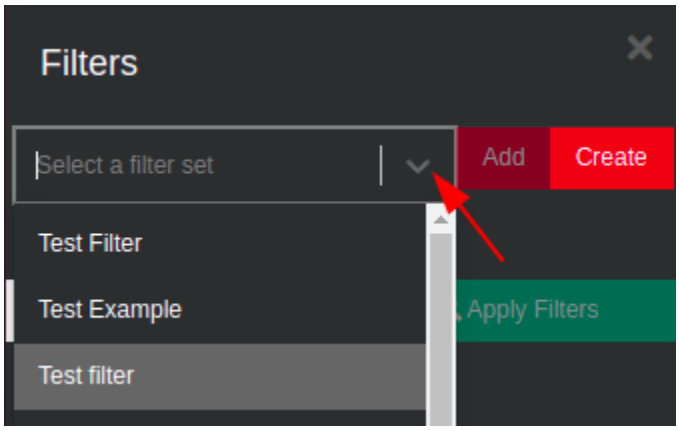
it opens the **Filters** menu where filter sets can be managed. To exit this menu, simply press "**Esc**" on your keyboard.

A drop-down list will show a list of all saved filter sets. Select a filter set from the already existing ones and add the selected filter set to the analysis with: 

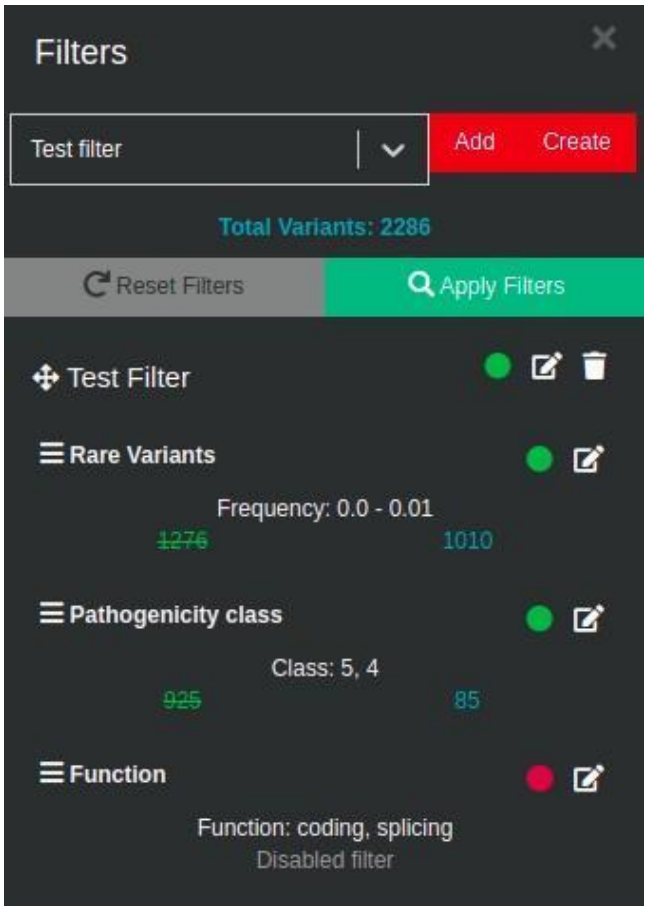
. In order to apply the

selected filter set to the analysis select: 

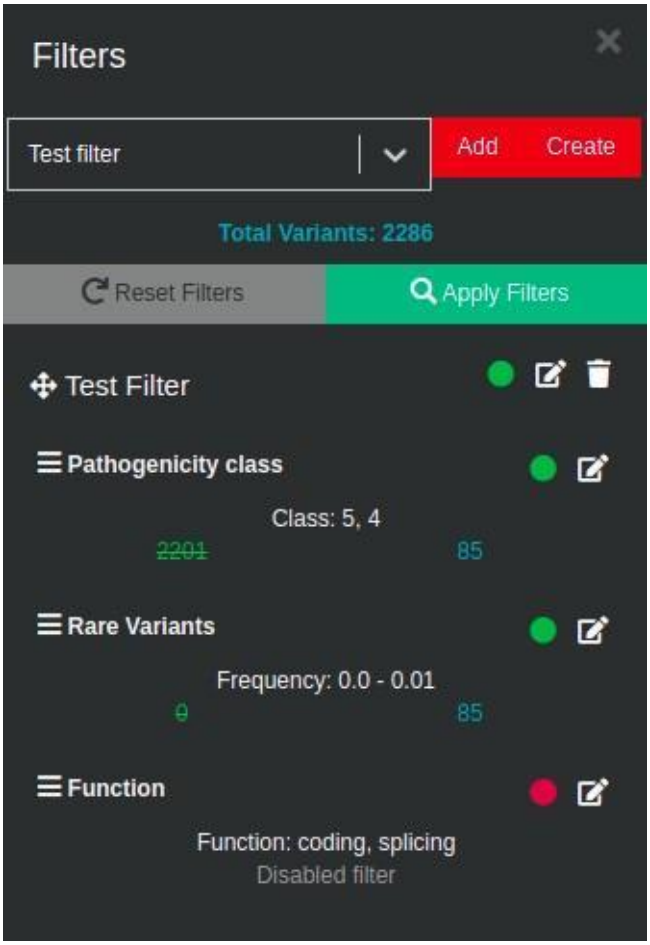
.



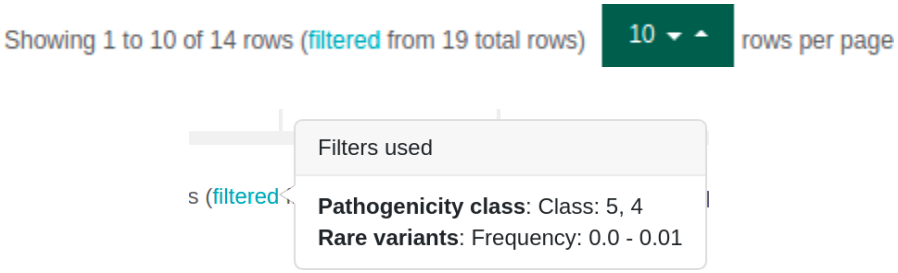
Each of the filters inside a filter set can be switched on and off by clicking on the green/red circles. Active filters have an associated green circle, and inactive filters have an associated red circle. Variants passing the filter are shown in blue, and variants that fail the filter are shown in green and crossed out. In the example below, all filters are active except the Function filter.Filters can be edited  or removed  .



The order of the filters can be changed by moving the mouse on the top of  , then dragging the filter and dropping it at the desired position. In the example below, the “Pathogenicity class” filter has been moved to the top position:

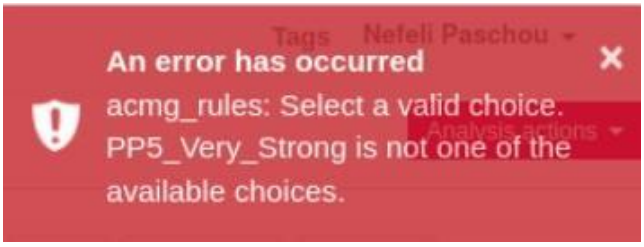


Click on Apply Filters in order to filter the original results of the variant table. The number of variants that pass the filter is also shown under the variant table The number of rows of the variant table can be adjusted by selecting: . The filters used are shown when placing the cursor on "filtered".



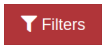
When downloading all the filtered variants from the Variant table in excel format there will be a separate tab called "Filters used" that mentions which filters were used.

⚠️ Please note that if existing filters with ACMG rules do not work, you will see an error message returned, like the following:

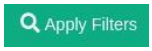


In this case you should edit (as shown above) the filter, by adding the rules you would like to have from the available options.

8.3 Shared Filters

Users of the same group can now share filters among them. To share a filter with other members of your group, click the Filters link at the navigation bar and then click the share icon on any of the filters you wish to share.
Filters that have been shared among members of the same group appear both on the Filters screen and when filtering a specific analysis. Although a shared filter can be used by anyone in the group, it can only be edited by the user that initially shared the filter. Other group members can still duplicate the filter and edit it. In the analysis results page, click on the Filters icon 

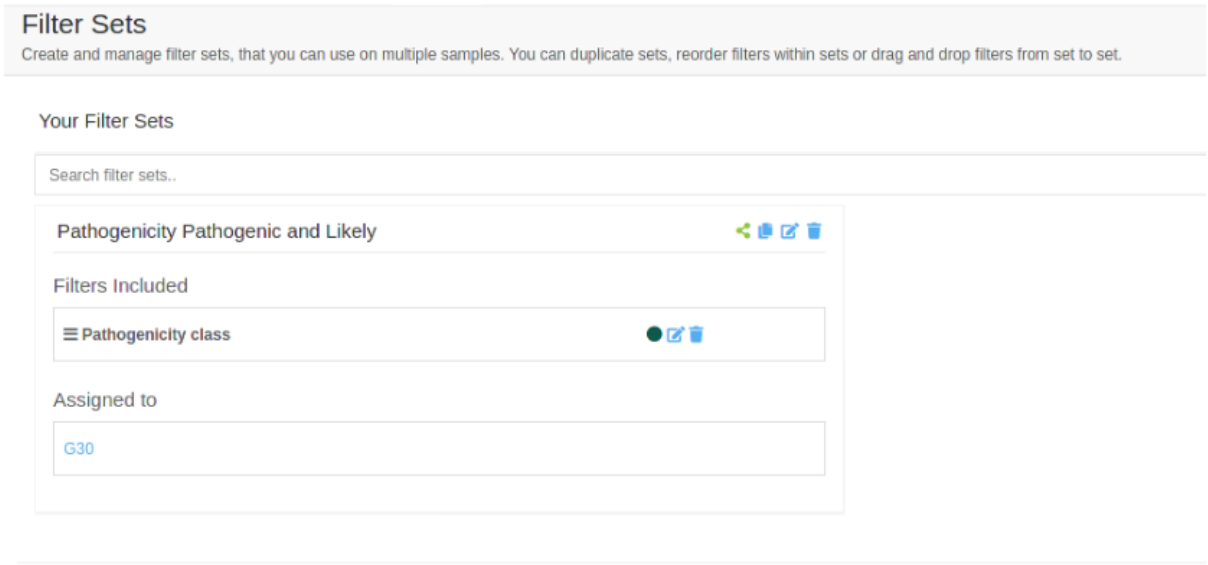
, use the drop down list of the search box to select the shared filter of interest, click on 

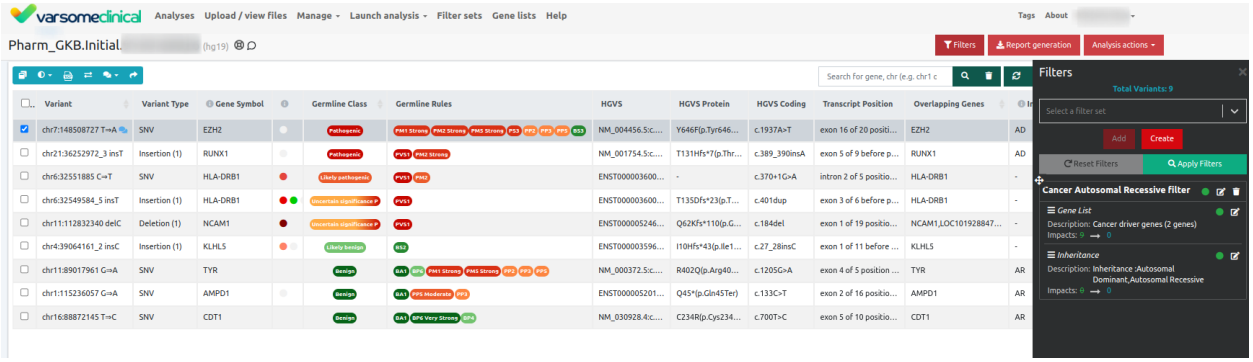
to add the shared filter set to the analysis and then click on 

.

The sharing icon is a click toggle to switch between sharing status. Filter sets shared with you will show up normally on analysis side-panel alongside your owned filter sets (if they are shared, it will say by whom). Note that filter-sets sharing is only within VarSome Clinical organizations, you cannot share them, for example, with people outside your organization.

Care must be taken for shared filters because user-created filter sets are essentially composed of filter instances (e.g. a filter instance is the filtering of pathogenic variants), and filter instances can be turned on or off by the creator of the filter set. So, if the filter set is shared, other users who use this filter set may be able to apply it to their analyses, but may see that not all filter instances are enabled, and they can do nothing about it. If the original creator of the filter set chooses to, for example, disable a filter instance, then all other users who use that shared filter set will see changes in their results because of that. To avoid this you can duplicate the shared filter set as one of your own so that you have direct access to what is enabled or not.

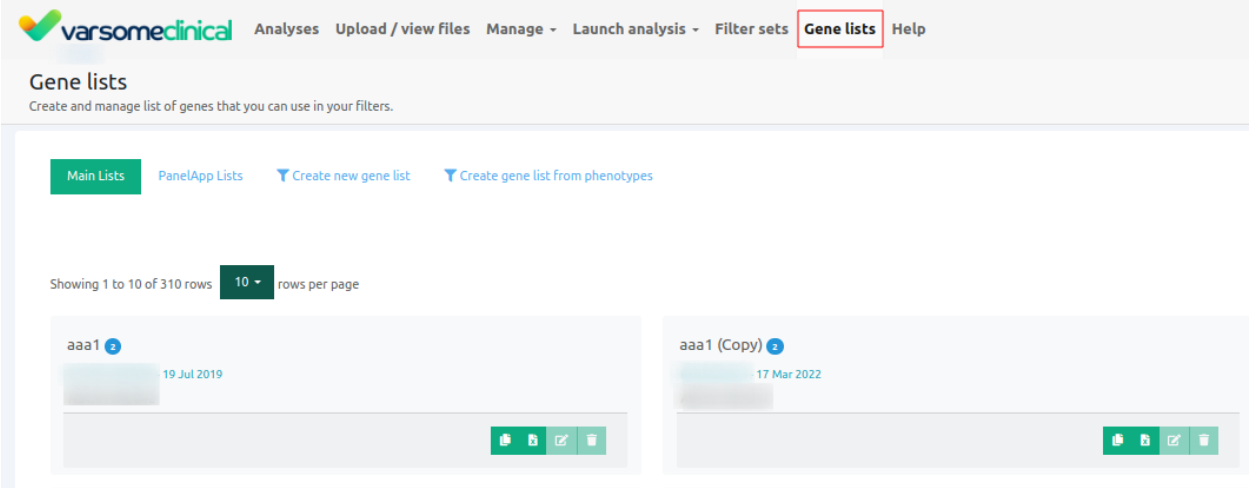

Filter Sets
Create and manage filter sets, that you can use on multiple samples. You can duplicate sets, reorder filters within sets or drag and drop filters from set to set.
Your Filter Sets
Search filter sets..
Pathogenicity Pathogenic and Likely
Filters Included
Pathogenicity class
Assigned to
G30
Shared Filter Sets (Saphetor)
Users can toggle sharing if Filter Sets by clicking the icon
x-link
(Shared by:)
Filters Included
CGD Inheritance
Assigned to
_fastq
_fastq - Analysis filtered by Intellectual disability PanelApp 20201112 as of 27-Nov-2020 (27 Nov 2020 Intellectual_disability.Pan...)
_fastq - Analysis filtered by Gene list from (any) phenotypes Autistic Behavior, Stereotypy, Absent Speech, Large For Gestational Age, Intellectual Disability, Severe as of 27-Nov-2020 (27 Nov 2020 Gene list from (any) phenot...)


varsomeclinical
Analyses
Upload / view files
Manage
Launch analysis
Filter sets
Gene lists
Help
Tags
About
Pharm_GKB.Initial
Filters
Report generation
Analysis actions
Variant
Variant Type
Gene Symbol
Germline Class
Germline Rules
HGVS
HGVS Protein
HGVS Coding
Transcript Position
Overlapping Genes
chr7:148508727 T>A
chr21:36252972_3 insT
chr6:32551885 C>T
chr6:32549584_5 insT
chr11:112832340 delC
chr4:39064161_2 insC
chr11:89017961 G>A
chr1:11523605 7 G>A
chr16:88872145 T>C
Filters
Total Variants: 9
Select a filter set
Add
Create
Reset Filters
Apply Filters
Cancer Autosomal Recessive filter
Gene List
Description: Cancer driver genes (2 genes)
Impacts:
Inheritance
Description: Inheritance :Autosomal Dominant,Autosomal Recessive
Impacts:

9. Gene List

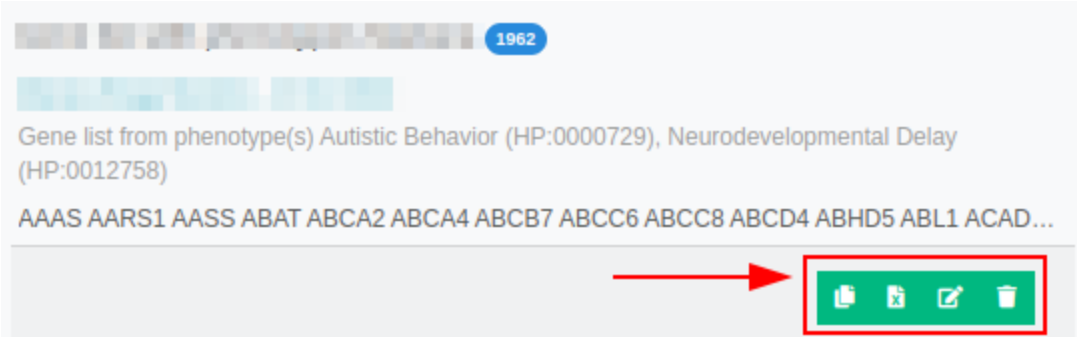
VarSome Clinical offers 3 ways of creating a new gene list:

1. Copy and modify an existing gene list
2. Create a new gene list
3. Create a gene list from phenotypes



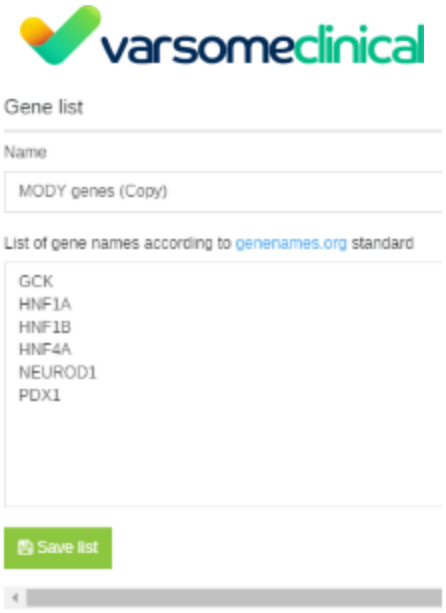
9.1 Copy and modify an existing gene list

The options to manage the existing gene lists are the following:



Gene lists can be copied by clicking on the "Copy" button  .

Genes can be added or removed by clicking on the "Edit" button  . Once you have finished editing the gene list, you can save the list with either the original name or with a new one.

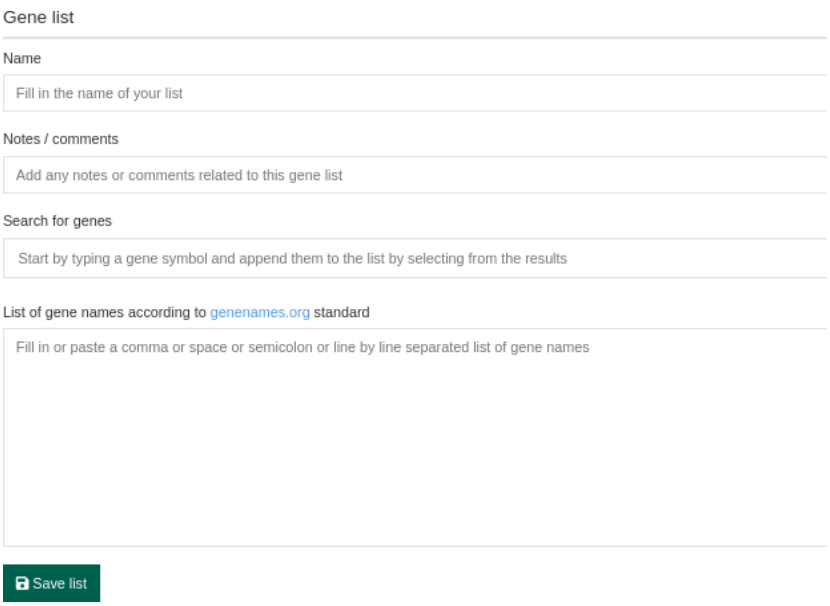


Gene lists can be exported to excel by clicking on the "Excel" button  .

Finally, these gene lists can also be deleted by clicking on the "Trash" button  .

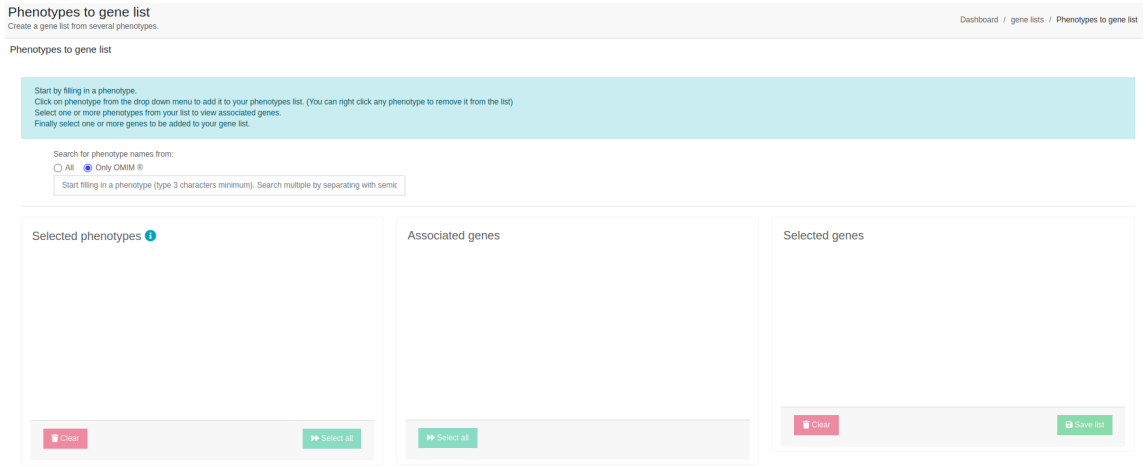
9.2 Create a new gene list

- Click on “Create new gene list”.
- A pop up form window appears, you just need to write a name for your gene list and paste the list of genes that you want to analyze. Important note: the gene names need to follow the HGNC (Hugo Gene Nomenclature Committee) standard (http://www.genenames.org/), if the gene names do not follow this rule you will get an error message after clicking on **Save list**.

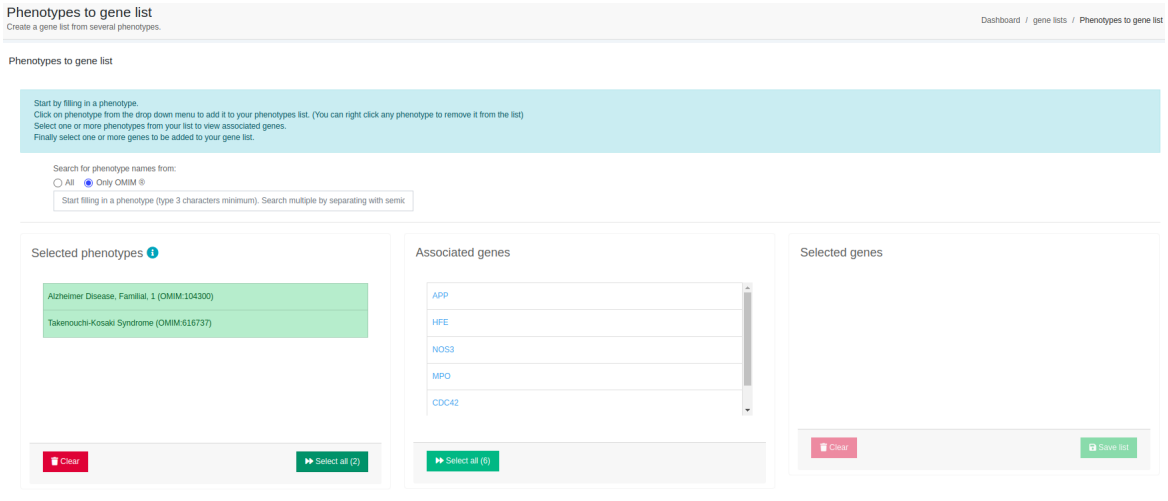


9.3 Create a gene list from phenotypes

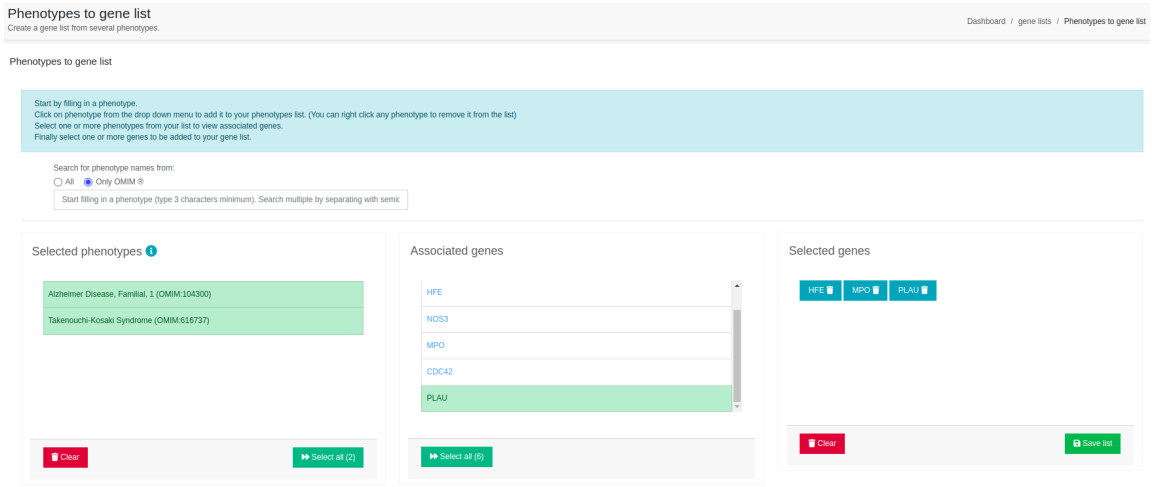
- Click on ”Create gene list from phenotypes”
- Start by filling in a phenotype. By selecting All, phenotype matching records from HPO, MONDO and OMIM® will be available



- Click on phenotype from the drop down menu to add it to your phenotypes list. (You can right click any phenotype to remove it from the list)
- Select one or more phenotypes from your list to view associated genes.



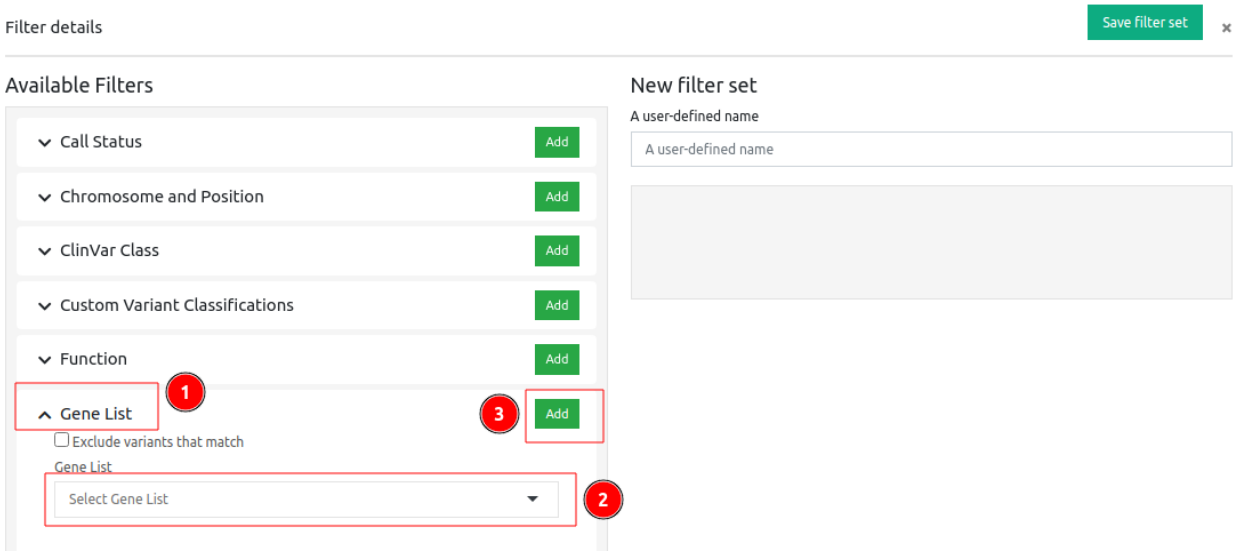
- Finally select one or more genes to be added to your gene list



Note: The methodology to generate a gene list from phenotype(s) has been changed. Previously, when making a gene list from phenotypes, we would include only those genes that are directly annotated with that phenotype. We have now extended this and instead first collect all diseases linked to the phenotype and then all genes linked to those diseases, as well as any genes directly linked to the phenotype. We already worked this way when adding phenotypes to analyses, so this change ensures we are consistent and also makes sure we don't miss any genes when creating gene lists.

If you want to filter your gene list, you have to go to the Variant table page of the CNV analysis you performed and click on the filter icon 

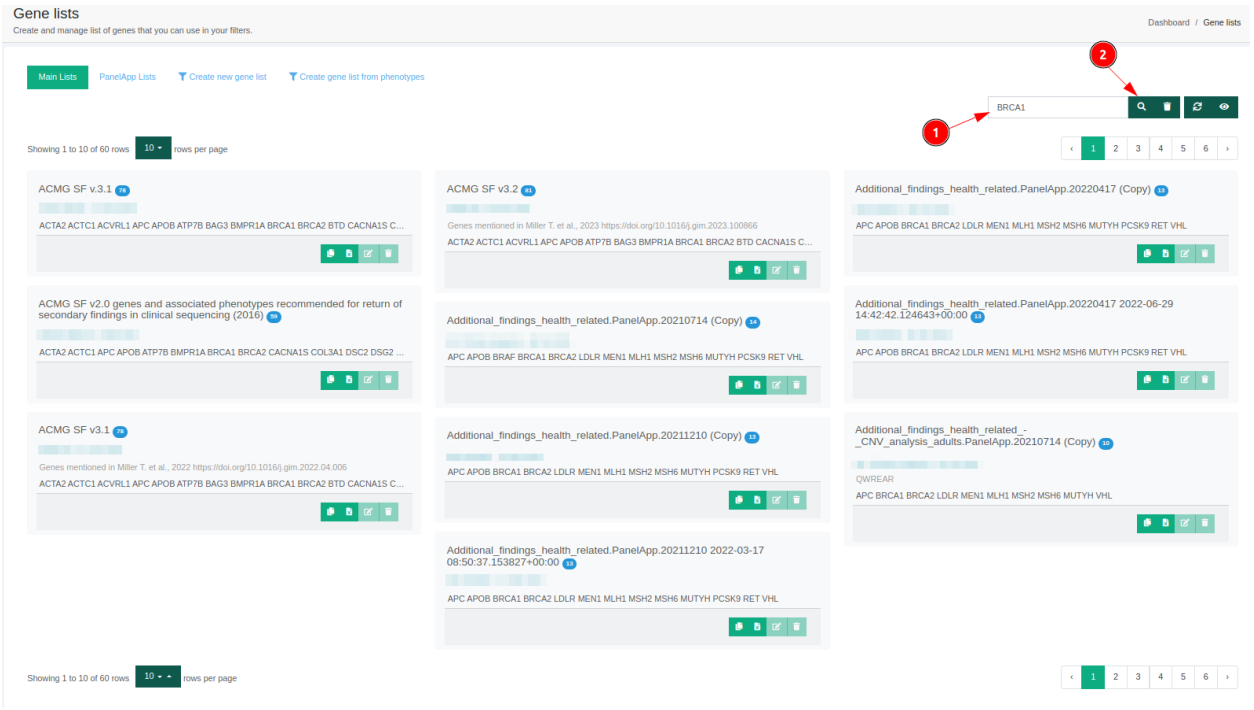
on the left. Click on "Create" to create a new one and you will see a screen like the one below:



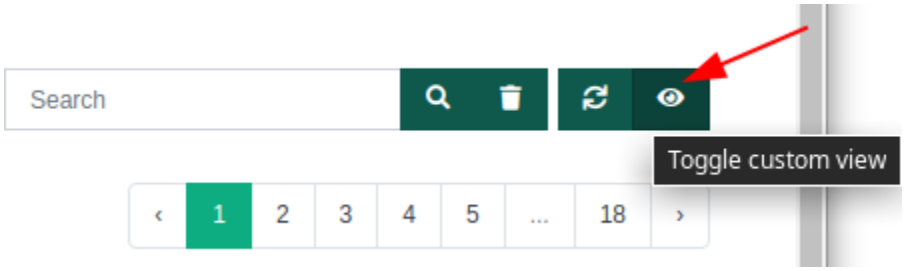
Click to add a Gene list filter for the specific phenotype you wish. Fill in the name of the filter set and save it in order to be applied on your analysis.

9.4 Browsing through gene lists

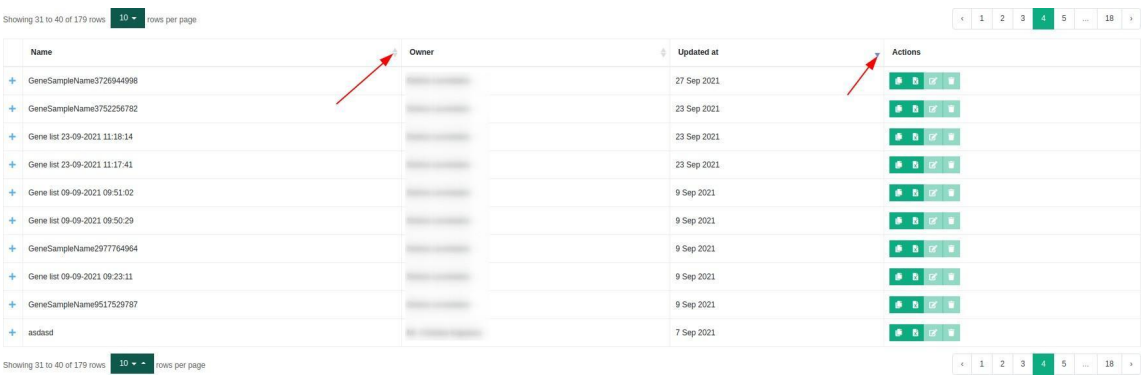
- Search gene lists by name:



- Toggle custom view:



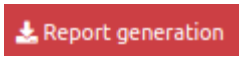
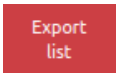
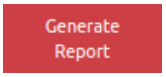
- Display gene lists sorted by name or update date:

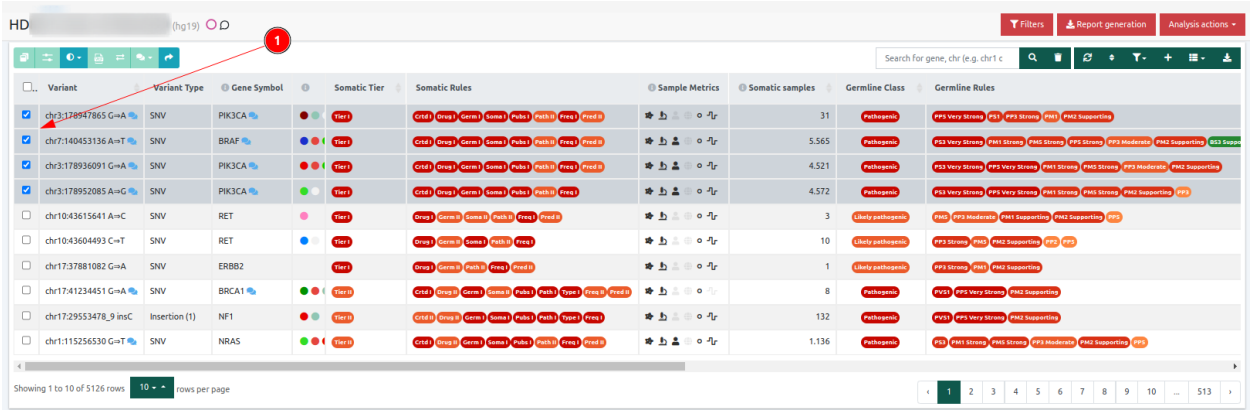


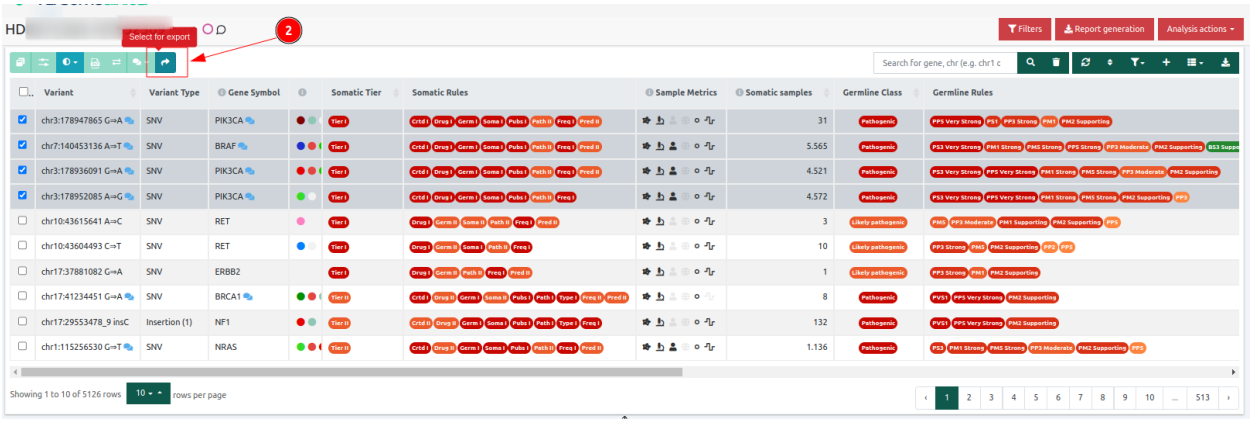
10. Final Report Generation

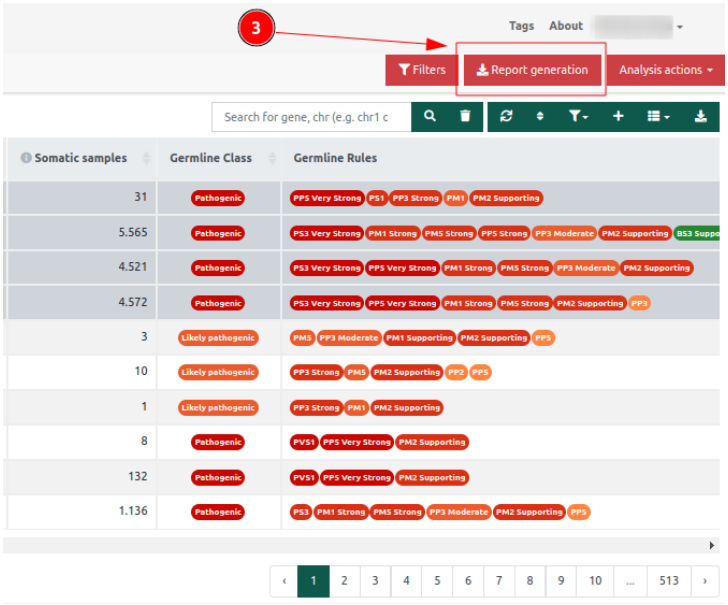
When it comes to reporting, in addition to the options provided by the "Analysis Actions", VarSome Clinical offers users the option to generate a report of specific variants in PDF or Word format.

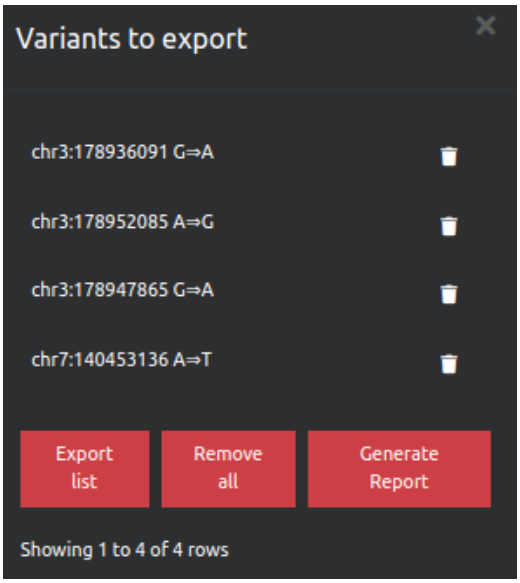
Variant reporting works on a sample level, i.e. you can report multiple variants belonging to the same sample, but you can't report one variant of the sample A and another of the sample B within the same report.

In the *Variant Table* page, you can either select one variant at a time just by clicking on it or you can select multiple variants at once (mass selection) by holding Ctrl and clicking on each variant you wish to include to your list and then click on the "Select for export"  option. On the left, click on the "Report generation"  to see the list of the variants selected for export, and then either choose  or  .


1
Filters
Report generation
Analysis actions
Search for gene, chr (e.g. chr1 c
Variant
Variant Type
Gene Symbol
Somatic Tier
Somatic Rules
Sample Metrics
Somatic samples
Germline Class
Germline Rules
chr3:178947865 G⇒A SNV PIK3CA 31 Pathogenic
chr7:140453136 A⇒T SNV BRAF 5.565 Pathogenic
chr3:178936091 G⇒A SNV PIK3CA 4.521 Pathogenic
chr3:178952085 A⇒G SNV PIK3CA 4.572 Pathogenic
chr10:43615641 A⇒C SNV RET 3 Likely pathogenic
chr10:43604493 C⇒T SNV RET 10 Likely pathogenic
chr17:37881082 G⇒A SNV ERBB2 1 Likely pathogenic
chr17:41234451 G⇒A SNV BRCA1 8 Pathogenic
chr17:29553478_9 insC Insertion (1) NF1 132 Pathogenic
chr1:115256530 G⇒T SNV NRAS 1.136 Pathogenic
Showing 1 to 10 of 5126 rows
10
rows per page
1 2 3 4 5 6 7 8 9 10 ... 513


2
Select for export
Filters
Report generation
Analysis actions
Search for gene, chr (e.g. chr1 c
Variant
Variant Type
Gene Symbol
Somatic Tier
Somatic Rules
Sample Metrics
Somatic samples
Germline Class
Germline Rules
chr3:178947865 G⇒A SNV PIK3CA 31 Pathogenic
chr7:140453136 A⇒T SNV BRAF 5.565 Pathogenic
chr3:178936091 G⇒A SNV PIK3CA 4.521 Pathogenic
chr3:178952085 A⇒G SNV PIK3CA 4.572 Pathogenic
chr10:43615641 A⇒C SNV RET 3 Likely pathogenic
chr10:43604493 C⇒T SNV RET 10 Likely pathogenic
chr17:37881082 G⇒A SNV ERBB2 1 Likely pathogenic
chr17:41234451 G⇒A SNV BRCA1 8 Pathogenic
chr17:29553478_9 insC Insertion (1) NF1 132 Pathogenic
chr1:115256530 G⇒T SNV NRAS 1.136 Pathogenic
Showing 1 to 10 of 5126 rows
10
rows per page
1 2 3 4 5 6 7 8 9 10 ... 513

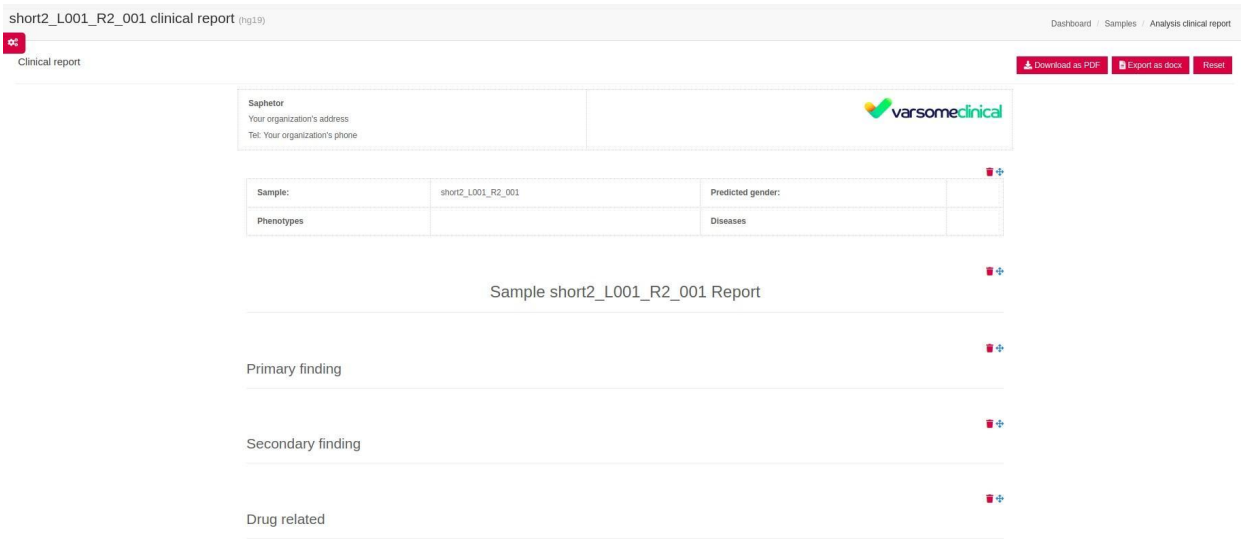

3
Tags
About
Filters
Report generation
Analysis actions
Search for gene, chr (e.g. chr1 c
Somatic samples
Germline Class
Germline Rules
31 Pathogenic
5.565 Pathogenic
4.521 Pathogenic
4.572 Pathogenic
3 Likely pathogenic
10 Likely pathogenic
1 Likely pathogenic
8 Pathogenic
132 Pathogenic
1.136 Pathogenic
1 2 3 4 5 6 7 8 9 10 ... 513



Selecting Export list will export the variants in an .xlsx file as shown below:

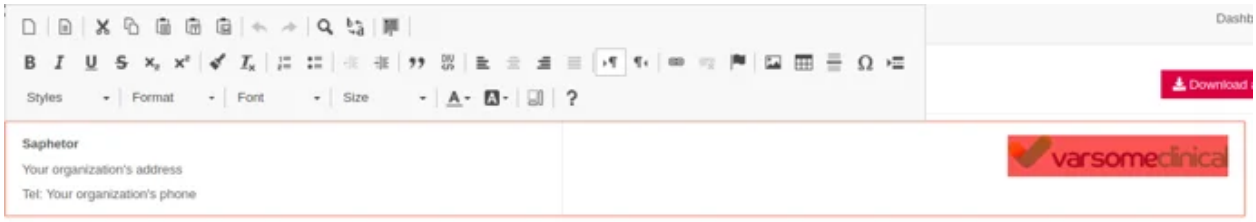
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Variant	Chromoso	Position	RS ID	Ref seq	Var seq	Type	HGVS	Genes	Phenotype	Number of	Transcript	OMIM phe	OMIM inhe	Function	Functions	Coding im	Inheritanc	ClinVar cla	ClinVar dis	Cosmic pri	Allelic bala
2	chr3:1789	chr3	178936091	rs1048860	G	A	SNV	NM_006218	PIK3CA			View	Gastric cancer	Nonsmall	coding	coding	missense	AD	Pathogenic		Breast	0.0732600
3	chr3:1789	chr3	178952085	rs1219132	A	G	SNV	NM_006218	PIK3CA			View	Gastric cancer	Nonsmall	coding	coding	missense	AD	Pathogenic		Breast	0.1323
4	chr3:1789	chr3	178947865	rs5877769	G	A	SNV	NM_006218	PIK3CA			View	Gastric cancer	Nonsmall	coding	coding	missense	AD	Pathogenic		Large Intes	0.2897
5	chr7:1404	chr7	140453136	rs1134880	A	T	SNV	NM_004333	BRAF			View	Nonsmall ce	Autosomal	coding,non	coding	missense	AD	Pathogenic		Thyroid	0.1013
6																						

Selecting Generate Report direct you to the following screen:

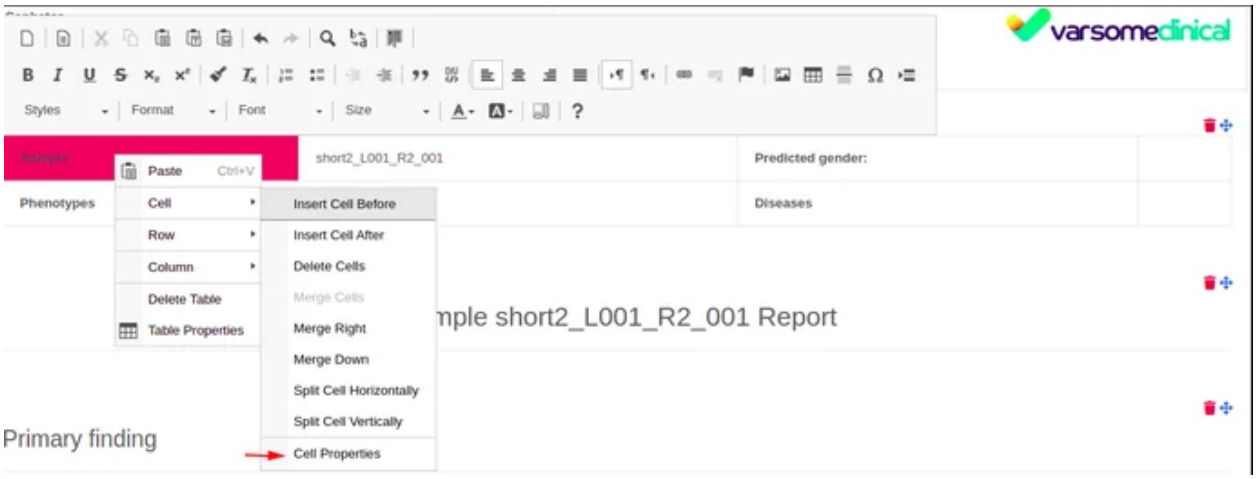


The form we provide as a template to create your report is completely customizable, where all section tables can be edited according to the user's needs . The final report layout you choose to generate will be saved for future use.

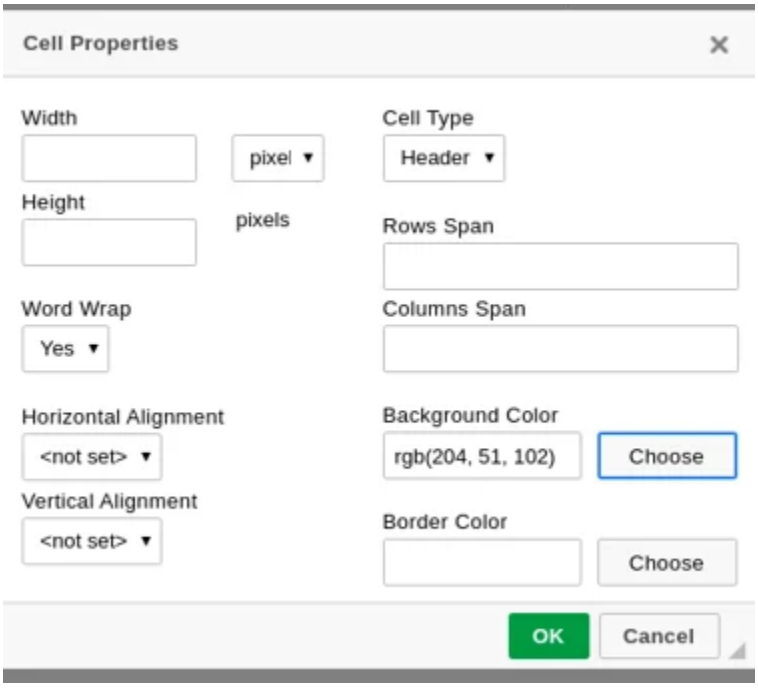
The report template consists of table sections, where all fields can be changed, along with the colors of both text and background and even the logo can be replaced with the logo of your organization.



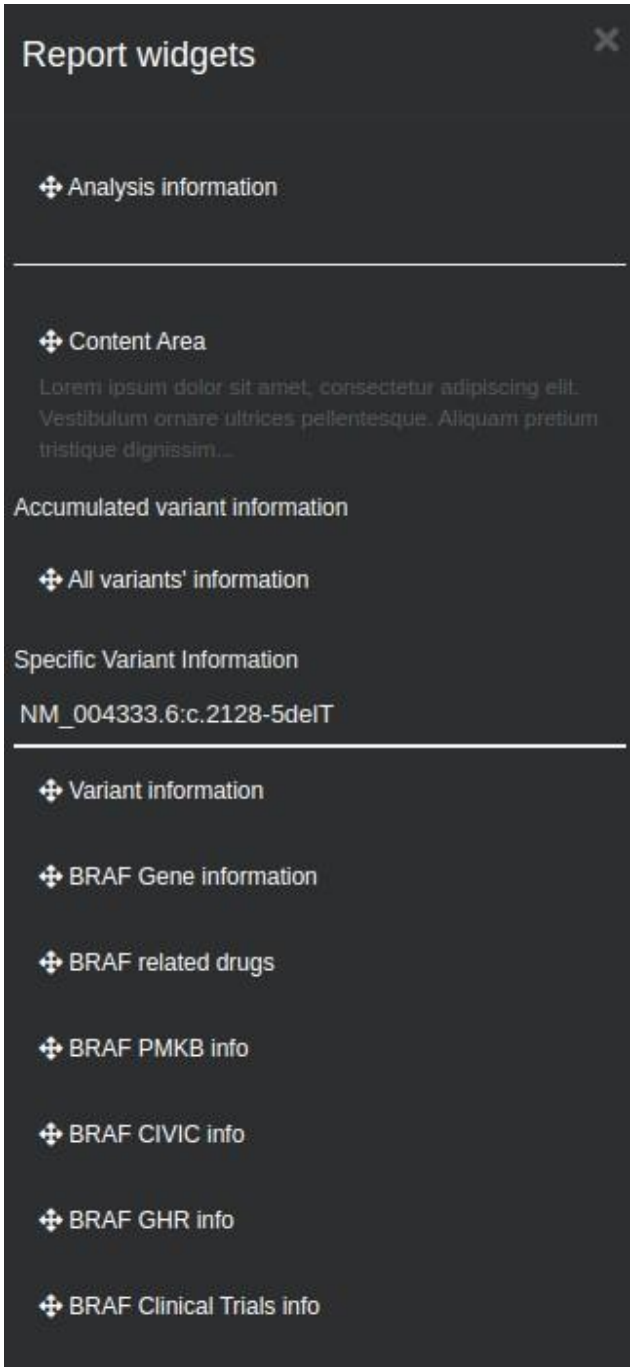
In each section table, there is the option of changing the background color of a cell by right-clicking on it and then going to Cell >> Cell properties as shown below:



In the Cell properties menu you can choose to change the background color among other available styling options:



What is more, by clicking on the  icon, a Report widgets menu will be shown, with the following options:



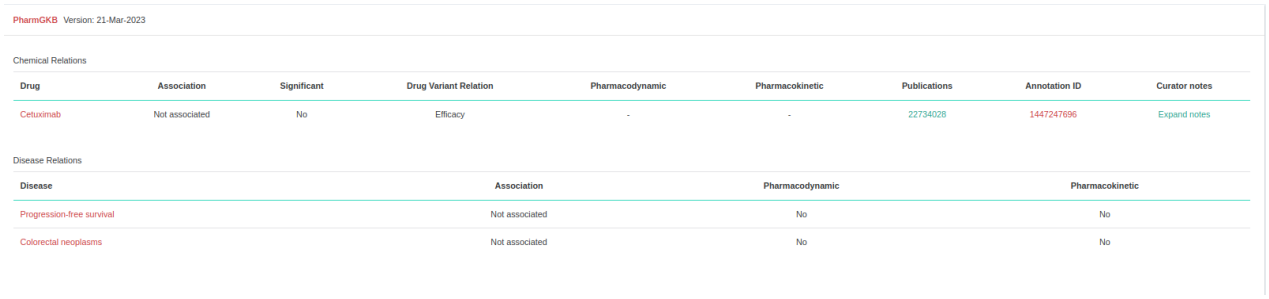
- **Analysis information:** General information including sample name, allelic balance, predicted gender, phenotypes, and diseases.
- **Content area:** Field to write content-related information of the analysis.
- **All variants' information**: Table with all selected variant's information.
- **Variant information:** Mutation identification, followed by its classification and a short description of the variant. More information is included containing HGVS notation, related gene, exon, variant type, rs ID, zygosity, frequency, and coverage.
- **Variant references:** List of scientific papers referring to a specific variant.

- **Gene information:** Disease-related information from CGD, about the gene containing a specific variant.
- **Related drugs (if applicable):** Information regarding the drug-gene interactions from DGIdb, and other data resources.
  For example:
  - **Pharmacogenomics Knowledge Base (PharmGKB)**

  PharmGKB is a pharmacogenomics knowledge resource including clinical information such as clinical guidelines, drug labels, potentially clinically actionable gene-drug associations, and genotype-phenotype relationships. Information on the impact of human genetic variation on drug responses is aggregated through various approaches which include, among others: literature review for the annotation of genetic variants and gene-drug-disease relationships, curation of FDA labels, and drug dosing guidelines in a pharmacogenomic approach.

  VarSome integrates information from PharmGKB that may concern:

  - The relationship between a variant selected from the table and corresponding medications along with related supporting publications. This information, if available, is located in tabular format in the “PharmGKB” Tab under the Variants Table. An example is shown in the following screenshot.


PharmGKB Version: 21-Mar-2023

Chemical Relations

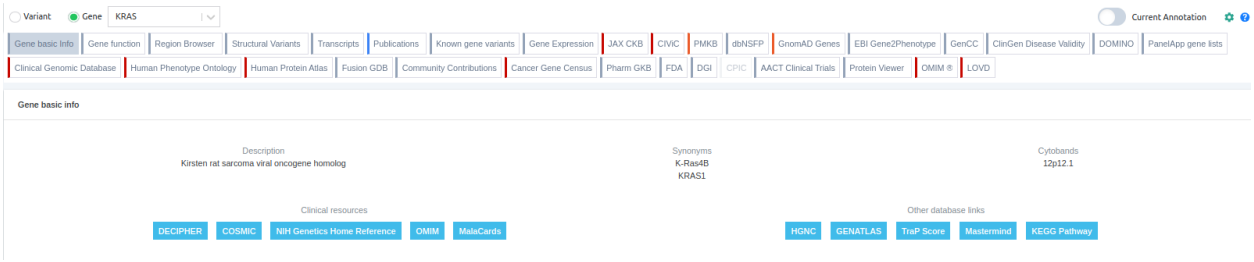
Drug	Association	Significant	Drug Variant Relation	Pharmacodynamic	Pharmacokinetic	Publications	Annotation ID	Curator notes
Cetuximab	Not associated	No	Efficacy	-	-	22734028	1447247696	Expand notes

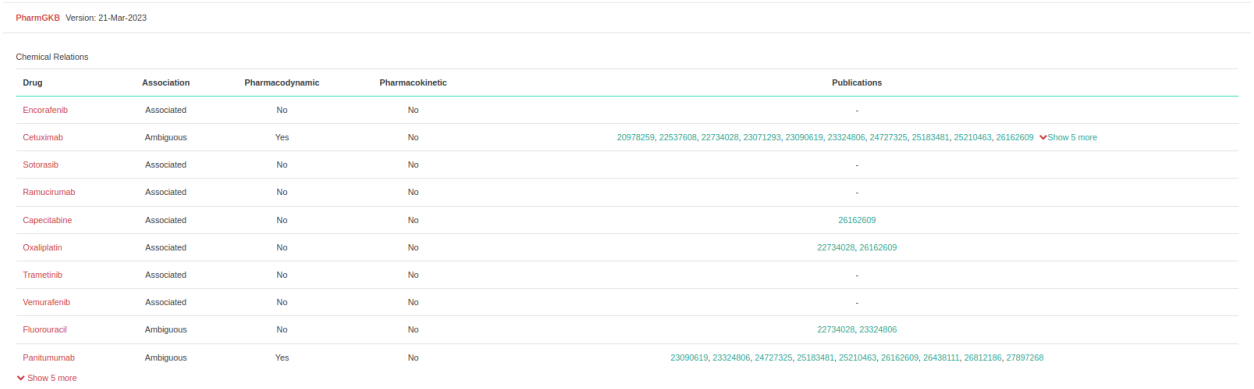
Disease Relations

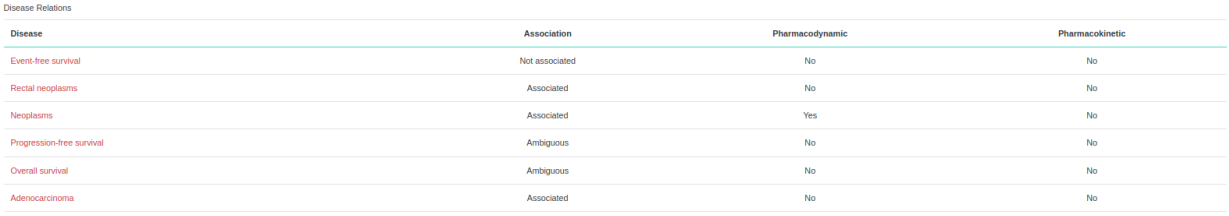
Disease	Association	Pharmacodynamic	Pharmacokinetic
Progression-free survival	Not associated	No	No
Colorectal neoplasms	Not associated	No	No

This Table holds the Clinical Annotations of the variant which can be also accessed at the PharmGKB site through the respective ID. The "Annotation" column refers to manually curated genotype-based summaries describing the phenotypic impact of the variant. The user can also see an overview of the corresponding Publication. If the corresponding data are available, the user may also see a table for chemical relations or disease relations.

- This section concerns the relations of the *gene* containing the variant with drug substances. All the information can be found under the "Gene basic info" tab on the right menu of the Variant Table page. If the corresponding information is available the user may see a table for Guidelines, Chemical relations, Drug Label Annotations and Disease Relations. The user can also see an overview of the corresponding Publications.







The gene-related tables are also available for the user during the report generation process which is described in more detail in the Final Report Generation above. The addition of the PharmGKB section in the report will generate a table similar to the following example:

BRAF related drugs from PharmaGKB

Chemical Reactions

Drug ID	Drug Name	Association	Pharmacodynamic	Pharmacokinetic	PUBMED References	Drug Label Annotations
PA166179872	encorafenib	associated	False	False		{'name': 'Annotation of EMA Label for encorafenib and BRAF', 'testing_level': 'Testing required', 'id': 'PA166182948'} {'id': 'PA166179878', 'testing_level': 'Testing required', 'biomarker_flag': 'On', 'name': 'Annotation of FDA Label for encorafenib and BRAF'}
PA166129522	nivolumab	associated	False	False		{'id': 'PA166160121', 'testing_level': 'Informative PGx', 'biomarker_flag': 'On', 'name': 'Annotation of FDA Label for nivolumab and ALK,BRAF,CD274,EGFR'}
PA165946873	vemurafenib	associated	False	False		{'id': 'PA166114482', 'testing_level': 'Testing required', 'biomarker_flag': 'Off (Never On)', 'name': 'Annotation of EMA Label for vemurafenib and BRAF'} {'id': 'PA166104858', 'testing_level': 'Testing required', 'biomarker_flag': 'On', 'name': 'Annotation of FDA Label for vemurafenib and BRAF'} {'name': 'Annotation of PMDA Label for vemurafenib and BRAF', 'testing_level': 'Testing required', 'id': 'PA166160851'} {'name': 'Annotation of HCSC Label for vemurafenib and BRAF', 'testing_level': 'Testing required', 'id': 'PA166127728'}
PA166114911	dabrafenib	associated	True	False		{'id': 'PA166119826', 'testing_level': 'Testing required', 'biomarker_flag': 'Off (Never On)', 'name': 'Annotation of EMA Label for dabrafenib and BRAF'} {'name': 'Annotation of HCSC Label for dabrafenib and BRAF', 'testing_level': 'Testing required', 'id': 'PA166127656'} {'id': 'PA166114912', 'testing_level': 'Testing required', 'biomarker_flag': 'On', 'name': 'Annotation of FDA Label for dabrafenib and BRAF'}
PA166115364	trametinib	associated	False	False		{'id': 'PA166115365', 'testing_level': 'Testing required', 'biomarker_flag': 'On', 'name': 'Annotation of FDA Label for trametinib and BRAF'} {'name': 'Annotation of HCSC Label for trametinib and BRAF', 'testing_level': 'Testing required', 'id': 'PA166127722'}

- **Precision Medicine Knowledgebase (PMKB)**

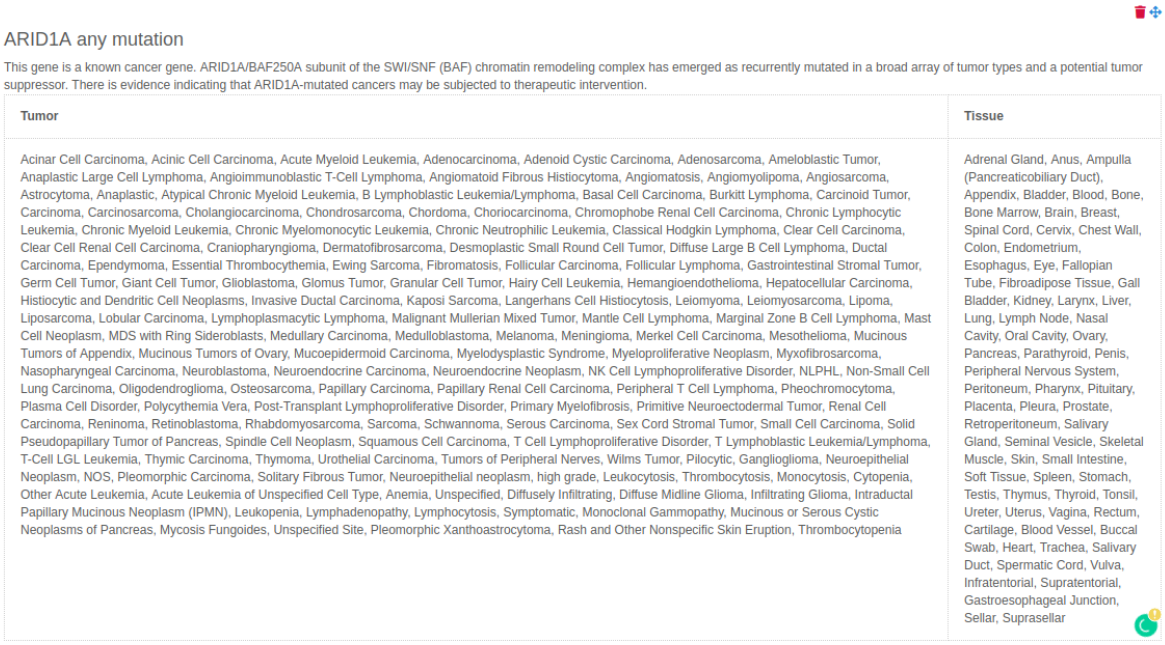
PMKB is a knowledge database that presents clinical interpretations of cancer variants in a structured way. It allows user contributions in terms of browsing, submitting, and editing existing entries. In addition, all changes are reviewed by molecular pathologists and oncologists.

VarSome presents the information retrieved from PMKB in the "Genes" Tab on the right menu of the Variant Table page:

PMKB Version: 21-Sep-2022

Tier	Tumor	Tissue	Definition	Interpretation	PUBMED references
1	Adenocarcinoma	Breast Ovary Prostate	BRCA1 any mutation	Inactivating mutations in BRCA1 may be associated with sensitivity to PARP inhibitors. Drug Rucaparib Niraparib Olaparib	-

The aforementioned information is also available for the user during the report generation process which is described in more detail in section 15. Final Report Generation. The addition of the PMKB section in the Report will generate a table similar to the following example:



ARID1A any mutation

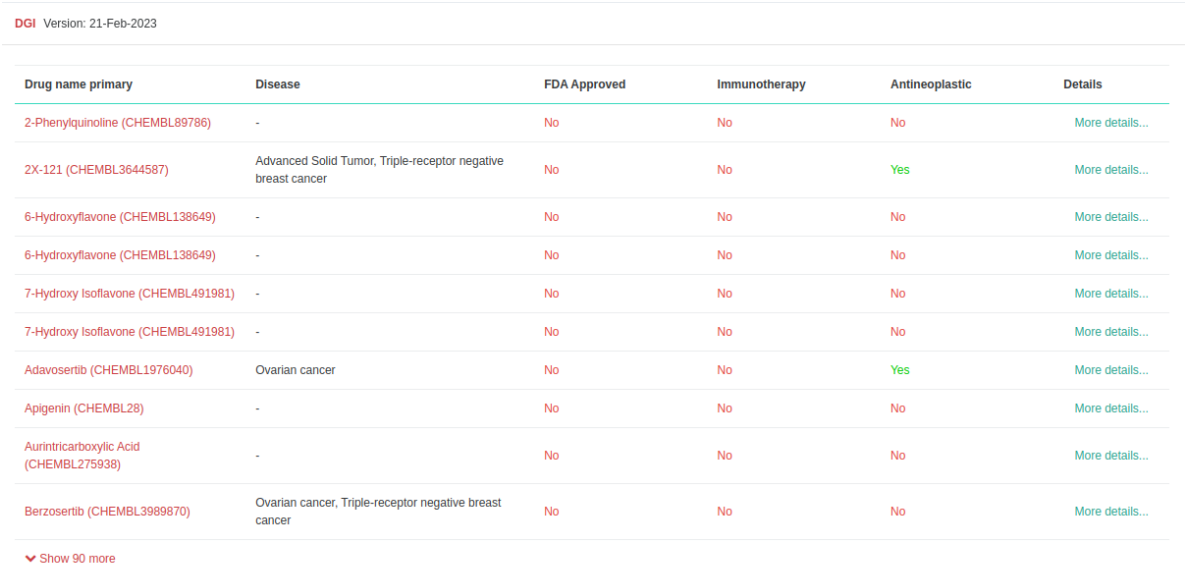
This gene is a known cancer gene. ARID1A/BAF250A subunit of the SWI/SNF (BAF) chromatin remodeling complex has emerged as recurrently mutated in a broad array of tumor types and a potential tumor suppressor. There is evidence indicating that ARID1A-mutated cancers may be subjected to therapeutic intervention.

Tumor	Tissue
Acinar Cell Carcinoma, Acinic Cell Carcinoma, Acute Myeloid Leukemia, Adenocarcinoma, Adenoid Cystic Carcinoma, Adenosarcoma, Ameloblastic Tumor, Anaplastic Large Cell Lymphoma, Angioimmunoblastic T-Cell Lymphoma, Angiomatoid Fibrous Histiocytoma, Angiomatosis, Angiomyolipoma, Angiosarcoma, Astrocytoma, Anaplastic, Atypical Chronic Myeloid Leukemia, B Lymphoblastic Leukemia/Lymphoma, Basal Cell Carcinoma, Burkitt Lymphoma, Carcinoid Tumor, Carcinoma, Carcinosarcoma, Cholangiocarcinoma, Chondrosarcoma, Chordoma, Choriocarcinoma, Chromophobe Renal Cell Carcinoma, Chronic Lymphocytic Leukemia, Chronic Myeloid Leukemia, Chronic Myelomonocytic Leukemia, Chronic Neutrophilic Leukemia, Classical Hodgkin Lymphoma, Clear Cell Carcinoma, Clear Cell Renal Cell Carcinoma, Craniopharyngioma, Dermatofibrosarcoma, Desmoplastic Small Round Cell Tumor, Diffuse Large B Cell Lymphoma, Ductal Carcinoma, Ependymoma, Essential Thrombocythemia, Ewing Sarcoma, Fibromatosis, Follicular Carcinoma, Follicular Lymphoma, Gastrointestinal Stromal Tumor, Germ Cell Tumor, Giant Cell Tumor, Glioblastoma, Glomus Tumor, Granular Cell Tumor, Hairy Cell Leukemia, Hemangioendothelioma, Hepatocellular Carcinoma, Histiocytic and Dendritic Cell Neoplasms, Invasive Ductal Carcinoma, Kaposi Sarcoma, Langerhans Cell Histiocytosis, Leiomyoma, Leiomyosarcoma, Lipoma, Liposarcoma, Lobular Carcinoma, Lymphoplasmacytic Lymphoma, Malignant Mullerian Mixed Tumor, Mantle Cell Lymphoma, Marginal Zone B Cell Lymphoma, Mast Cell Neoplasm, MDS with Ring Sideroblasts, Medullary Carcinoma, Medulloblastoma, Melanoma, Meningioma, Merkel Cell Carcinoma, Mesothelioma, Mucinous Tumors of Appendix, Mucinous Tumors of Ovary, Mucoepidermoid Carcinoma, Myelodysplastic Syndrome, Myeloproliferative Neoplasm, Myxofibrosarcoma, Nasopharyngeal Carcinoma, Neuroblastoma, Neuroendocrine Carcinoma, Neuroendocrine Neoplasm, NK Cell Lymphoproliferative Disorder, NLPHL, Non-Small Cell Lung Carcinoma, Oligodendroglioma, Osteosarcoma, Papillary Carcinoma, Papillary Renal Cell Carcinoma, Peripheral T Cell Lymphoma, Pheochromocytoma, Plasma Cell Disorder, Polycythemia Vera, Post-Transplant Lymphoproliferative Disorder, Primary Myelofibrosis, Primitive Neuroectodermal Tumor, Renal Cell Carcinoma, Reninoma, Retinoblastoma, Rhabdomyosarcoma, Sarcoma, Schwannoma, Serous Carcinoma, Sex Cord Stromal Tumor, Small Cell Carcinoma, Solid Pseudopapillary Tumor of Pancreas, Spindle Cell Neoplasm, Squamous Cell Carcinoma, T Cell Lymphoproliferative Disorder, T Lymphoblastic Leukemia/Lymphoma, T-Cell LGL Leukemia, Thymic Carcinoma, Thymoma, Urothelial Carcinoma, Tumors of Peripheral Nerves, Wilms Tumor, Pilocytic, Ganglioglioma, Neuroepithelial Neoplasm, NOS, Pleomorphic Carcinoma, Solitary Fibrous Tumor, Neuroepithelial neoplasm, high grade, Leukocytosis, Thrombocytosis, Monocytosis, Cytopenia, Other Acute Leukemia, Acute Leukemia of Unspecified Cell Type, Anemia, Unspecified, Diffusely Infiltrating, Diffuse Midline Glioma, Infiltrating Glioma, Intraductal Papillary Mucinous Neoplasm (IPMN), Leukopenia, Lymphadenopathy, Lymphocytosis, Symptomatic, Monoclonal Gammopathy, Mucinous or Serous Cystic Neoplasms of Pancreas, Mycosis Fungoides, Unspecified Site, Pleomorphic Xanthoastrocytoma, Rash and Other Nonspecific Skin Eruption, Thrombocytopenia	Adrenal Gland, Anus, Ampulla (Pancreaticobiliary Duct), Appendix, Bladder, Blood, Bone, Bone Marrow, Brain, Breast, Spinal Cord, Cervix, Chest Wall, Colon, Endometrium, Esophagus, Eye, Fallopian Tube, Fibroadipose Tissue, Gall Bladder, Kidney, Larynx, Liver, Lung, Lymph Node, Nasal Cavity, Oral Cavity, Ovary, Pancreas, Parathyroid, Penis, Peripheral Nervous System, Peritoneum, Pharynx, Pituitary, Placenta, Pleura, Prostate, Retroperitoneum, Salivary Gland, Seminal Vesicle, Skeletal Muscle, Skin, Small Intestine, Soft Tissue, Spleen, Stomach, Testis, Thymus, Thyroid, Tonsil, Ureter, Uterus, Vagina, Rectum, Cartilage, Blood Vessel, Buccal Swab, Heart, Trachea, Salivary Duct, Spermatic Cord, Vulva, Infratentorial, Supratentorial, Gastroesophageal Junction, Sellar, Suprasellar

- **Drug-Gene Interaction Database (DGIdb)**

DGIdb provides information about known or potential drug-gene interactions using expert curation and text-mining methods. The presented drug-gene interactions have been mined from DrugBank, therapeutic target database (TTD), PharmGKB, a list of targeted agents in lung cancer, and ClinicalTrials.gov. In addition, retrieved data from Gene Ontology, dGene, and “druggable genome” lists from Hopkins and Groom (2002) and Russ and Lampel (2005) are utilized for the categorization of the genes as potentially druggable.

VarSome presents the information retrieved from DGIdb at the bottom of the "Genes" Tab on the right menu of the Variant Table page:



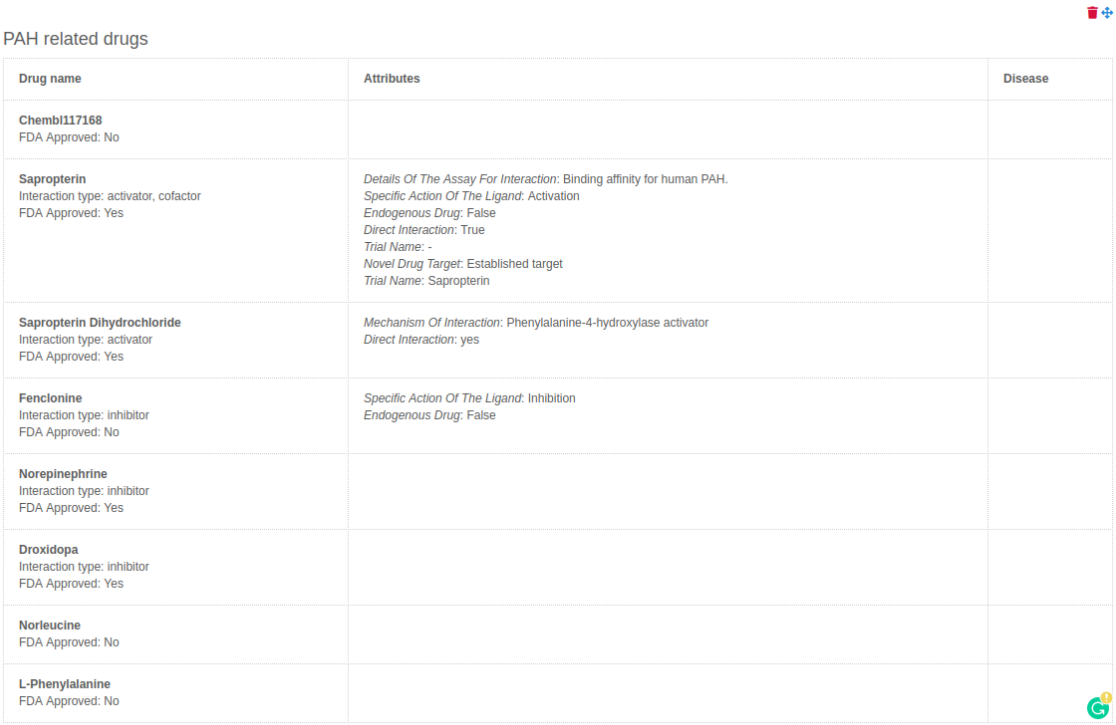
DGI Version: 21-Feb-2023

Drug name primary	Disease	FDA Approved	Immunotherapy	Antineoplastic	Details
2-Phenylquinoline (CHEMBL89786)	-	No	No	No	More details...
2X-121 (CHEMBL3644587)	Advanced Solid Tumor, Triple-receptor negative breast cancer	No	No	Yes	More details...
6-Hydroxyflavone (CHEMBL138649)	-	No	No	No	More details...
6-Hydroxyflavone (CHEMBL138649)	-	No	No	No	More details...
7-Hydroxy Isoflavone (CHEMBL491981)	-	No	No	No	More details...
7-Hydroxy Isoflavone (CHEMBL491981)	-	No	No	No	More details...
Adavosertib (CHEMBL1976040)	Ovarian cancer	No	No	Yes	More details...
Apigenin (CHEMBL28)	-	No	No	No	More details...
Aurintricarboxylic Acid (CHEMBL275938)	-	No	No	No	More details...
Berzosertib (CHEMBL3989870)	Ovarian cancer, Triple-receptor negative breast cancer	No	No	No	More details...

Show 90 more

VarSome Clinical retrieves information from DGIdb and makes it available in the Report. By

adding the DGIdb section to your Report, the information will be shown in tabular form, as below. This table holds aggregated information from DGIdb on the interaction of the respective gene with different drug substances and the various attributes of the interaction.

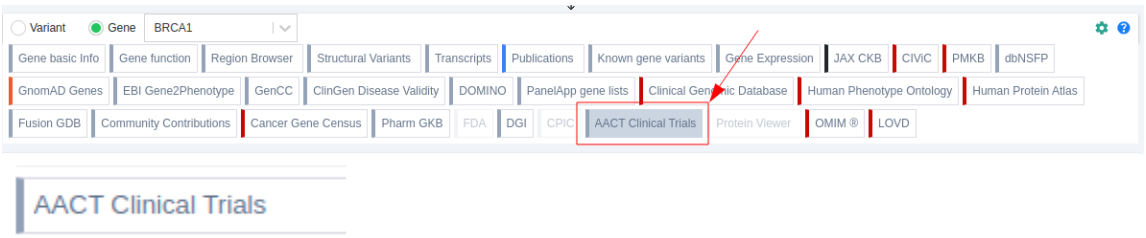


PAH related drugs

Drug name	Attributes	Disease
Chembl117168 FDA Approved: No		
Sapropterin Interaction type: activator, cofactor FDA Approved: Yes	*Details Of The Assay For Interaction*: Binding affinity for human PAH. *Specific Action Of The Ligand*: Activation *Endogenous Drug*: False *Direct Interaction*: True *Trial Name*: - *Novel Drug Target*: Established target *Trial Name*: Sapropterin	
Sapropterin Dihydrochloride Interaction type: activator FDA Approved: Yes	*Mechanism Of Interaction*: Phenylalanine-4-hydroxylase activator *Direct Interaction*: yes	
Fenclonine Interaction type: inhibitor FDA Approved: No	*Specific Action Of The Ligand*: Inhibition *Endogenous Drug*: False	
Norepinephrine Interaction type: inhibitor FDA Approved: Yes		
Droxidopa Interaction type: inhibitor FDA Approved: Yes		
Norleucine FDA Approved: No		
L-Phenylalanine FDA Approved: No		

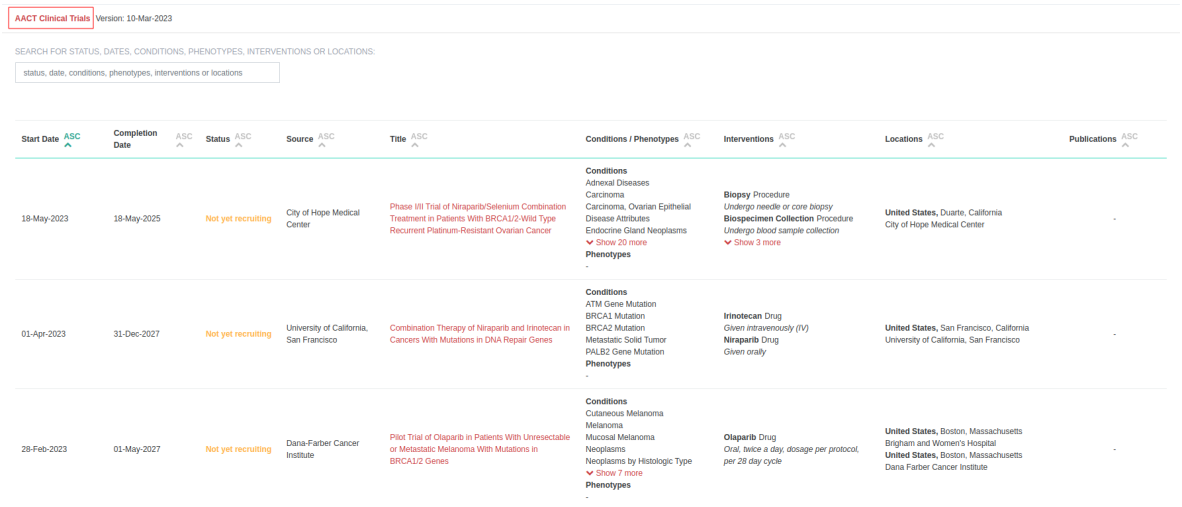
- **PMKB info (if applicable):** Information about the gene containing a specific variant, retrieved from PMKB database.
- **CIVIC info (if applicable):** Information about mutations shown to predict response to targeted therapy from CIViC database.
- **GHR info (if applicable):** gene information from GHR.
- **Clinical Trials info (if applicable):** Information about related clinical studies from AACT database. The AACT Database is a publicly available relational database and its purpose is to include all information (protocol and result data elements) about every study registered in ClinicalTrials.gov. VarSome Clinical retrieves information from AACT concerning clinical trials related to the gene that contains the selected variant. The information, if available, is displayed under the "Gene" option on the bottom of the variant table in the Variant Table results page, as shown in the screenshot below.

The card:



opens a table that holds an overview (protocol title, trial status, research sites, supporting

references, etc.) of the clinical trials related to the gene, providing to the user direct and complete insight into the interest around the gene in the clinical research field.



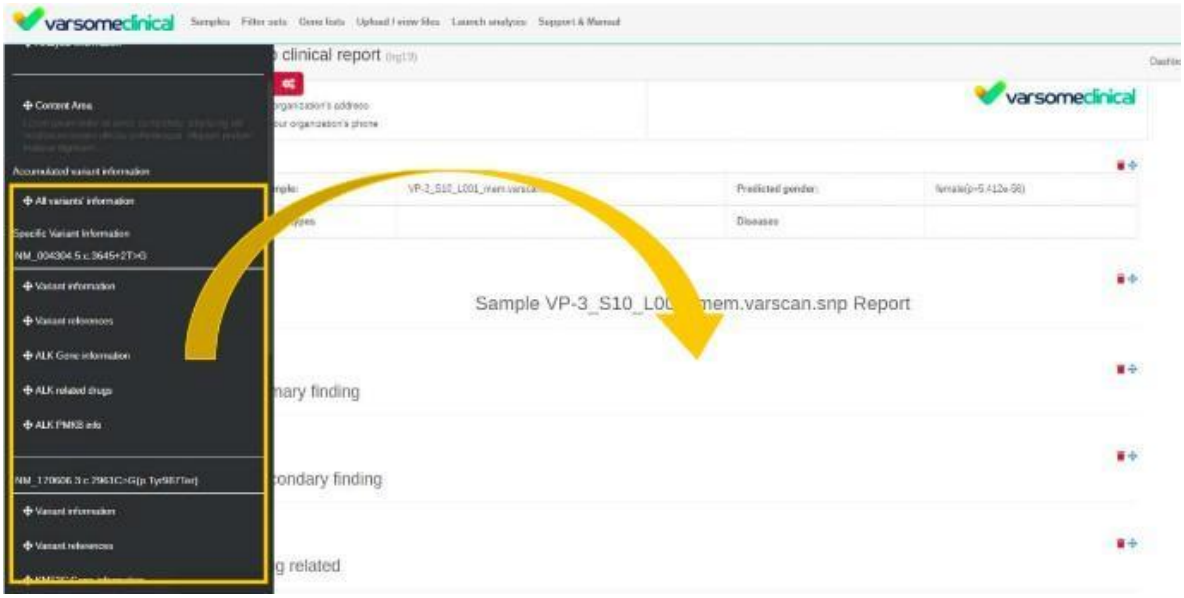
AACT Clinical Trials Version: 10-Mar-2023

SEARCH FOR STATUS, DATES, CONDITIONS, PHENOTYPES, INTERVENTIONS OR LOCATIONS:

status, date, conditions, phenotypes, interventions or locations

Start Date ASC	Completion Date ASC	Status ASC	Source ASC	Title ASC	Conditions / Phenotypes ASC	Interventions ASC	Locations ASC	Publications ASC
18-May-2023	18-May-2025	Not yet recruiting	City of Hope Medical Center	Phase I/II Trial of Niraparib/Selenium Combination Treatment in Patients With BRCA1/2-Wild Type Recurrent Platinum-Resistant Ovarian Cancer	**Conditions** Adnexal Diseases Carcinoma Carcinoma, Ovarian Epithelial Disease Attributes Endocrine Gland Neoplasms Show 20 more **Phenotypes** -	**Biopsy** Procedure *Undergo needle or core biopsy* **Biospecimen Collection** Procedure *Undergo blood sample collection* Show 3 more	**United States,** Duarte, California City of Hope Medical Center	-
01-Apr-2023	31-Dec-2027	Not yet recruiting	University of California, San Francisco	Combination Therapy of Niraparib and Irinotecan in Cancers With Mutations in DNA Repair Genes	**Conditions** ATM Gene Mutation BRCA1 Mutation BRCA2 Mutation Metastatic Solid Tumor PALB2 Gene Mutation **Phenotypes** -	**Irinotecan** Drug *Given intravenously (IV)* **Niraparib** Drug *Given orally*	**United States,** San Francisco, California University of California, San Francisco	-
28-Feb-2023	01-May-2027	Not yet recruiting	Dana-Farber Cancer Institute	Pilot Trial of Olaparib in Patients With Unresectable or Metastatic Melanoma With Mutations in BRCA1/2 Genes	**Conditions** Cutaneous Melanoma Melanoma Mucosal Melanoma Neoplasms Neoplasms by Histologic Type Show 7 more **Phenotypes** -	**Olaparib** Drug *Oral, twice a day, dosage per protocol, per 28 day cycle*	**United States,** Boston, Massachusetts Brigham and Women's Hospital **United States,** Boston, Massachusetts Dana Farber Cancer Institute	-

This information is also available for integration into the Report.



11. Visualization

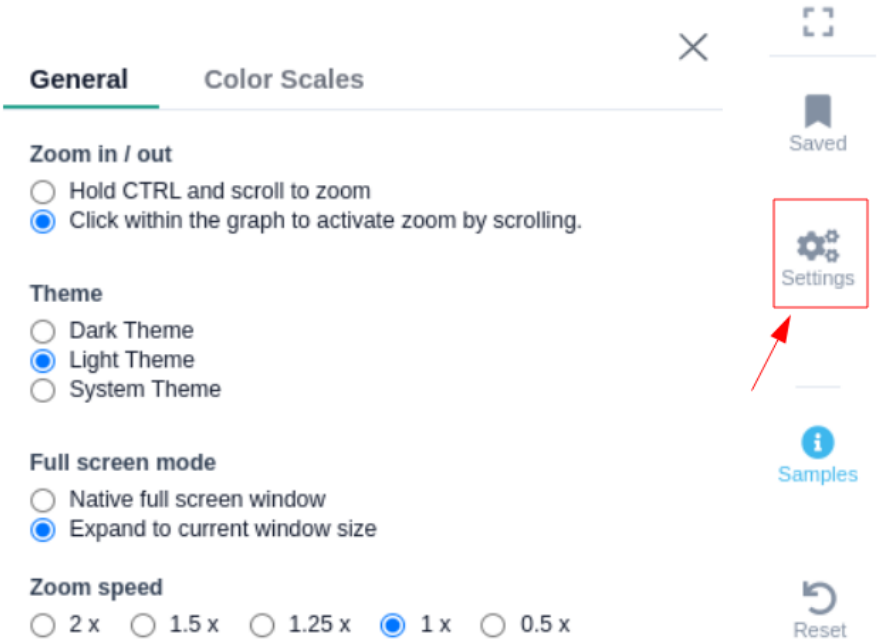
11.1 Browsers

VarSome Clinical offers different browsers to visualize the data. These are the Region Browser, CNV Browser and Sample View Browser. Each browser provides different information. The region browser displays the variant in the context of its genomic "neighborhood", including transcripts and other, nearby variants and their pathogenicity as well as frequencies from GnomAD.
The CNV Browser provides an interactive graph to visualize the CNV call region in all samples of the cohort.
Finally, the Sample View presents a genome region browser, but one that is focused on displaying the variants (SNPs, indels, CNVs) identified in the current sample. This provides a global overview of the variants identified in the current sample.

11.1.1 Browser interface settings

There are four icons on the right hand side which can be used to maximize the browser, save the current position on the browser to use in any of the 3 browsers, and change various display settings as shown in the image below. The reset button will bring the browser back to the original position in case the browser has been moved.

General settings:



Theme: You can select the theme of the genome browser component. Note that by selecting

the system theme, the theme is adjusted to the one you have selected from the setting background option on your browser.

Gesture control options:

- Hold ctrl and scroll to zoom : Use CTRL + SCROLL to zoom in/out
- Click to activate scroll zoom : Use SCROLL to zoom in/out

Full screen mode: You have the option to select the size of the genome browser window.

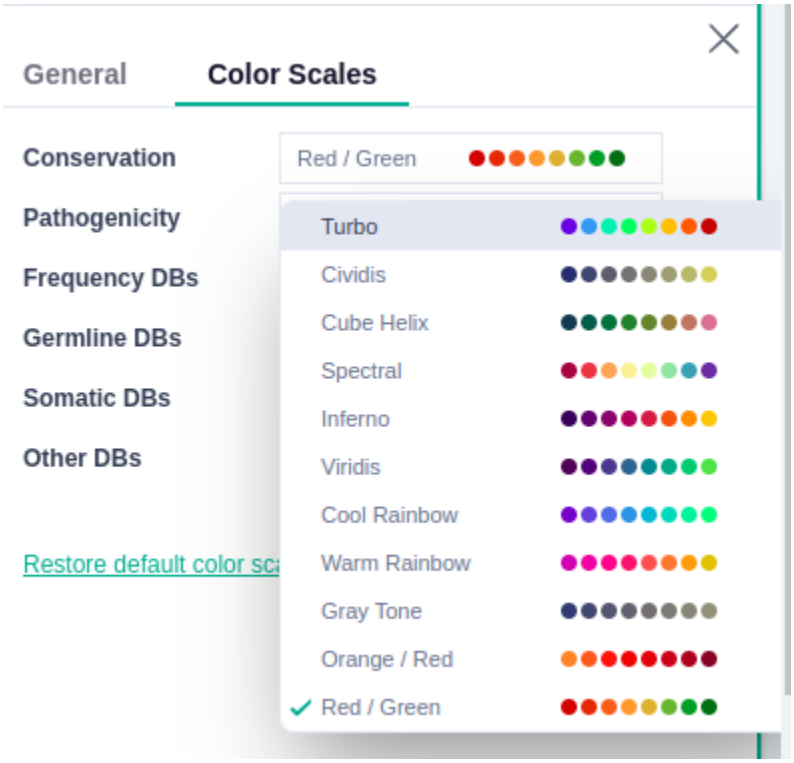
- Native full screen window
- Expand to current window size

Zoom speed: The option to adjust the speed while zooming in/out inside the genome browser window (default = 1x).

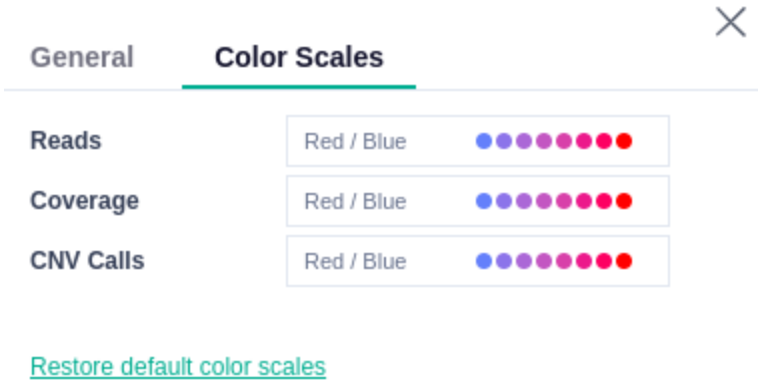
Color scales:

You can choose the colors/color scales for every track in the different browsers, e.g. Conservation and Pathogenicity on the genome region browser and sample view. Click on the drop-down menu of the track that you want to change, and select one of the predefined palette options.

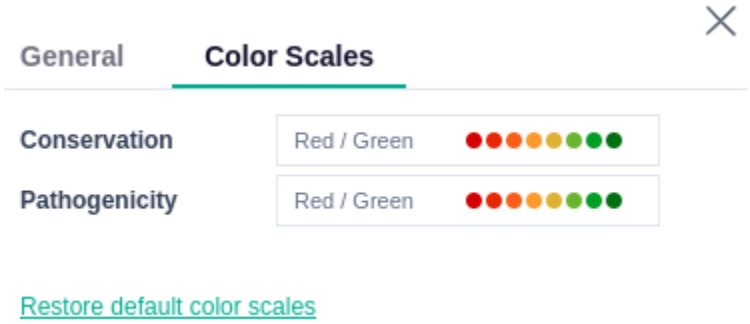
Genome Browser color scales:



CNV Browser color scales:

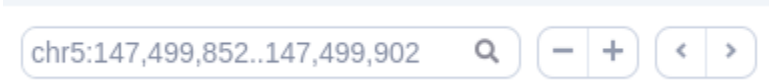


Sample View color scales:



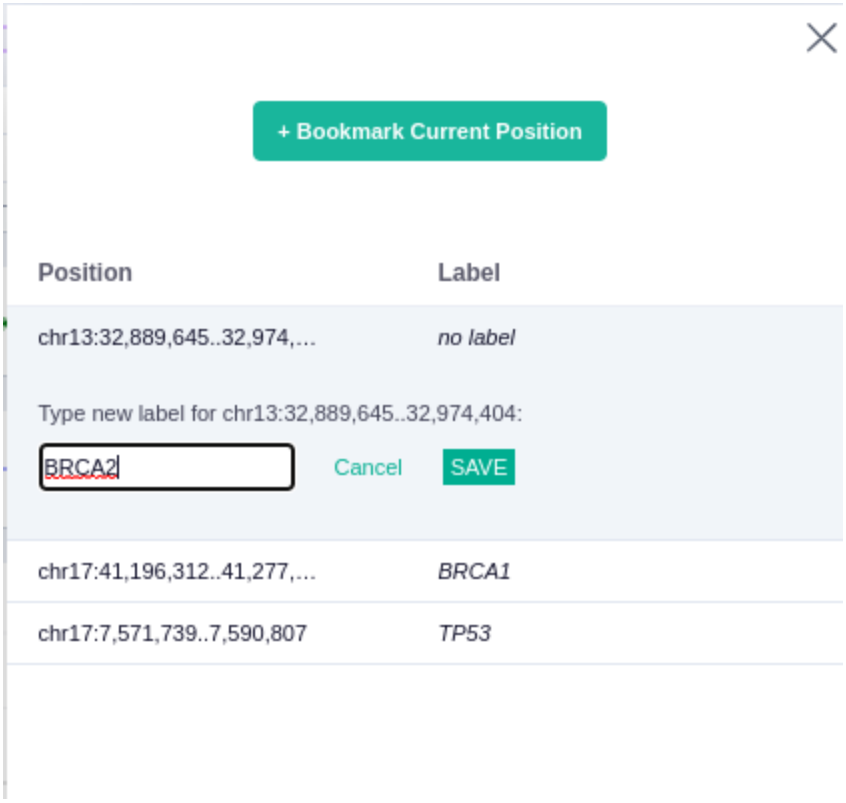
Search for a position or a gene

You can use the search box on the top right of the browser to search for a position in the genome (e.g. chr1) or for a specific gene (e.g. BRCA1). To search for a gene, you need to delete the genomic location displayed and type in the name of your gene of interest.



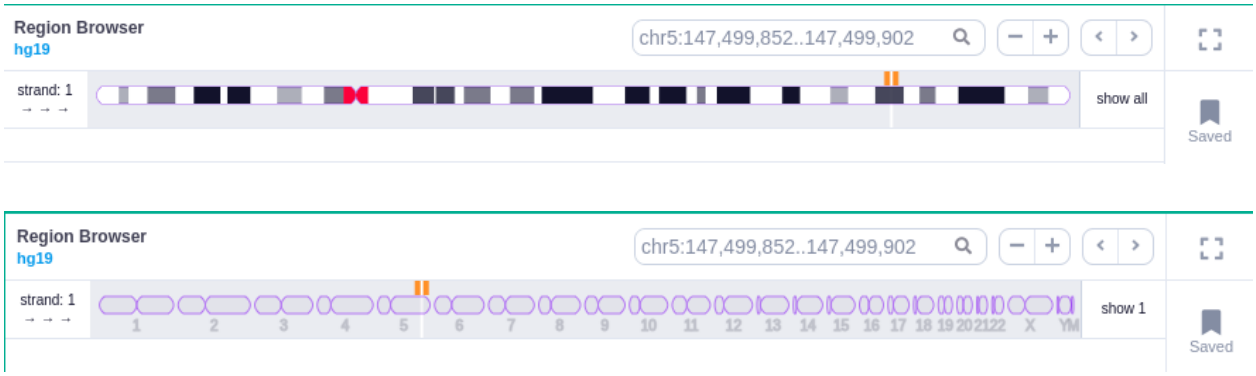
The saved icon 

on the right can be used to bookmark the current position, providing a label that can be used later.



Chromosome track

In the chromosome track, the user can switch strands by clicking on the "STRAND" button. STRAND 1 (default) refers to the positive strand (5'-3' direction), while STRAND -1 to the negative one (3'-5' direction). Additionally, the browser offers the option to expand the chromosome-level visualization from the current chromosome to all chromosomes (and *vice versa*), by clicking on the "show all" button on the right-hand side of the bar. Each chromosome can be selected for visualization to inspect the identified variants.



Zoom in and out area

It is possible to zoom in and out on the different tracks of the browser. If the area is marked with a green rectangle this means the focus is on that area and it is possible to use the mouse scroll

to zoom in and out. Otherwise no green rectangle is displayed and there is a need to use Ctrl + scroll to zoom in and out.

Green rectangle

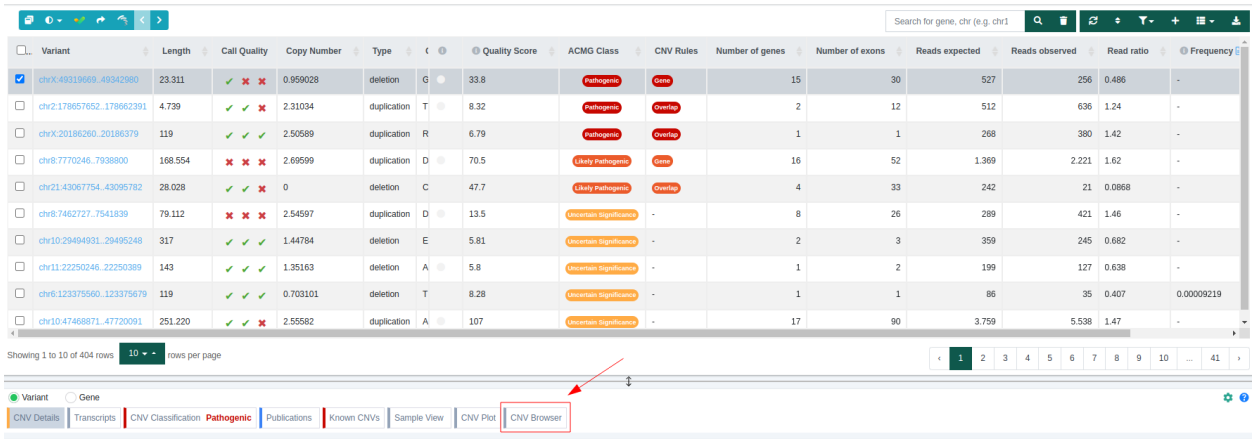
Without green rectangle

11.2 CNV Visualizations

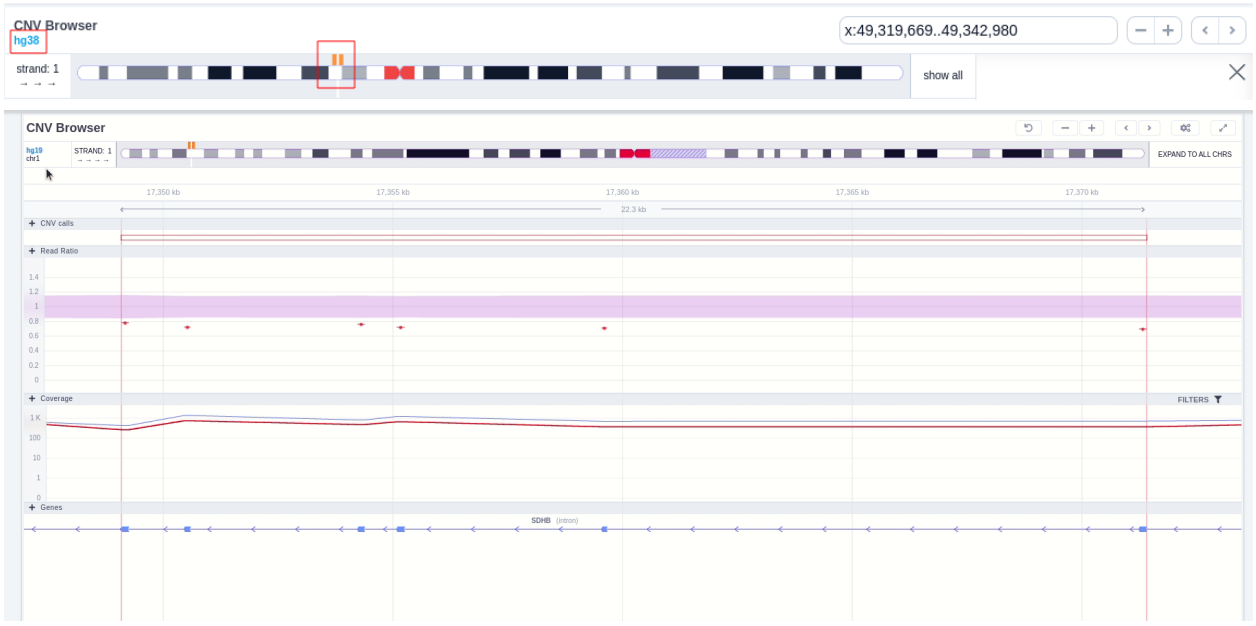
11.2.1 CNV Browser

The CNV Browser can be found under the variant table. It provides an interactive graph to

visualize the CNV call region in all samples of the cohort.

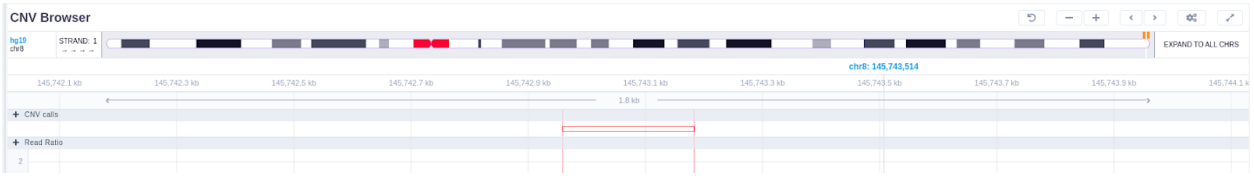


The position of the CNV call in the reference genome is marked at the top of the CNV Browser and the chromosome name is shown on the left. The user can change the size of the region shown by dragging:

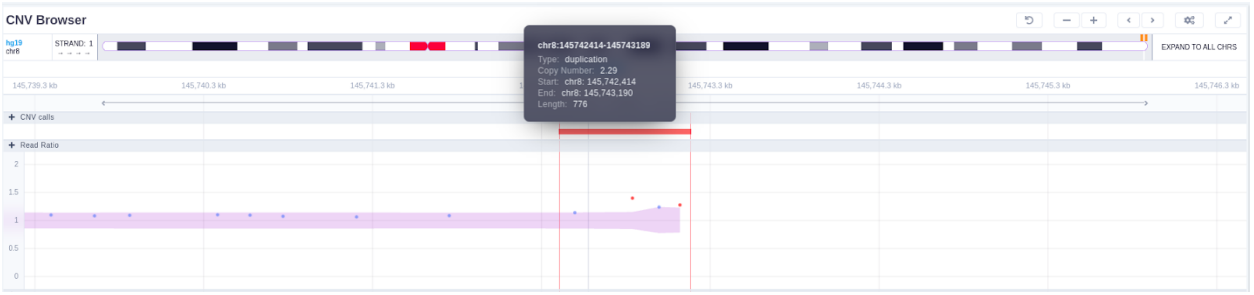


The CNV calls track shows if it is a deletion or a duplication. Deletions are represented by an empty rectangle and duplications by a full one. The position in the reference genome and the length of the CNV are shown when the user hovers over the browser with the mouse on the rectangle.

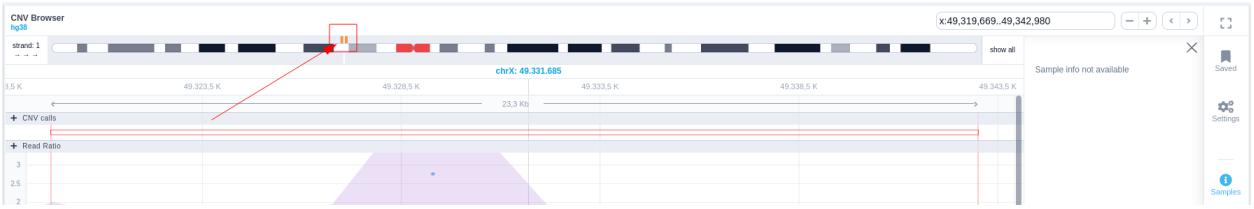
Deletion



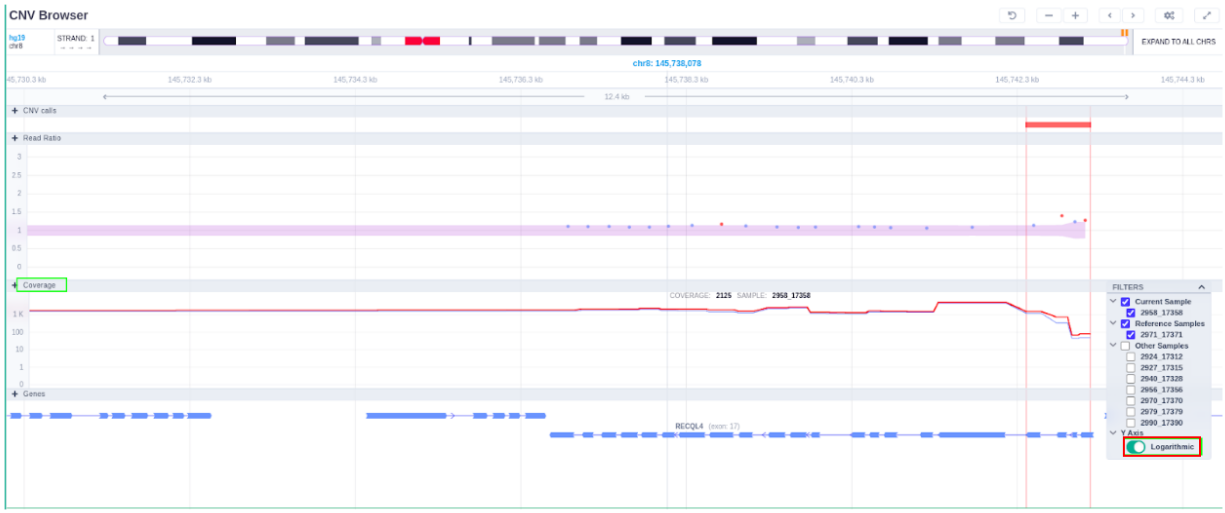
Duplication



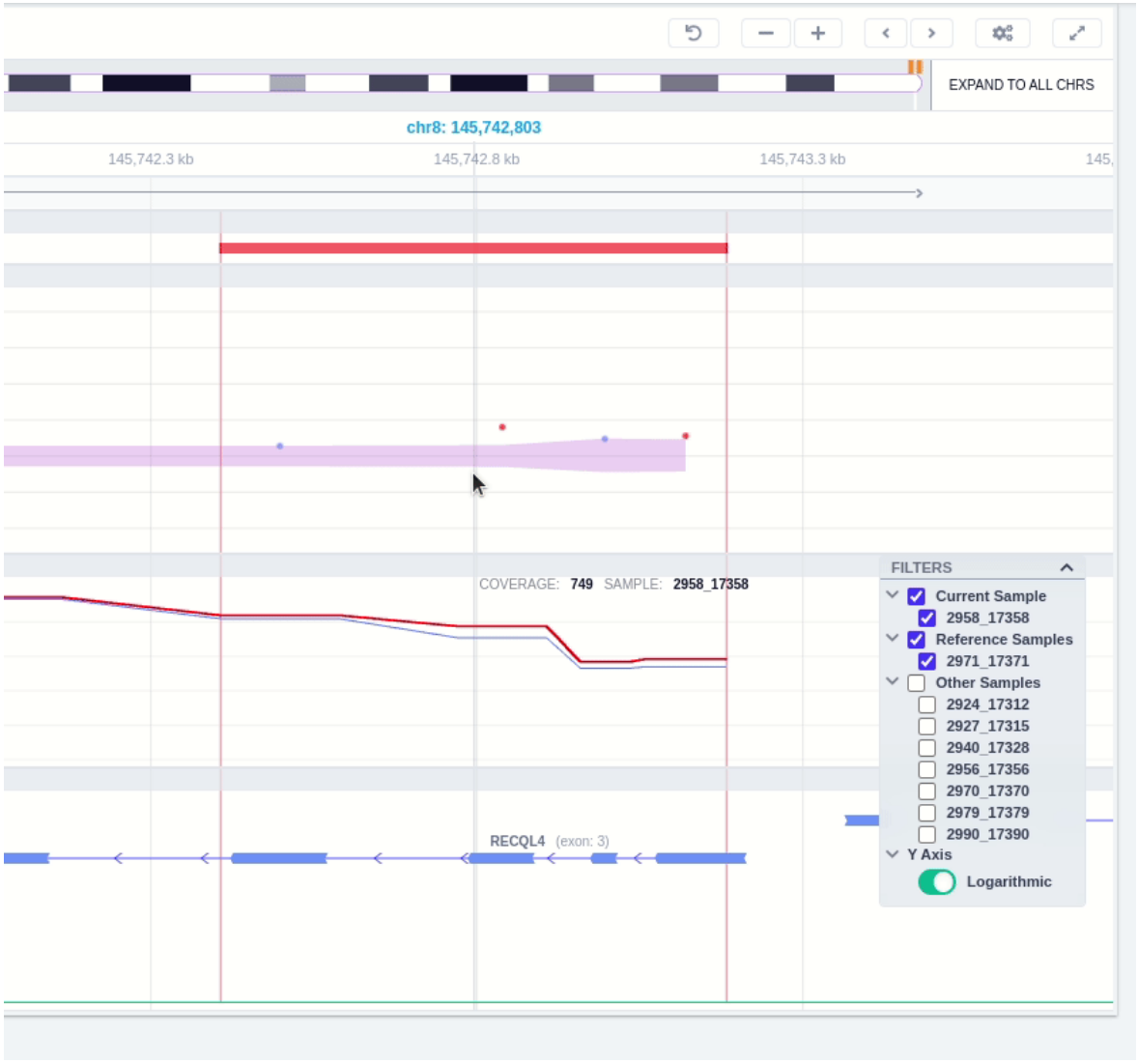
The region of the CNV call is highlighted in orange. The purple area shows the 95% confidence interval of the read ratio (observed/expected reads). Each colored rectangle represents one of the target regions of the assay used to sequence the sample, and its position on the vertical axis indicates the read ratio for this region. Target regions whose read ratio falls within the confidence interval will be colored blue, while those whose ratio falls outside the expected range will be red. The observed and expected read ratio is shown when the user hovers the mouse over the target region.



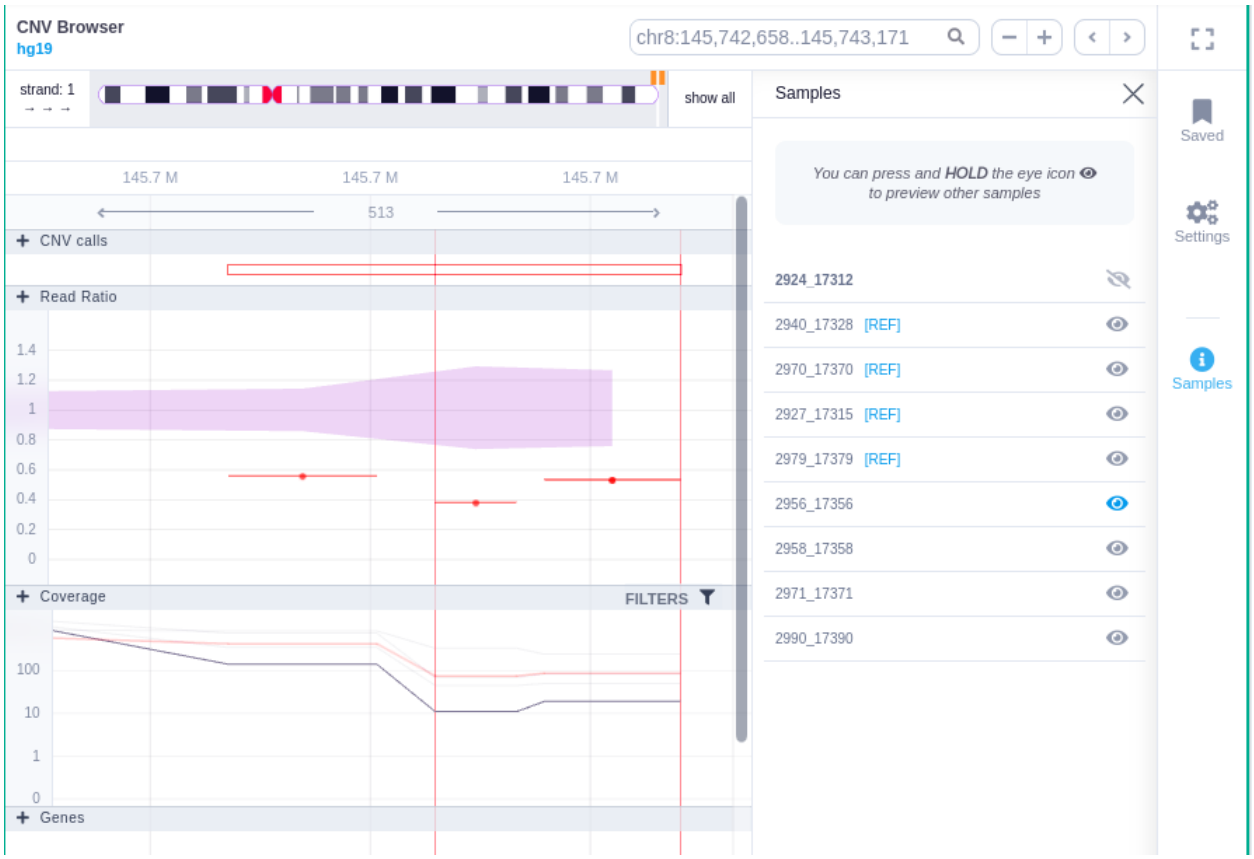
Below the read ratio is the coverage track which shows the depth of coverage across both the test sample (the sample currently under analysis) and the control samples used in this analysis. The test sample's coverage is shown in red and the others are in blue. The user can choose to view the coverage on a logarithmic or linear scale under the Filters panel.



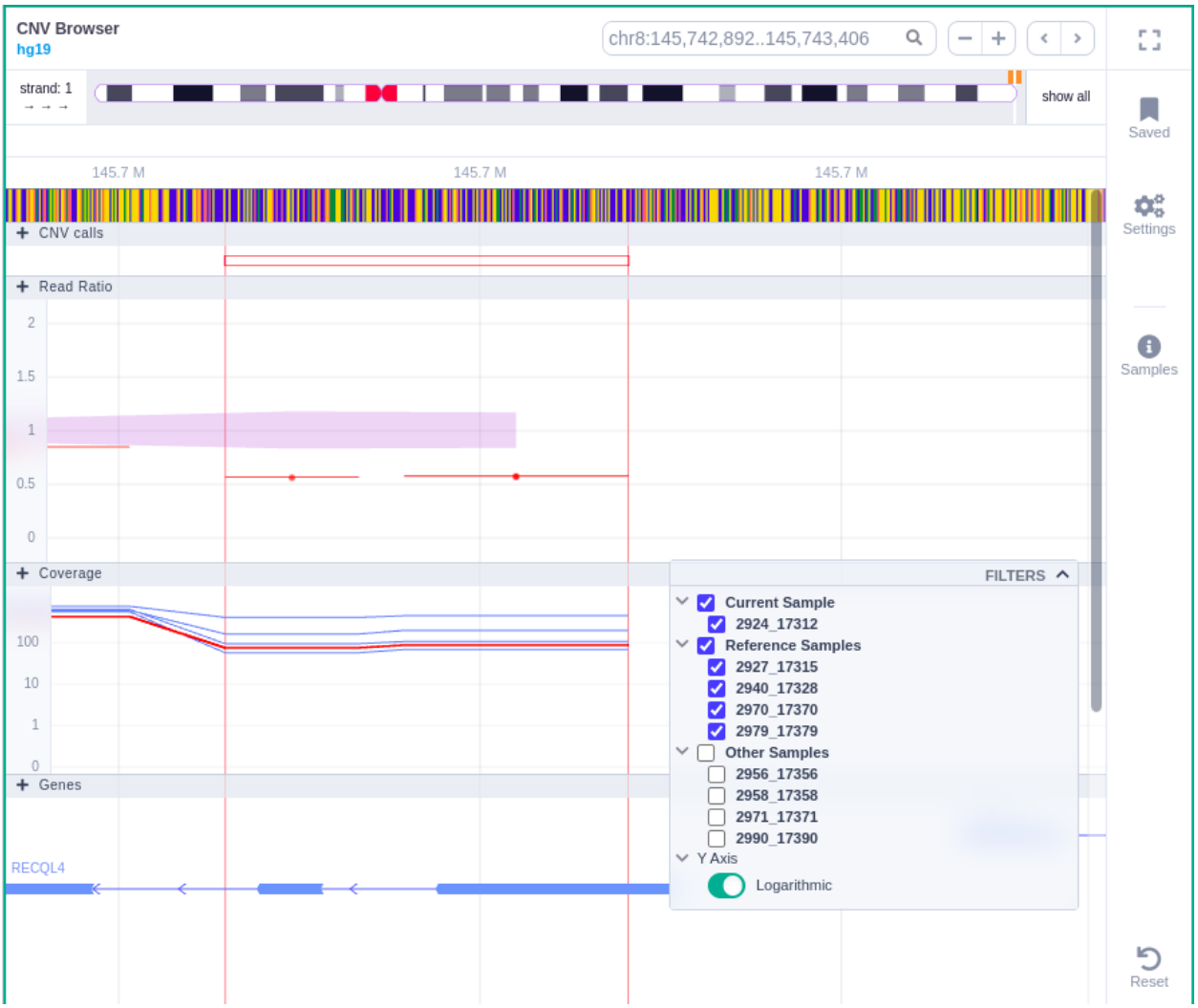
The user can hover along the CNV and see the coverage for the selected sample, the transcript and its exons under the read ratio track.



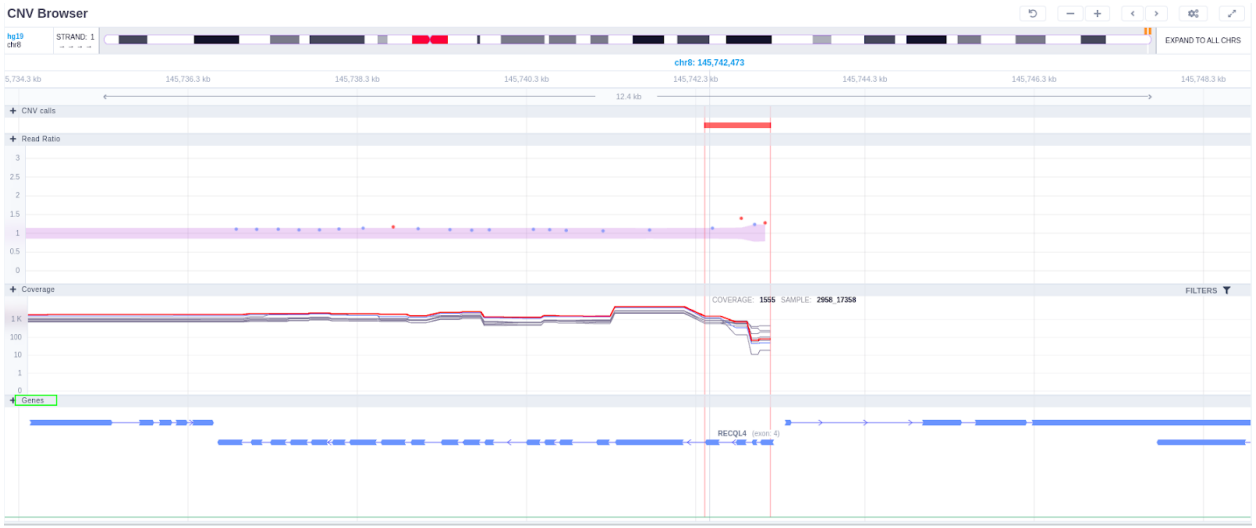
Clicking on the Samples icon on the right shows a new window where the user can choose to visualize the coverage of the other samples of the cohort by clicking and holding on the eye icon.



The filters panel allows the user to visualize the CNV call region in the rest of the samples of the CNV analysis by checking the box of the sample name. Once checked, hover over the sample names on the filters and only these will be shown in the Coverage track.

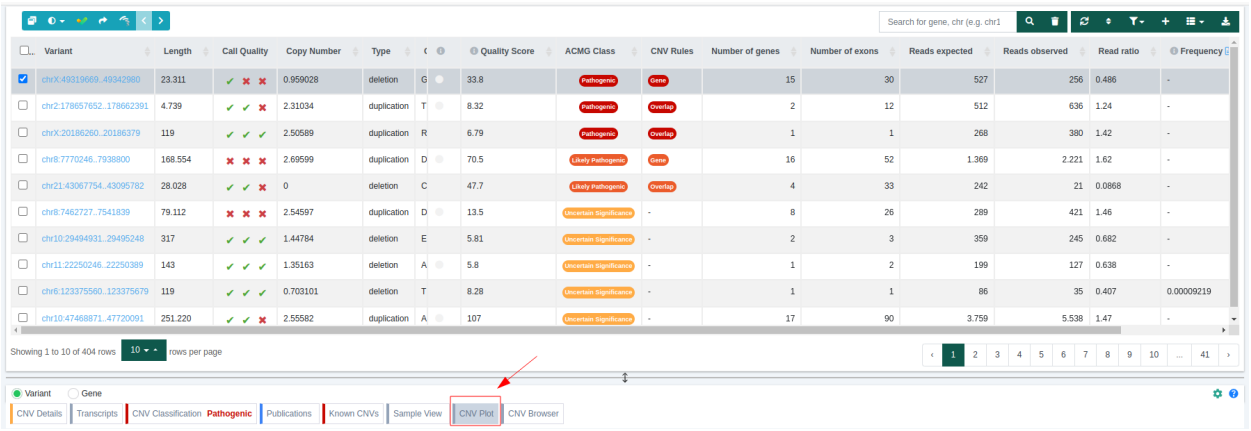


Finally, the exonic structure of any genes the CNV overlaps with is shown in the Genes track under the coverage.



11.2.2 CNV Plot

For CNV cohort analyses, VarSome Clinical provides a CNV plot, showing how the observed read depth in the area of the CNV differs from the expected. The plot can be found on its own tab, under the Variant Table.



The CNV plots are generated using a modified version of the ExomeDepth tool.

- The gray area indicates the 95% confidence interval of the observed/expected read ratio and the red crosses are specific read depth values at those positions. The genomic location of the CNV is given by the vertical dotted lines.
- The left Y-axis shows the "Observed vs expected read ratio" and the right Y-axis represents the "Copy number". The X-axis shows the chromosome coordinates.
- **For short CNVs** that encompass a few exons, the plot is displayed in a gene-centered view where the exon numbers and their position along the gene are represented in a horizontal axis above the gene name.
- **For large CNVs**, ⚠️ please note that, since the CNV region might encompass several genes, making it impossible to plot all of them, we only show the position of the canonical transcript for each gene. Therefore, if a gene has no canonical transcript (e.g. annotated pseudogene) or if its canonical transcript doesn't overlap with the CNV, you may see genes listed in the Variant Table that are not shown in the CNV plot since their canonical transcript isn't the one that overlaps with the CNV.

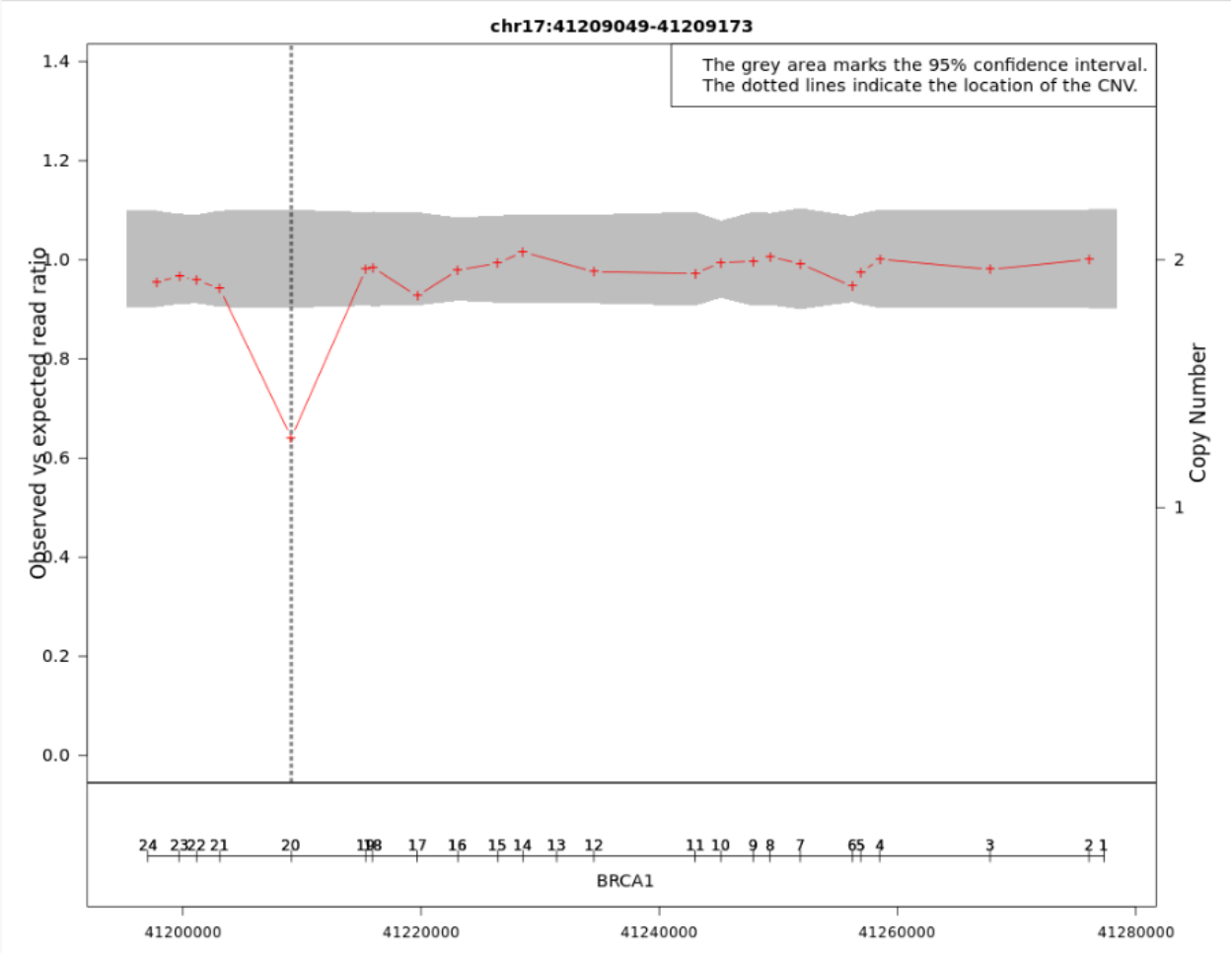
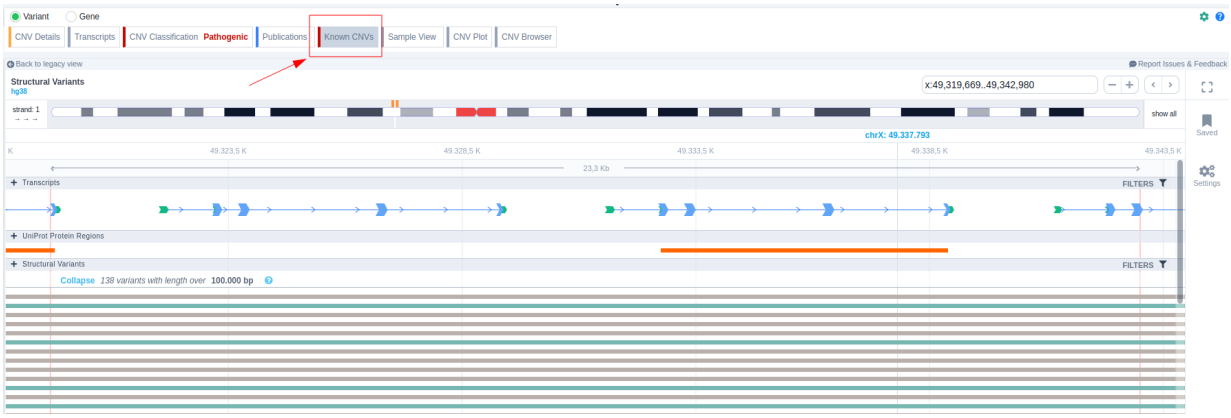


Figure: In this example, the observed to expected read depth ratio in the region of the CNV is lower than 1 and below the 95% confidence interval (grey area). These results support the hypothesis of a deletion in the exon 20 of the BRCA1 gene.

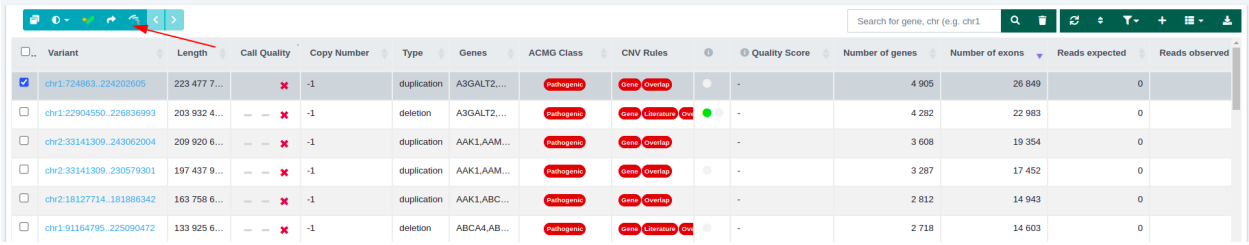
11.2.3 Known CNVs

We display only the relevant CNVs for the classification according to the following criteria:

- CNV deletions: we retain those that fully overlap with the given CNV for gnomAD variants. For CNVs coming from clinical sources (Decipher, DBVar, ClinVar CNVs) we use the overlapping CNVs if they are benign and the contained CNVs if they are pathogenic.
- CNV duplications: we keep only the CNVs encompassing the same coding genes. If the CNV is non-coding, then we retain the CNVs that have at least 85% of overlap.



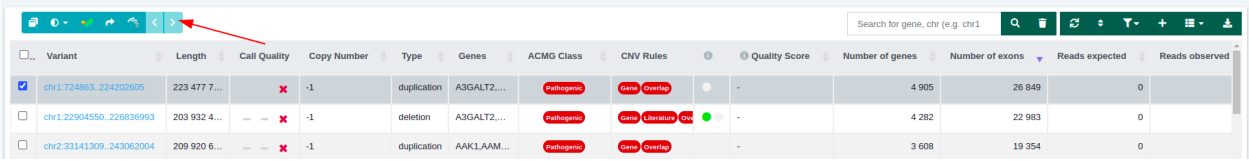
11.2.4 Read alignment visualization in JBrowse



Once you have selected a variant in the variant table, you can click on the JBrowse icon to view the alignment of the reads in the regions of the detected CNVs. The CNV called region is highlighted in yellow. The gene and the transcripts are represented above the aligned reads.

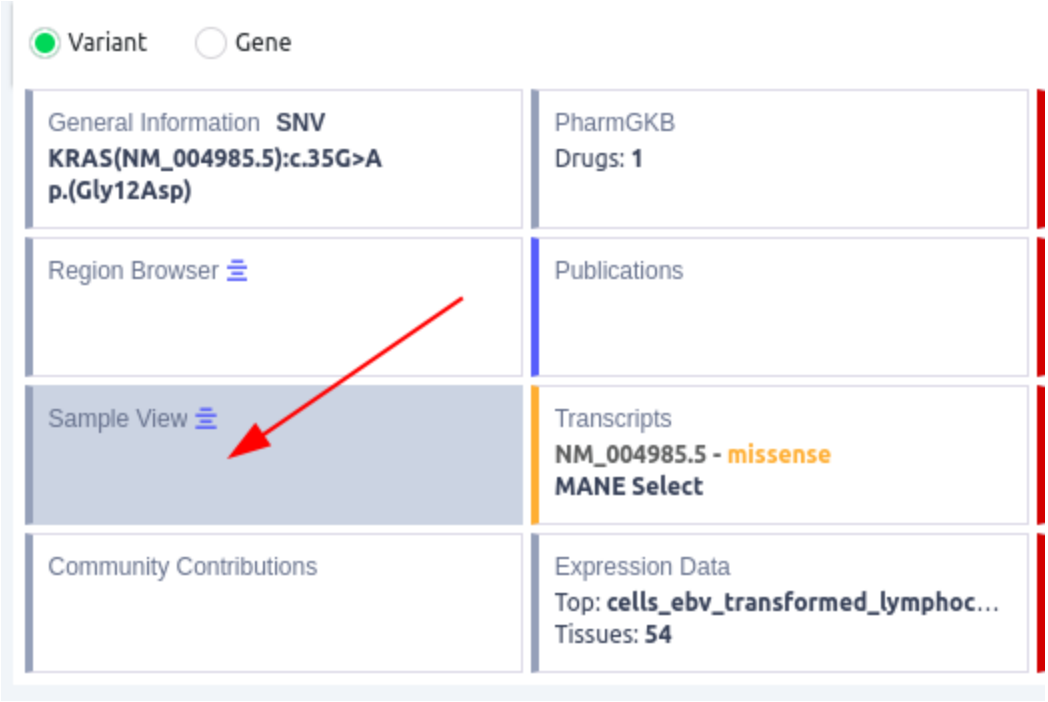
11.2.5 Browsing through the samples of a CNV analysis

You can search through the samples analyzed under the same CNV/SV analysis by visiting the results page of one and using the arrows to move to the next or previous sample:



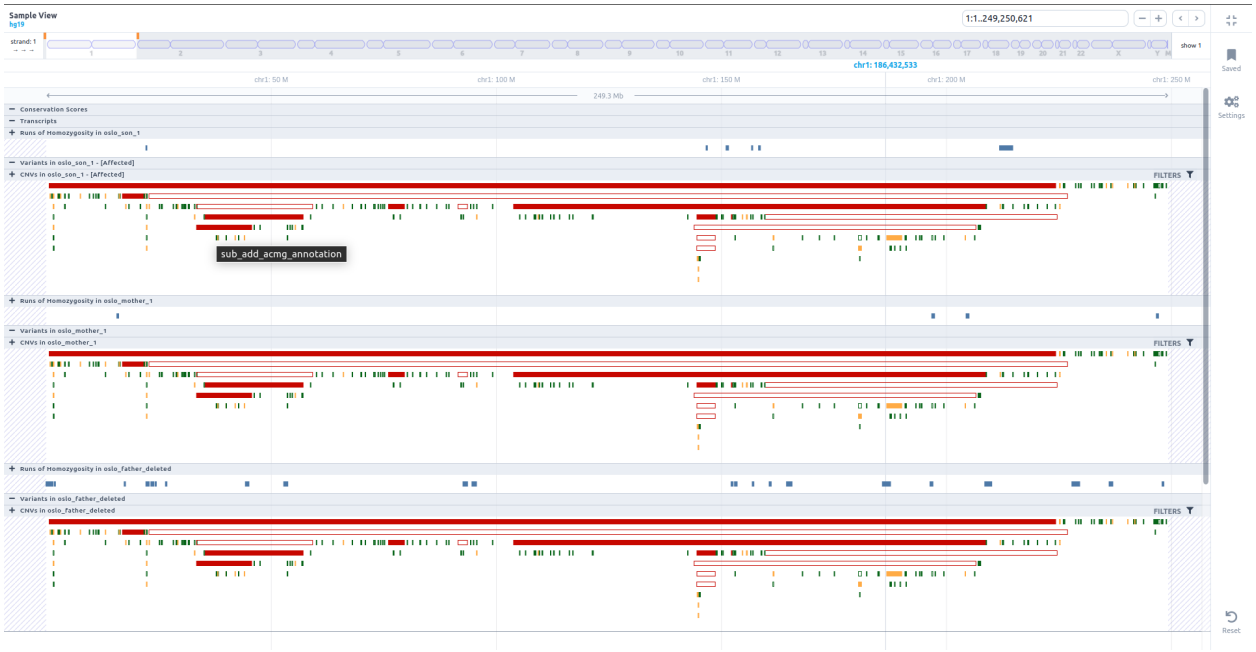
11.3 Sample View

A new "Sample View" card has been added to VarSome Clinical. This presents a genome region browser, but one that is focused on displaying the variants (SNPs, indels, CNVs) identified in the current sample.



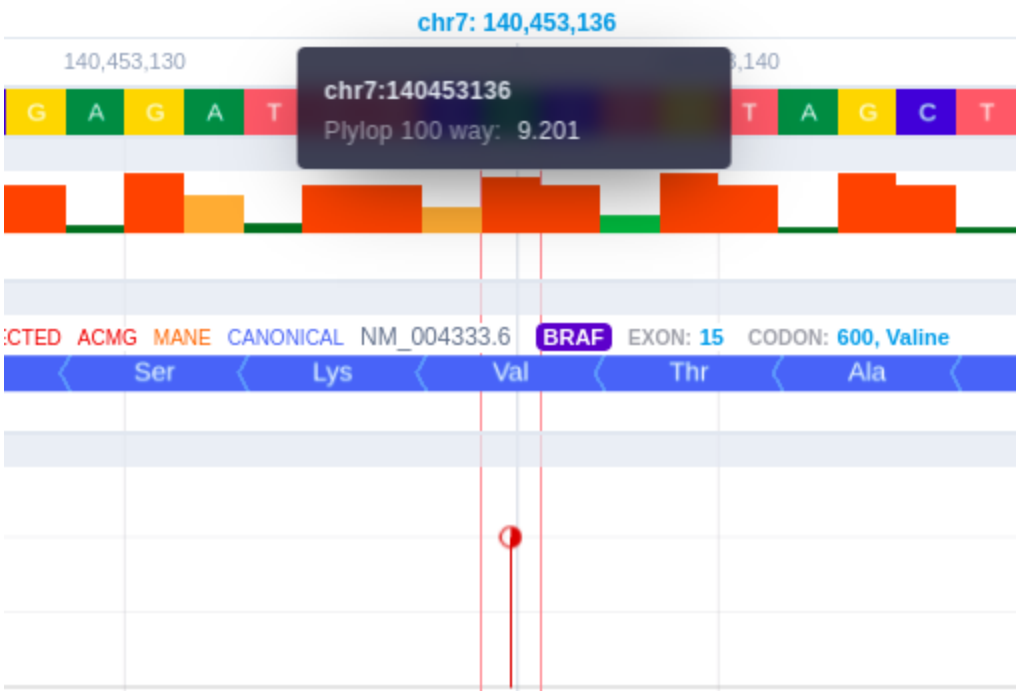
This sample browser incorporates the following features:

- Conservation scores per position
- A track showing the transcript that was used for the Germline Variant Classification of the selected variant
- More filtering options
- Ability to switch strand and expand the visualization to all chromosomes to see all analyses' data if the analyses are small enough (less than 10k variants).
- Touch and gesture support
- Customisable color palette
- **6 track levels:**
  Chromosome
  Conservation scores
  Transcripts
  Runs of Homozygosity
  Variants in the sample (Single or Multi)
  CNVs in the sample.

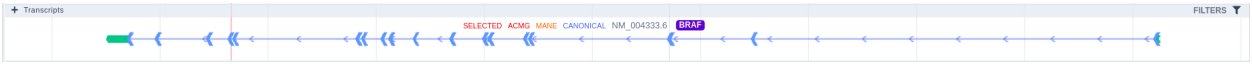


Conservation scores

The conservation score per chromosome position is displayed with different colors related to pathogenicity, and its value is presented as a bar. To inspect the conservation scores for a smaller region, the user can zoom in from the corresponding button on the upper right of the browser. When hovering over the conservation bars, the chromosome position is displayed inside a black box alongside the conservation score, provided by the PhyloP software.

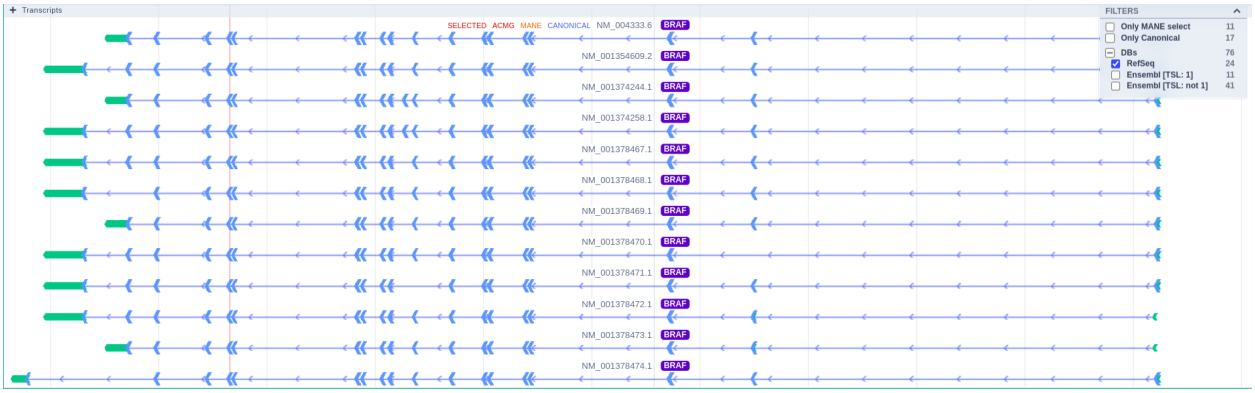


Transcripts



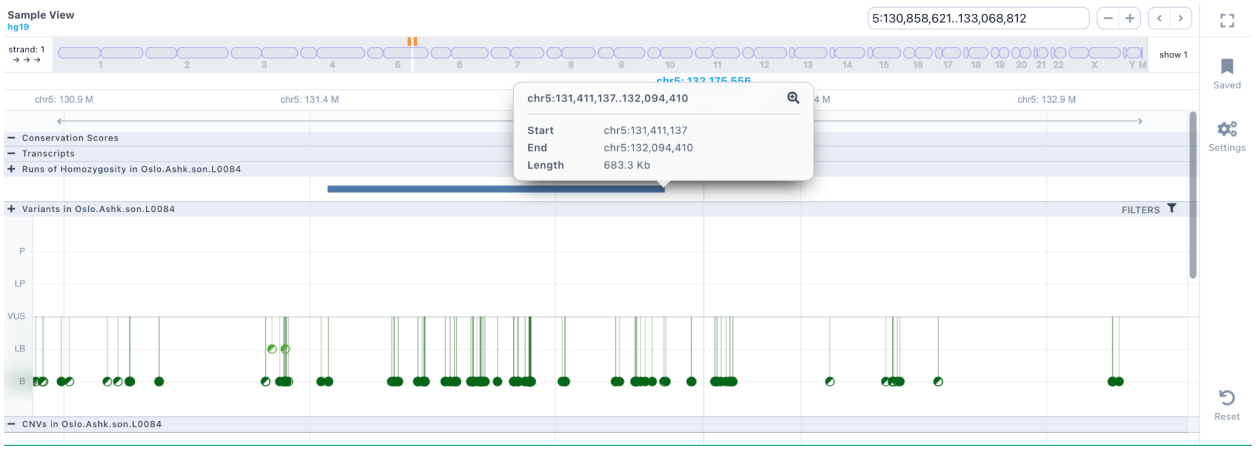
The Transcripts track shows the MANE RefSeq transcript and the transcript selected for the Germline Variant Classification by default. Hovering over the transcript track, the user is able to see useful transcript-related information such as the source, the transcript reference ID, gene name, exon number, number of codons and amino acids.

A user can select which transcripts of interest should be displayed by using the filtering options on the top right of the track. Moreover, a transcript of interest can be selected by clicking on the displayed option of the tooltip, and it will be visible regardless of the applied filters.



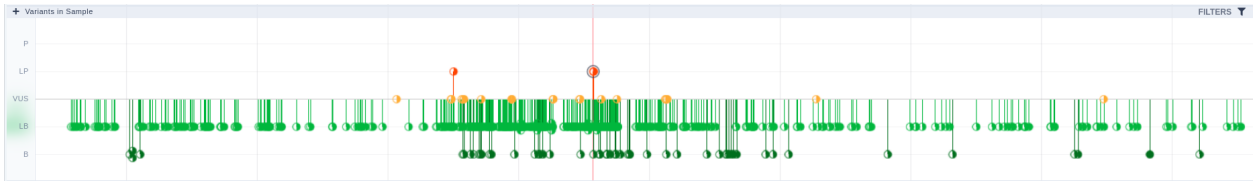
Runs of Homozygosity

The Runs of Homozygosity (ROH) are displayed as bars across the genome. Hovering over the bar will display the start, end and length of the ROH.

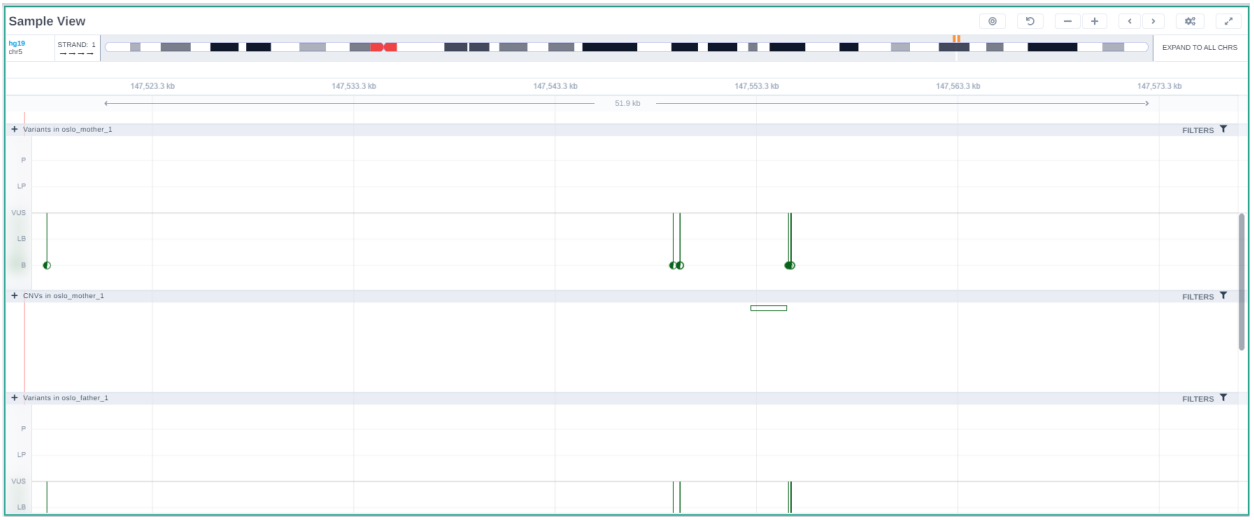


Variants in sample

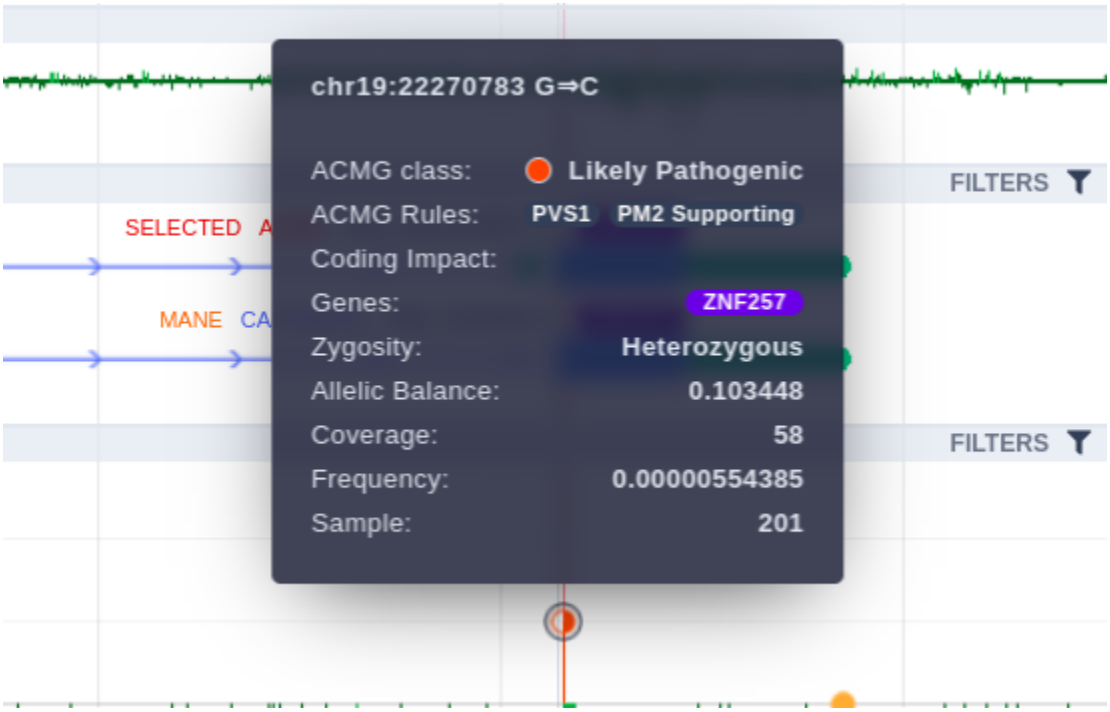
The SNVs and indels of the sample are displayed as a lollipop graph with different colors related to pathogenicity. The zygosity state of each variant is presented at the edge of the lollipop display (homozygous variants as filled disks, heterozygous as half-filled), and it corresponds to the one presented in the results table.



The Sample Browser can be used to visualize variants across all component samples of multi-sample analyses (up to 10 samples).

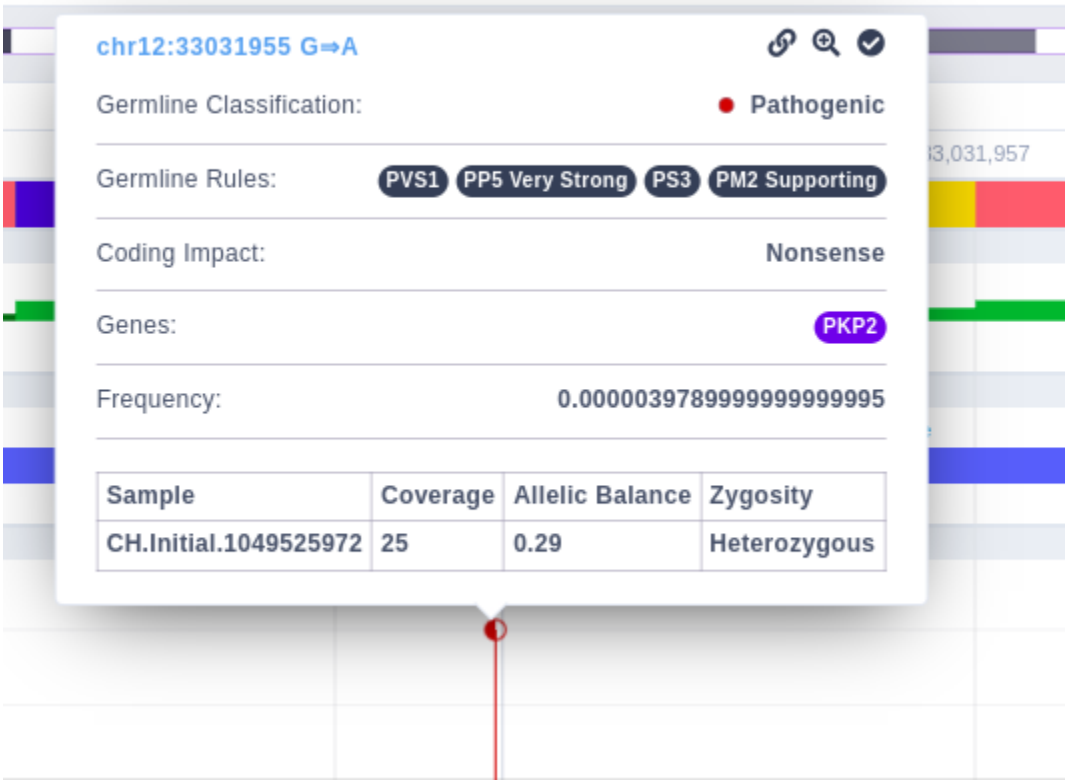


Hovering over a variant, a tooltip is presented containing information about the Germline Variant Classification class, the triggered rules, coding impact, zygosity, allelic balance etc.

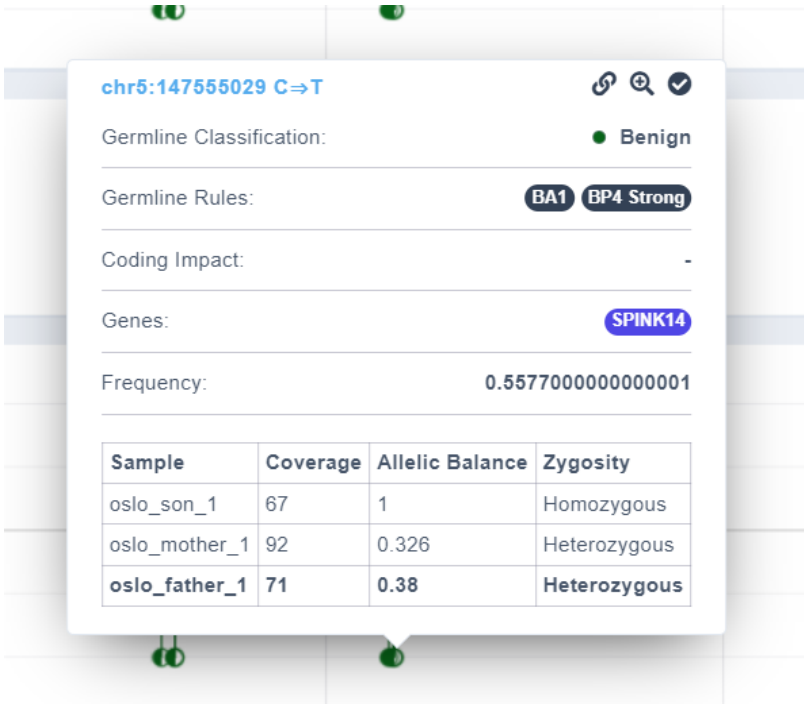


By clicking on the variant, a box is presented providing some additional options, such as a link to open the analysis in which the variant has been identified, a zoom-in option and one to highlight this variant. It will also display a table with the sample, coverage, allelic balance and zygosity information.

Single sample:



Multi sample:

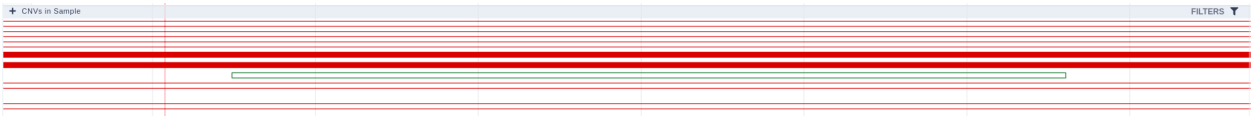


This track can be filtered based on pathogenicity class, zygosity, coding impact as well as CNVs

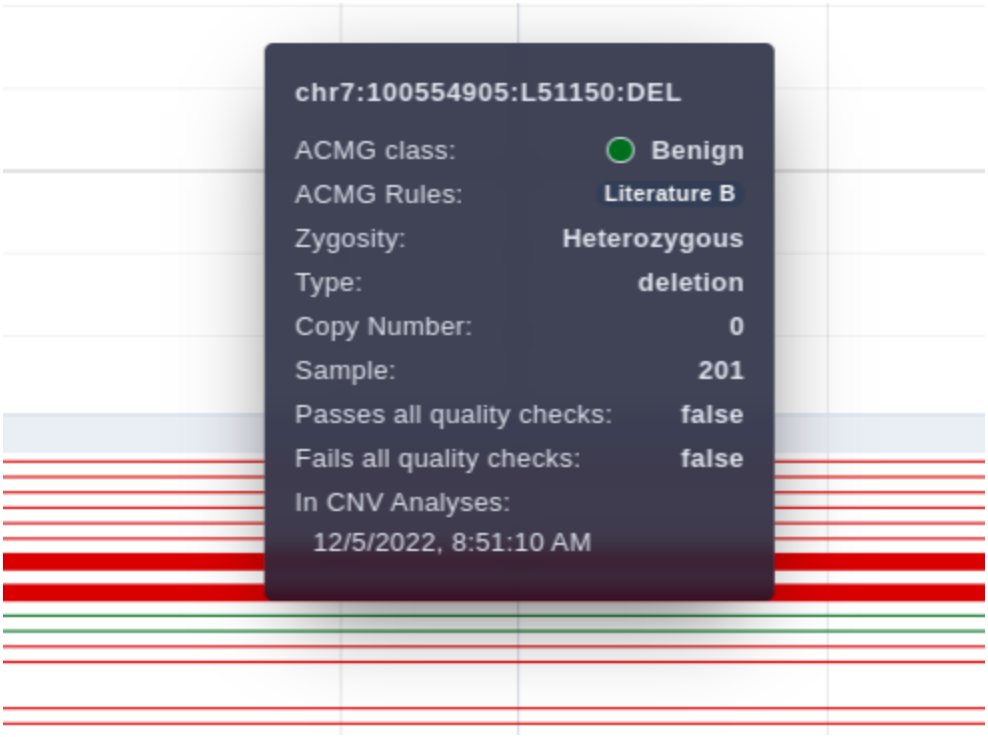
that are present in any other sample or are present in all other samples or are not present in the current sample and coverage. To see the available options, click on the filters button on the top right of the track.

CNVs in sample

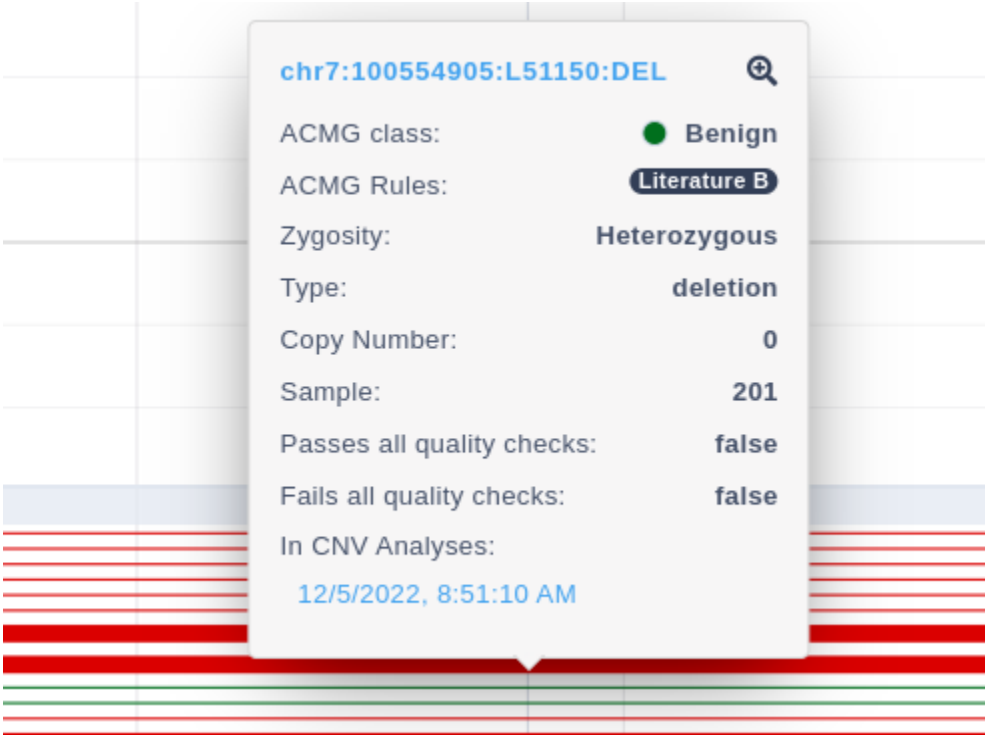
If a CNV sub-analysis has been launched for a sample, an additional track will be displayed for the identified CNVs. Each CNV is coloured based on the pathogenicity classification and either filled (insertions) or empty (deletions).



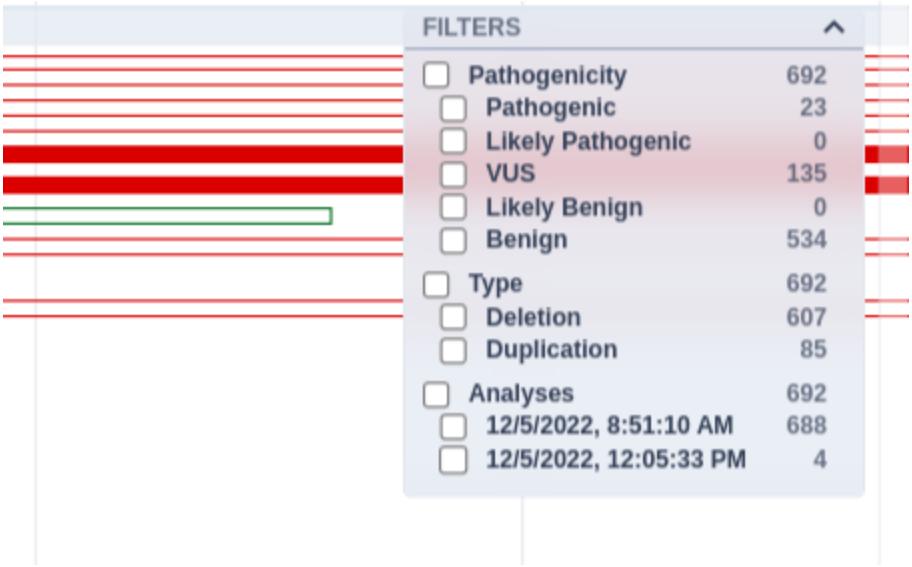
Hovering over a CNV, a tooltip is displayed containing information about the Germline Variant Classification class, zygosity, type etc.



By clicking on the CNV, a box is displayed providing a zoom option for it.



Additionally, this track can be filtered based on pathogenicity class, type and analyses in which the CNVs are identified. To see the available options, click on the filters button on the top right of the track.



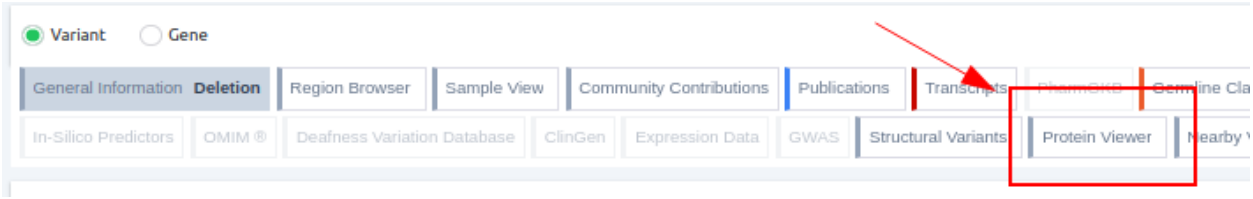
12. VarSome Clinical Frequently Asked Questions (FAQs)

12.1 3D protein viewer

How to access the protein viewer tool

This tool is accessible from the "Variant/Gene" table at the top of the window. Click on the "Protein Viewer" button to open a new window with the 3D Protein Viewer tool.

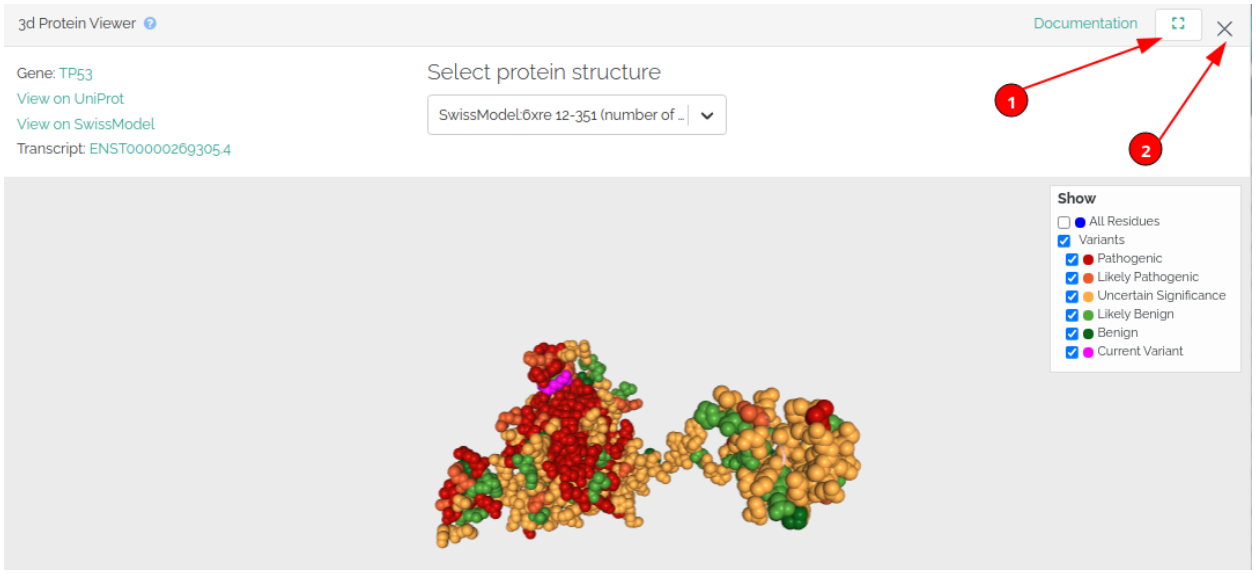
From the Variant table



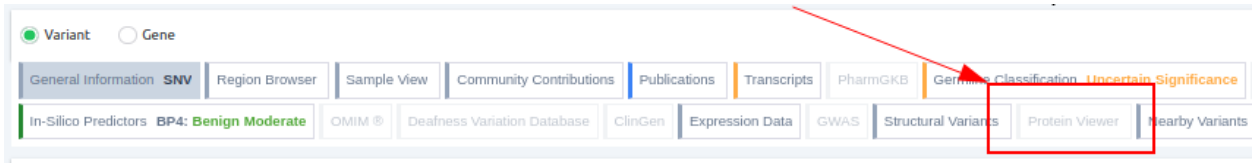
From the Gene table

Let's use the variant **TP53:R175L** as an example on the 3D Protein Viewer.

Go to the top right corner of the window to maximize the protein viewer page (1) or to close the window (2) and keep navigating through the VarSome website.



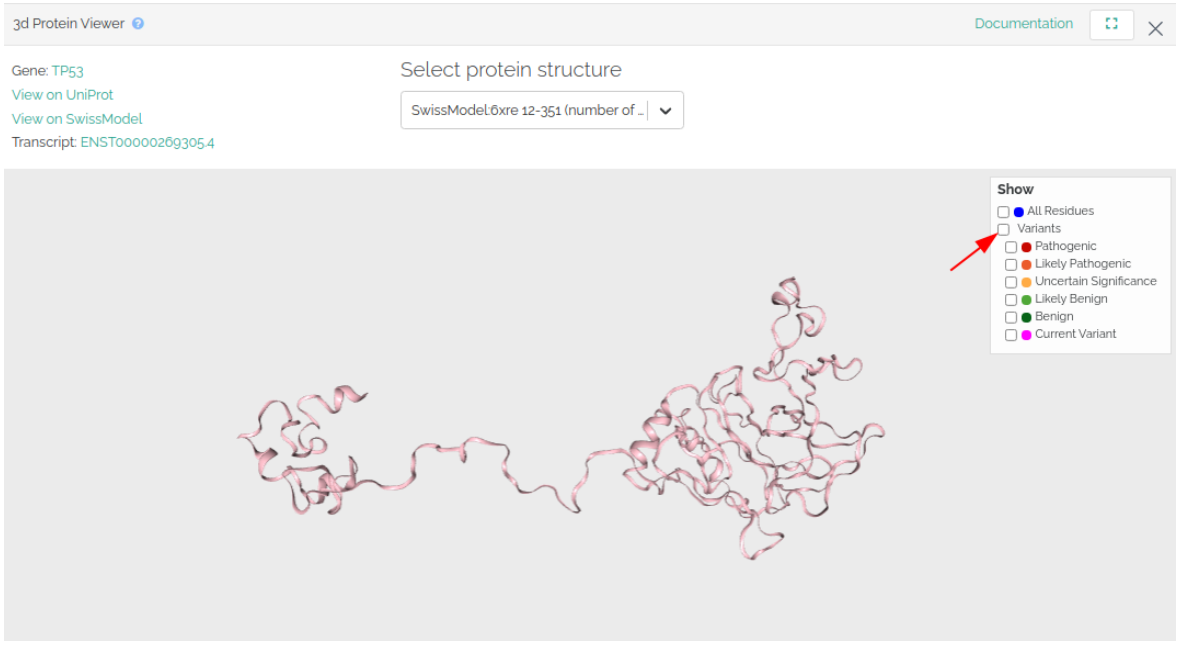
Please note that if there is no available structure, the "Protein Viewer" icon will be grayed out:



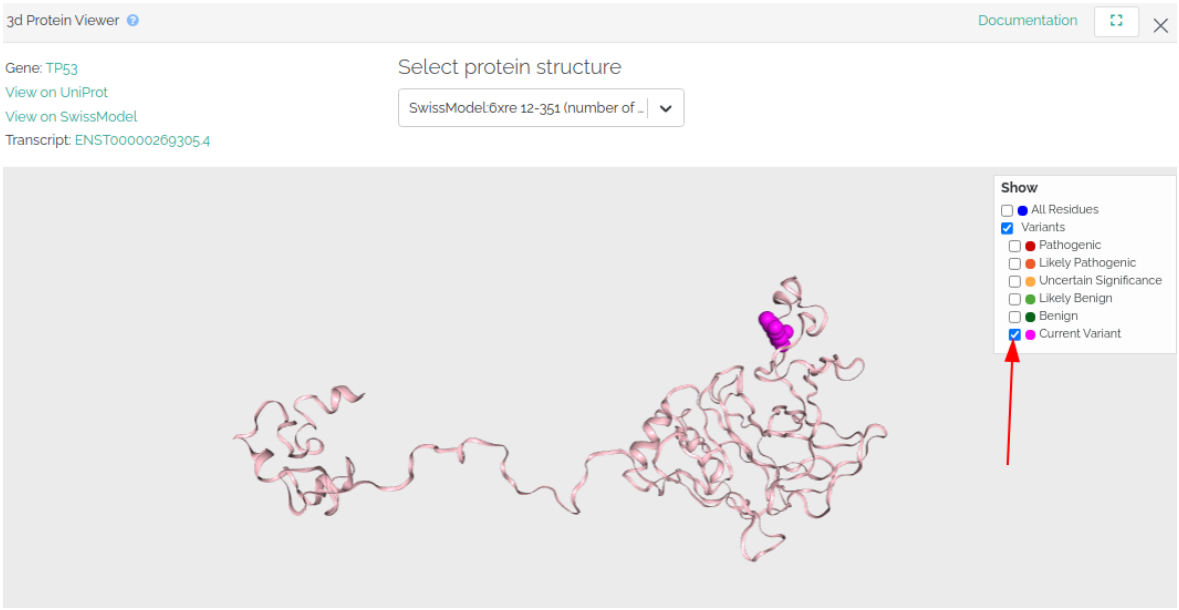
How to display variants on the protein structure

Variants reported from ClinVar, Uniprot and the VarSome Community are mapped onto the protein structure and are colored according to the Germline Variant Classification. Additionally, you will be able to see the variant of interest (the variant whose VarSome page you are currently visiting) in pink. You can decide which variants are shown in the "Show" menu on the top right corner.

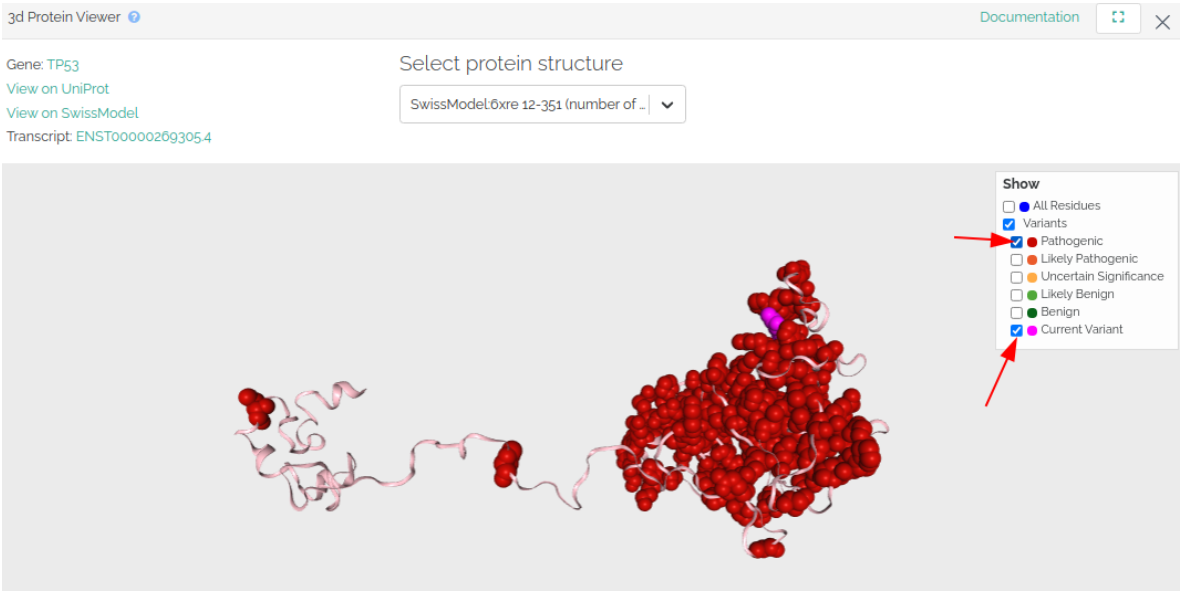
- Click on “Variants” to deselect all variants and display the protein structure without highlighting any residue position.



- Click on "current variant" to show the variant you have searched on VarSome.com.



- Click on the other variant options to start adding variants to the protein structure. For example, you can select "current variant" and "pathogenic" variants to check whether your variant clusters with the reported pathogenic variants in that gene.



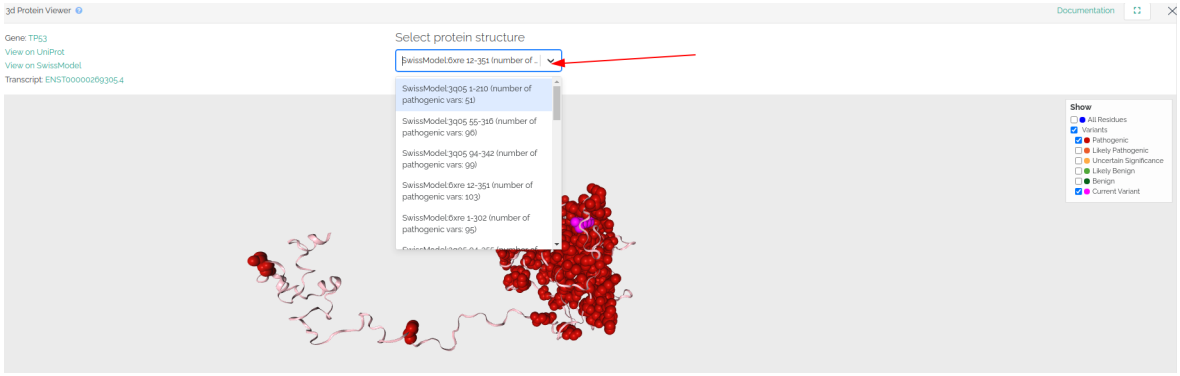
- Click on a residue of the protein structure to show a list of reported variants at that position below the protein structure.

How to select a different protein structure

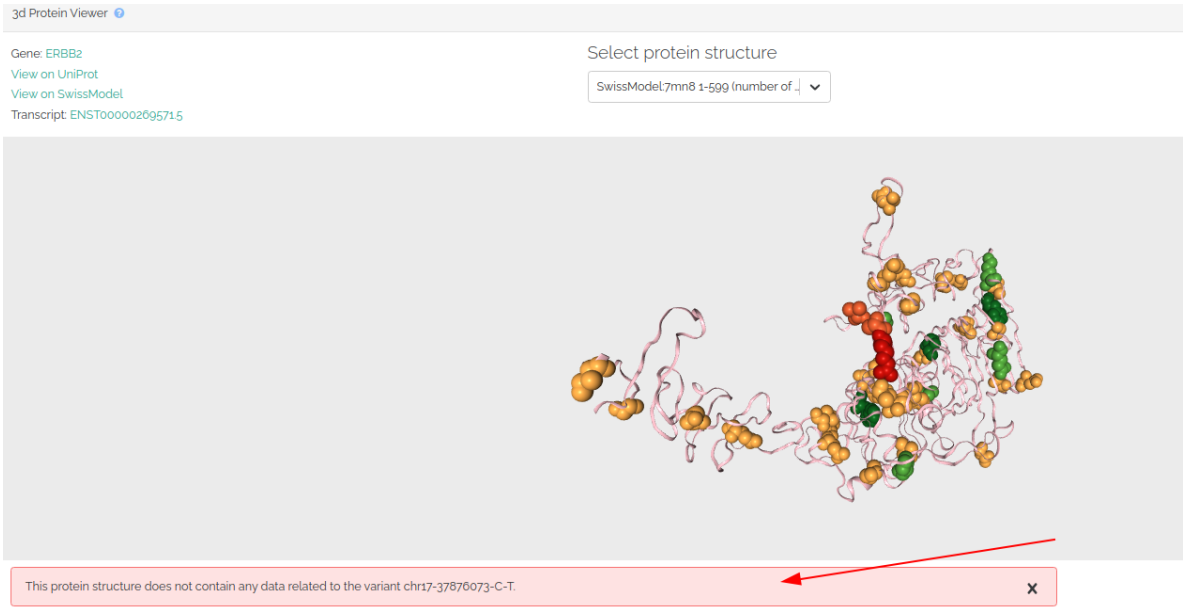
We import protein structures from Swiss-Model. The protein structure shown by default will be:

1. The structure that contains the variant of interest, if available.
2. The structure with the highest number of pathogenic variants.

Click the arrow next to the protein structure's name to select a different protein structure from the drop-down menu.



Sometimes the selected protein structure does not contain any data related to the variant you have searched. If this is the case, a message below the structure will be shown to notify the user.



⚠️ Please note that if you experience issues when using the 3D protein, please try to clearing your browser data (e.g. cookies) before continue navigating. Please, contact us if the issue persists.

12.2 ClinVar

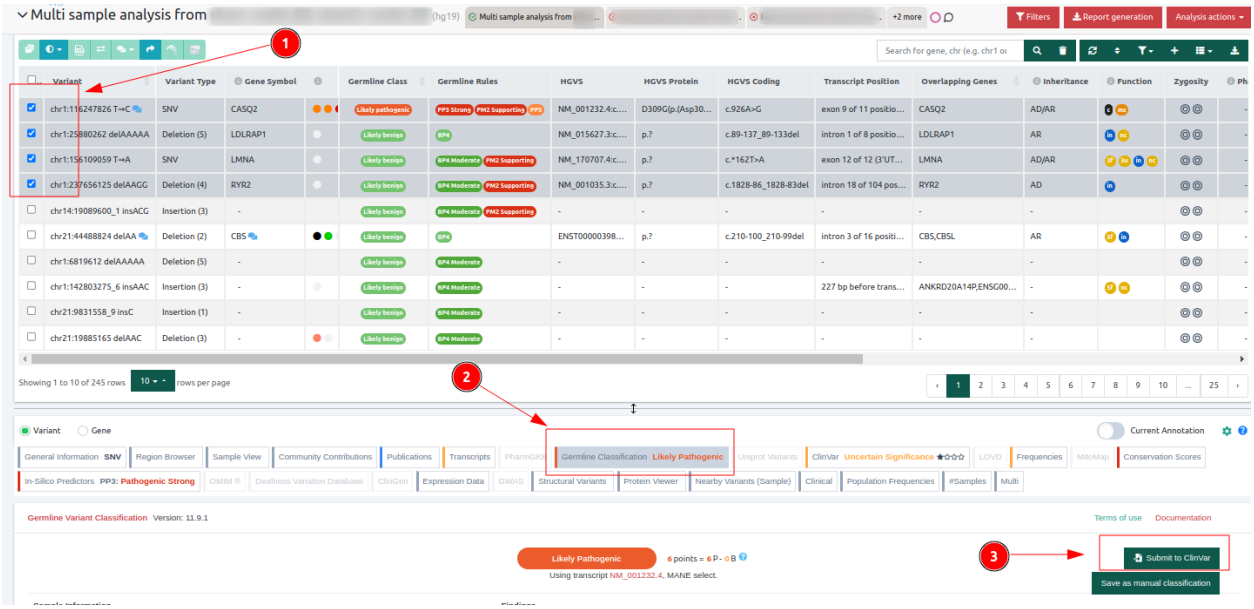
12.2.1 How can I submit to ClinVar?

Steps for the submission to ClinVar

Step 1
Select from the Dashboard or the Analyses page the analysis that you want.

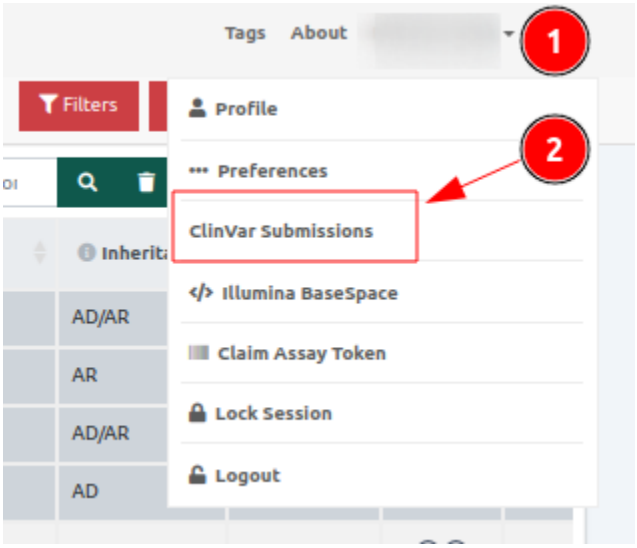
Step 2
Select the variant of interest from the results table and click on the “Germline Classification” component that is available on the Varsome Clinical results analysis page.

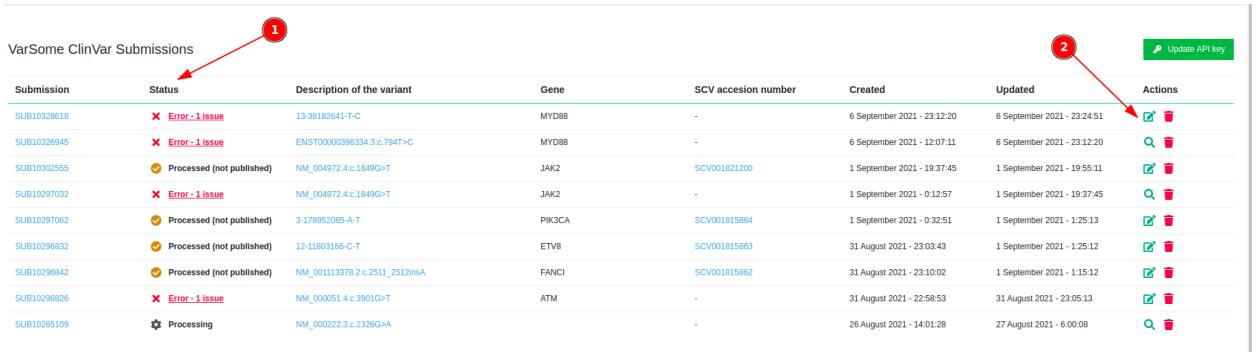
Then click on the ''Submit to ClinVar'' button, which is available inside the Germline Classification component (screenshot below).

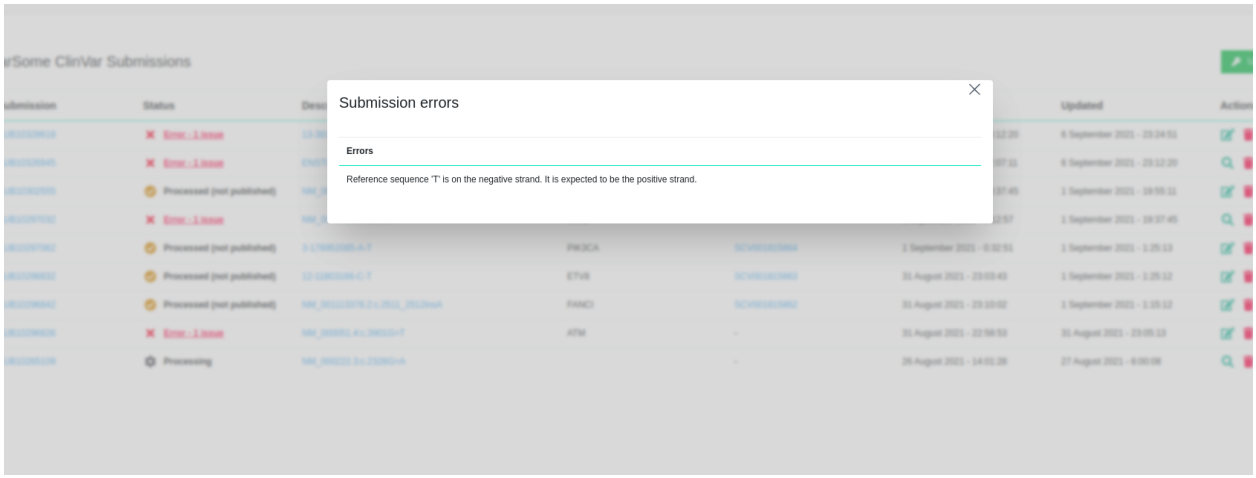


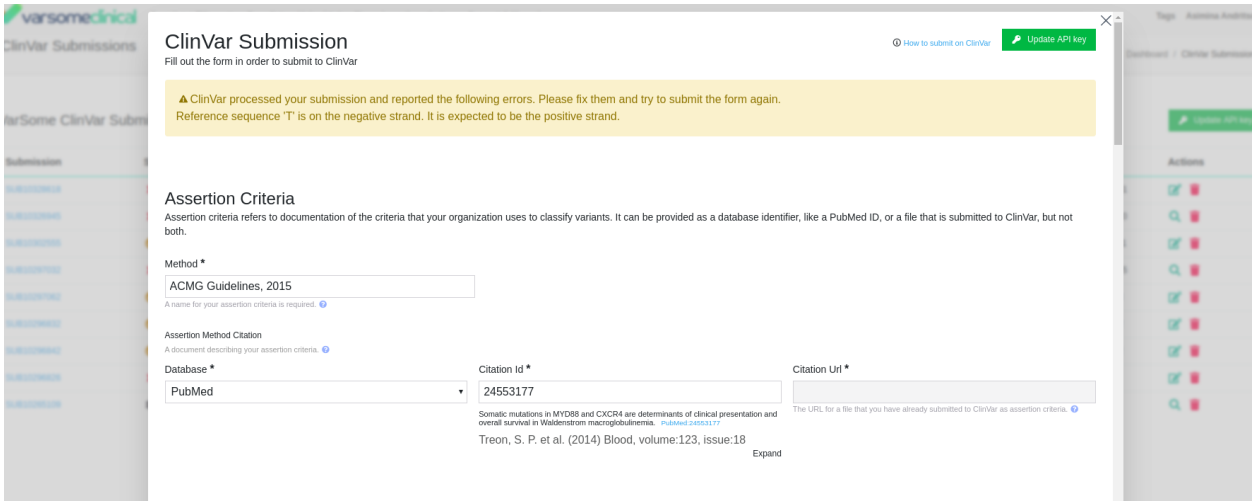
Step 3

After Submitting to ClinVar, you will be able to see all submissions you have made, track their status or tackle any potential submission errors.









12.2.2 What is ClinVar class?

Clinvar class is the clinical significance value given by ClinVar database: http://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/

Guidance for use in ClinVar SCV records

- **Benign** - As recommended by ACMG for variants interpreted for Mendelian disorders.
- **Likely benign** - As recommended by ACMG for variants interpreted for Mendelian disorders.
- **Uncertain significance** - As recommended by ACMG for variants interpreted for Mendelian

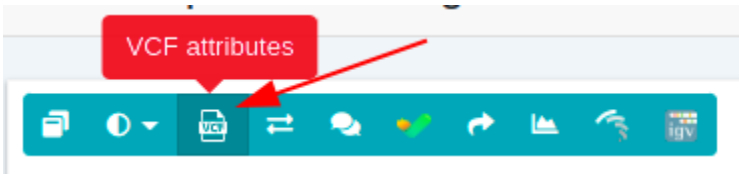
disorders.

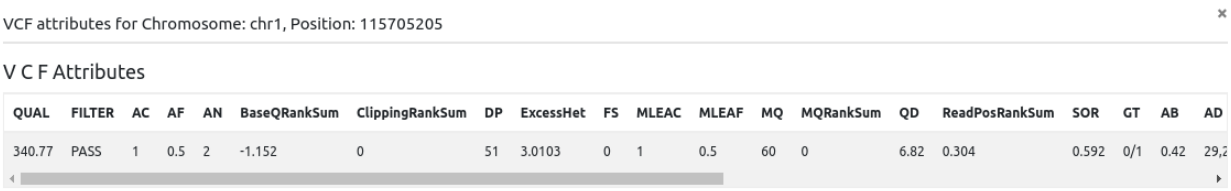
- **Likely pathogenic** - As recommended by ACMG for variants interpreted for Mendelian disorders.
- **Pathogenic** - As recommended by ACMG for variants interpreted for Mendelian disorders.
- **drug response** - A general term for a variant that affects a drug response, not a disease. We anticipate adding more specific drug response terms based on a recommendation by CPIC.
- **association** - For variants identified in a GWAS study and further interpreted for their clinical significance.
- **risk factor** - For variants that are interpreted not to cause a disorder but to increase the risk.
- protective - For variants that decrease the risk of a disorder, including infections.
- **affects** - For variants that cause a non-disease phenotype, such as lactose intolerance.
- **conflicting data from submitters** - Only for submissions from a consortium, where groups within the consortium have conflicting interpretations of a variant but provide a single submission to ClinVar.
- **other** - If ClinVar does not have the appropriate term for your submission, we ask that you submit "other" as clinical significance and contact us to discuss if there are other terms we should add.
- **not provided** - For submissions without an interpretation of clinical significance. The primary goal of ClinVar is to archive reports of clinical significance of variants. Therefore submissions with a clinical significance of "not provided" should be limited to:
  - "literature only" submissions that report a publication about the variant, without interpreting the clinical significance
  - "research" submissions that provide functional significance (e.g. undetectable protein level) but no interpretation of clinical significance
  - "clinical testing" submissions from clinics or physicians that provide additional information about individuals with the variant, such as observed phenotypes, but do not interpret the clinical significance

12.3 Explanation of VCF attributes

A VCF (Variant Calling Format) is a text file format that stores information about genetic variation. Each variant shown in the variant table corresponds to one line in the VCF file. Besides the variant position, reference and alternative alleles, each line contains additional information (attributes) such as different quality measures that can be accessed from VarSome Clinical. To see these, you need to select the variant in the table, then you can click on the VCF icon to display the VCF attributes as shown in the picture below.

The germline and somatic variant calling pipelines use different algorithms for variant calling and, therefore, the VCFs obtained contain different attributes. In the following lines we will describe the VCF attributes for germline and somatic VCFs.





The germline and somatic variant calling pipelines use different algorithms for variant calling and, therefore, the VCFs obtained contain different attributes. In the following lines we will describe the VCF attributes for germline and somatic VCFs.

Germline VCF

- QUAL: it is the phred-scaled probability that the site has no variant. This quality value is used in the pre-processing step (before the annotation) to decide which variants have a PASS call status and which have a FAIL.
- FILTER: filters that have been applied to the variant.
- AC: allele count in genotypes, for each ALT (alternative) allele, in the same order as listed
- AF: allele frequency for each ALT allele in the same order as listed.
- AN: total number of alleles in called genotypes.
- BaseQRankSum: a z-score for base qualities of reference and alternative alleles. For example, a BaseQRankSum close to 0 means the reference and alternative alleles have the same base qualities and a BaseQRankSum around 2 means they differ by 2 SDs. A positive value of 2 means that ALT alleles have higher qualities than REF (reference).
- ClippingRankSum: Z-score From Wilcoxon rank sum test of ALT vs. REF number of hard clipped bases.
- DP: approximate read depth; some reads may have been filtered.
- ExcessHet: phred-scaled p-value for exact test of excess heterozygosity.
- FS: Fisher strand. It is a measure of sequencing bias. This measures if one strand is preferred than the other when sequencing. Larger values means larger bias.
- MLEAC: maximum likelihood expectation of AC (Allele counts).
- MLEAF: maximum likelihood expectation of AF (Allele Frequency).
- MQ: mapping quality. Comparison quality value.
- MQRankSum: this is the u-based z-approximation from the Rank Sum Test for mapping qualities. It compares the mapping qualities of the reads supporting the reference allele and the alternate allele.
- QD: QUAL normalized by read-depth (QUAL/DP).
- ReadPosRankSum: this is the u-based z-approximation from the Rank Sum Test for site position within reads. It compares whether the positions of the REF and ALT alleles are different within the reads.
- SOR (StrandOddsRatio): this is another way to estimate strand bias using a test similar to the symmetric odds ratio test. FS tends to penalize variants that occur at the ends of exons and SOR does not. Reads at the ends of exons tend to only be covered by reads in one direction and FS gives those variants a bad score. SOR will take into account the ratios of reads that cover both alleles.
- GT: genotype. It is encoded as allele values separated by either of / (not phased) or | (phased). The allele values are 0 for the reference allele (what is in the REF field), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on.

- AD: allelic depths for the REF and ALT alleles in the order listed.
- GQ: phred-scaled probability that the call is incorrect.
- PGT: physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another.
- PID: physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group.
- PL: normalized, phred-scaled likelihoods for genotypes as defined in the VCF specification.
- SAC: number of reads on the forward and reverse strand supporting each allele (including reference).
- MIN_COVERAGE: minimum coverage threshold considered to give the variant a status of PASS.
- MIN_QUALITY_INDELS: minimum QUAL threshold considered to give the INDEL variant a status of PASS.
- MIN_QUALITY_SNV: minimum QUAL threshold considered to give the variant a status of PASS.

Somatic VCF

- FILTER: filters that have been applied to the variant.
- AS_FilterStatus: filter status for each allele, as assessed by ApplyRecalibration. Note that the VCF filter field will reflect the most lenient/sensitive status across all alleles.
- AS_SB_TABLE: allele-specific forward/reverse read counts for strand bias tests. Includes the reference and alleles separated by a '|'.
- DP: approximate read depth; some reads may have been filtered.
- ECNT: number of events in this haplotype.
- GERMQ: phred-scaled quality that ALT alleles are not germline variants
- MBQ: median base quality.
- MFRL: median fragment length.
- MMQ: median mapping quality.
- MPOS: median distance from end of read.
- POPAF: negative log 10 population allele frequencies of ALT alleles.
- TLOD: log 10 likelihood ratio score of variant existing versus not existing.
- GT: genotype, encoded as allele values separated by either of / (not phased) or | (phased). The allele values are 0 for the reference allele (what is in the REF field), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on.
- AD: allelic depths for the REF and ALT alleles in the order listed.
- AF: allele fractions of alternate alleles in the tumor.
- F1R2: count of reads in F1R2 pair orientation supporting each allele.
- F2R1: count of reads in F2R1 pair orientation supporting each allele.
- SB: per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.
- MIN_COVERAGE: minimum coverage threshold considered to give the variant a status of PASS.
- MIN_QUALITY_INDELS: minimum QUAL threshold considered to give the INDEL variant a status of PASS.
- MIN_QUALITY_SNV: minimum QUAL threshold considered to give the variant a status of PASS.

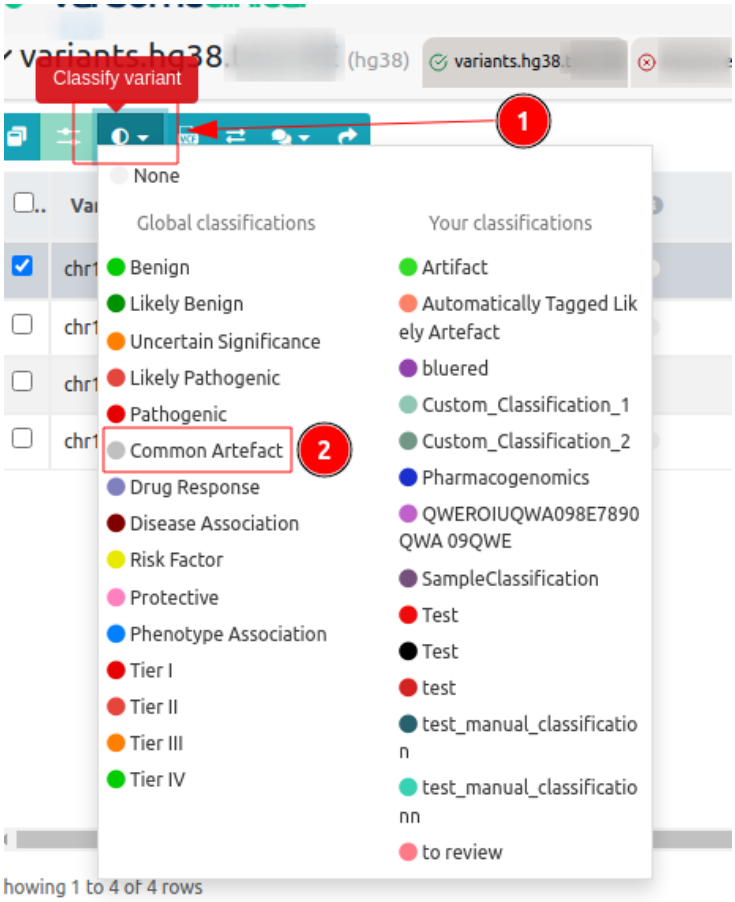
12.4 How can I filter out artifacts from my samples?

Please, follow the next steps in order to tag and filter out artifacts from the variant table.

1. Tag a variant as a "Common artifact"

First of all you need to identify all variants that are potential artifacts. To do this you can:

- Add a custom variant classification: click on the icon and classify the variant as "Common artifact":

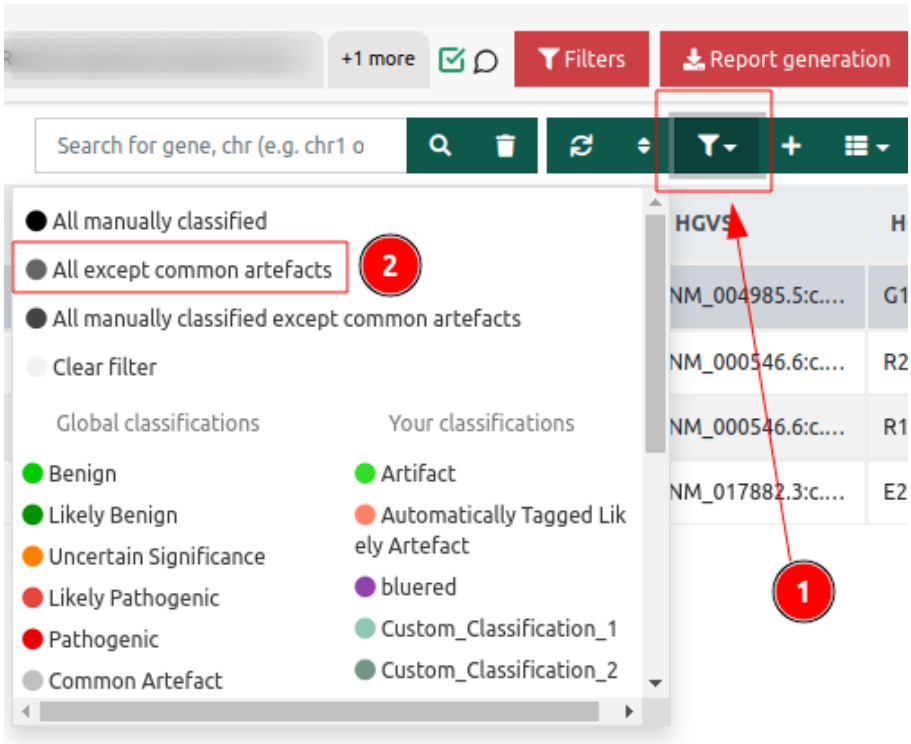


- If you have a pre-defined list of common artifacts, you can send us your list and we will label these variants as "Common artifacts". Please, find more information on section Custom Classifications and Comments.

⚠️ Please note that custom classifications and comments are associated to the variant. You only need to classify the variant once and this classification will be kept and shown in case the same variant is found in any other analysis of your VarSome Clinical account.

2. Filter out "Common artifacts"

Click on the icon above the variant table to select which variants would you like to show:



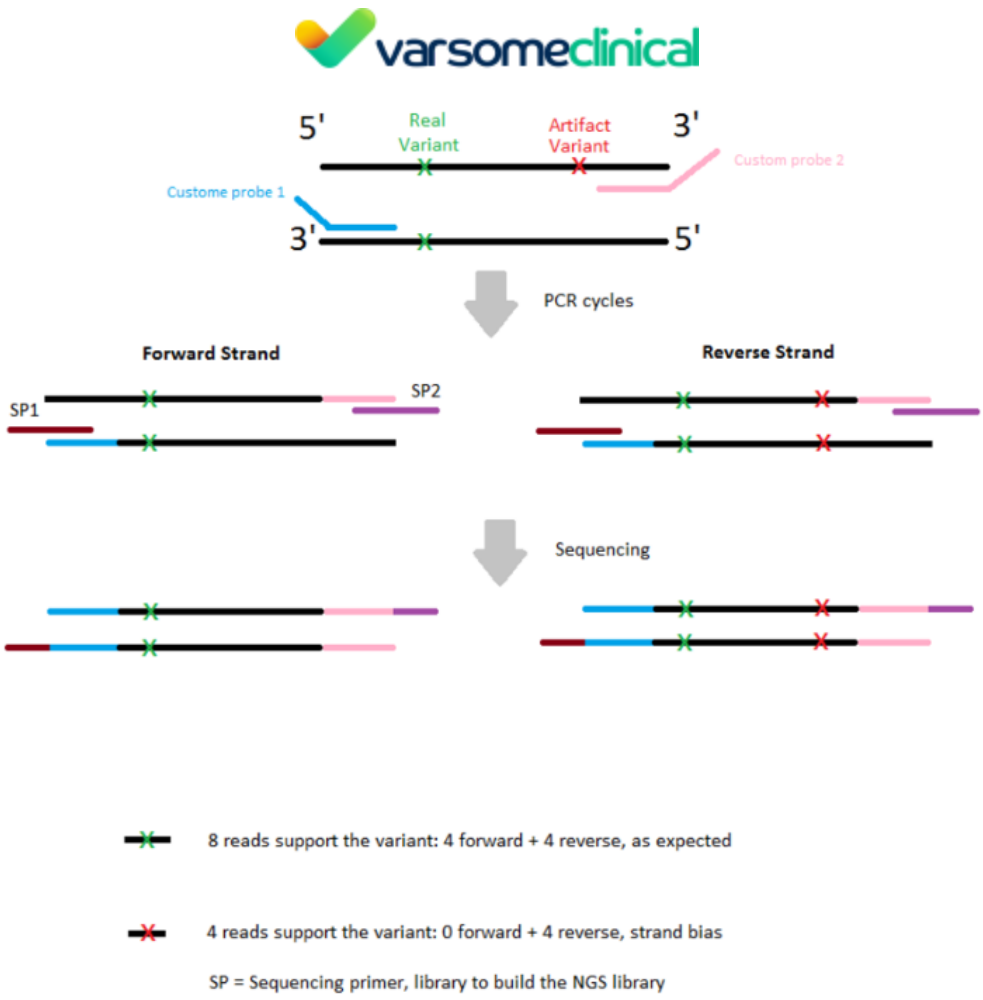
Click on show "All except common artifacts". This option will exclude the common artifacts from the variant table.

12.5 Why are variants identified in a single strand considered as artifacts, in particular when using amplicon kits?

Please see the following Technical Note from Illumina:

Figure 1: Detection and Differentiation of DNA Damage from Mutation—In panel A (left), cytosine deamination results in a nucleotide change in one strand of a DNA molecule, but does not alter the complementary nucleotide on the opposite strand. Sequencing each strand independently will yield base calls that differ between the 2 strands. In panel B (right), a true DNA mutation results in a nucleotide change in both strands of a DNA molecule. Sequencing each strand independently will yield the same variant call for both strands. The Amplicon - DS somatic variant caller filters false positives and reports confirmed variants that were identified using TruSight Tumor 26.

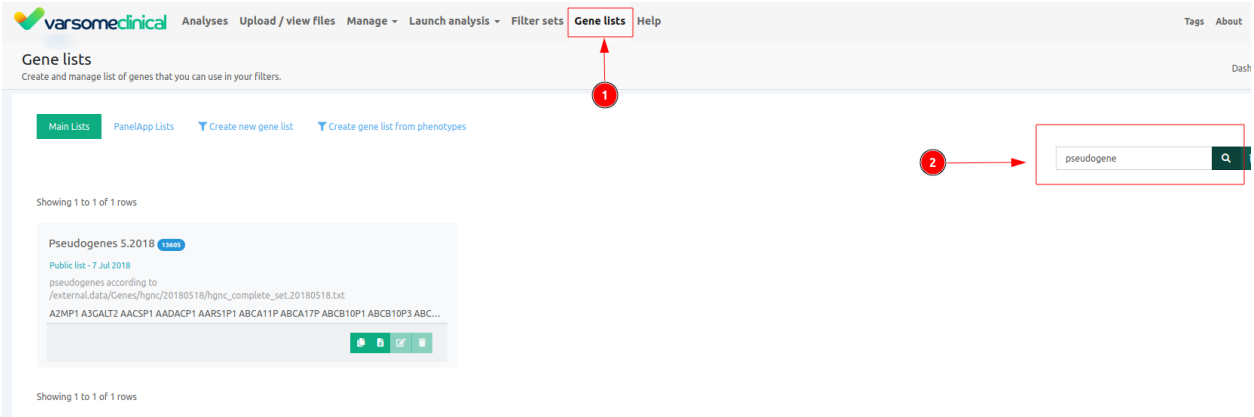
Each amplicon is the PCR product from one DNA strand, but since DNA is double stranded, a real variant is supposed to be found in both stands:



12.6 Pseudogenes

The list of pseudogenes is obtained from https://www.genenames.org.
For filtering with a pseudogene list you can use dynamic filters (the funnel icon on the left) and select a Gene List filter.
Or you can perform a Gene List analysis, which creates a sub-analysis like an algorithmic filter.

Where to find the gene list for pseudogenes?



Can VarSome Clinical recognize if a variant was identified on the true transcript or the

pseudogene? How can I be sure that the variants called in pseudogene rich regions do not derive from the pseudogene?

That's one of the limitations of NGS sequencing. Pseudogenes are real sequences and are present in the genome. When a read can be perfectly aligned to multiple regions of the genome, as can be the case if a region of a pseudogene is an exact copy of a region of a functioning gene, then what usually happens is that the read is either discarded or marked with a very low score since the aligner cannot know which of the two (or more) matching regions it should be aligned against.

And this isn't a problem that is specific to pseudogenes. There are multiple, real and functioning genes that are very similar to one another. SMN1 and SMN2, for example. One way of dealing with such duplicate sequences is to mask all such regions and only keep one of them unmasked. Please see Ebbert et al for more details. There is no good solution to this problem, it isn't a limitation of VarSome Clinical, but a direct result of how NGS sequencing works (see, for instance the following two links: https://www.fulgentgenetics.com/Neurofibromatosis and https://blueprintgenetics.com/pseudogene/).

12.7 Variant Sorting

We sort by pathogenicity, and the most pathogenic variants are usually coding. In the rare cases where that isn't the case, then an intronic variant that is more pathogenic than a coding one will be sorted first. The simplest way around this is to use a soft filter and select only coding variants. That doesn't affect the results of the analysis, the user can simply activate and deactivate the filter with a single click and can choose to see only coding or everything each time.

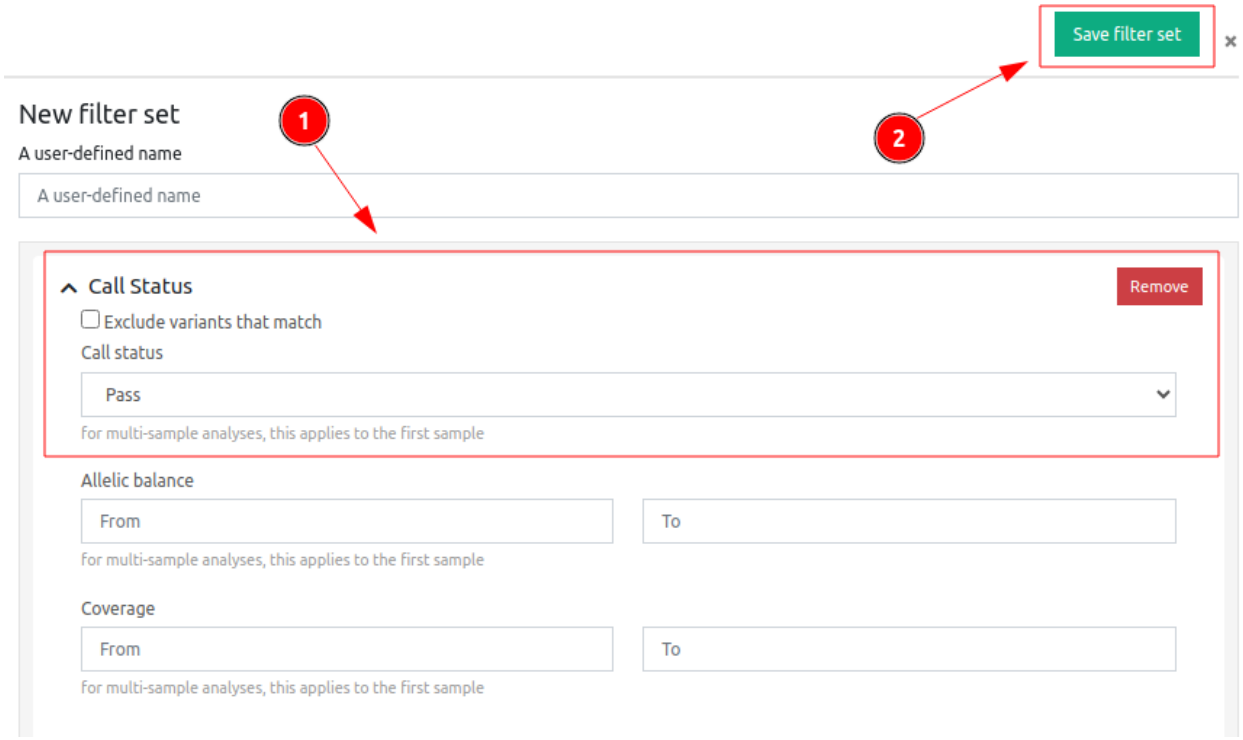
12.8 Tumor Mutational Burden (TMB)

The Tumor Mutational Burden (TMB) or tumor mutation load is considered an effective and clinically verifiable biomarker in multiple cancer types. Unfortunately, because there are many different ways employed by different clinics and labs to measure TMB, and there is no clear standard, VarSome Clinical currently does not offer a calculation of TMB in somatic samples as it would be impossible to provide a TMB metric that could satisfy all of our customers. However, VarSome Clinical offers the necessary tools to get the approximate number of germline and somatic variants and so calculate the TMB using your own definition.

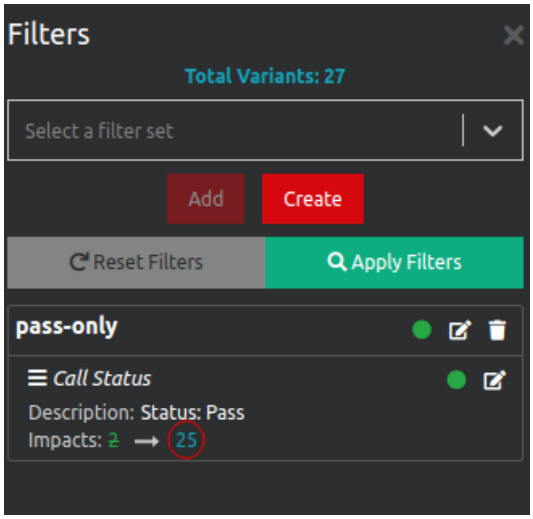
First, you need to know the total number of variants found in your sample. This metric can be found, for example, in the Analyses menu, under the column Variants.

If you are starting your somatic analysis from FASTQ, it is important to bear in mind that VarSome Clinical reports all variants found in the sample no matter their quality. Therefore, as a

first step to calculate the TMB, we highly recommend you to get the total number of variants found with a Call status of PASS, meaning that you will be counting only those variants passing the quality filters. To do this, please go to "Filters" > "Create" > "Call status", and create a filter to keep only the variants that have passed the quality filters.

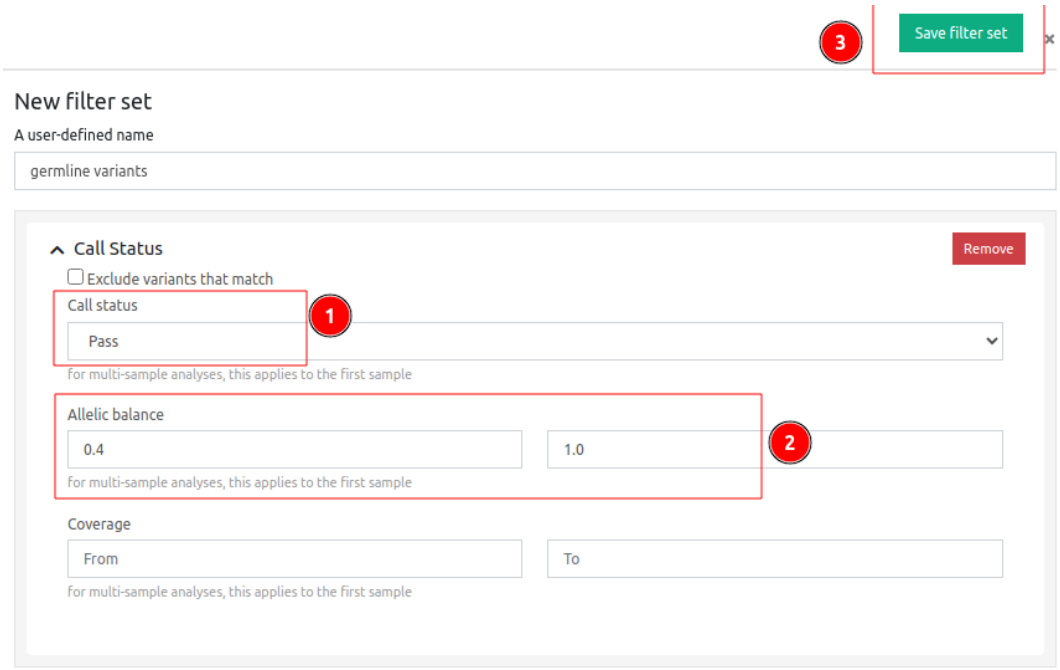


Once the filter is applied to your sample, you will get the total number of variants found with a PASS call status:



Then, you can filter by allelic balance as an approach to differentiate between germline and somatic variants in your sample. This can be assessed through the "Call Status" filter too. This filter contains the allelic balance field where you can include a specific range of allelic balance to retain and count the number of potential germline variants in your somatic sample.

To do this, please, create a new Call Status filter, selecting the variants with a PASS call status and the allelic balance range for germline variants.

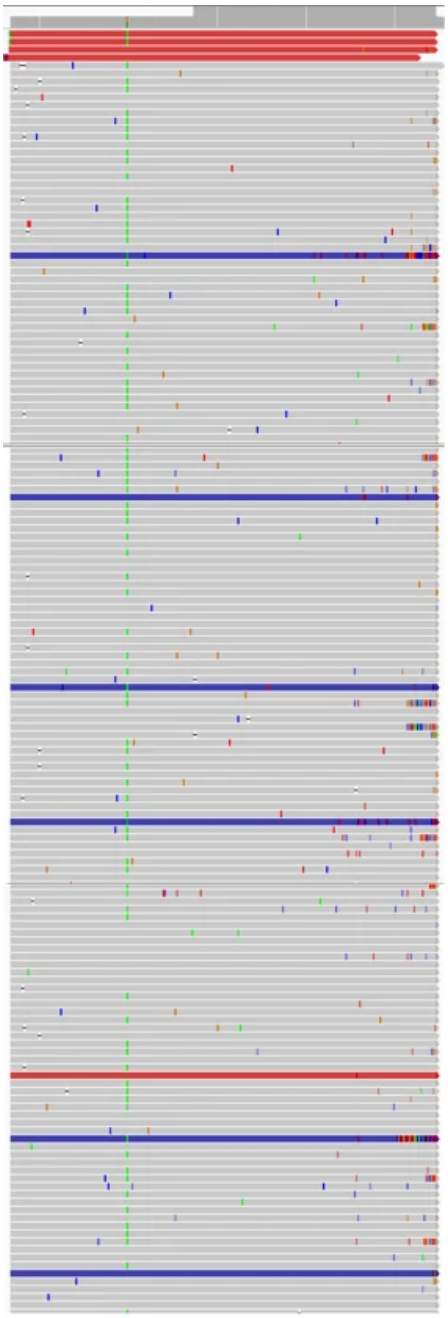


⚠️ Please note that the allelic balance range shown in the screenshot above might not be appropriate in your case. Please provide your own allelic balance range.

Finally, once you have obtained the number of germline variants, you can use this number and the total number of variants previously obtained to calculate the number of somatic variants found in your sample. These metrics can be used to calculate the TMB using your own definition.

12.9 Variant not found

The most common problems related to missing variants are usually twofold. First, at positions with high coverage (e.g. >3000), the Variant Caller will "down-sample" and select a random group of reads to perform the calling on. It is therefore possible that the selection doesn't accurately represent the real data and the variant might be missed. However, in cases where a good percentage (50% or more) of the reads support the variant, most likely the actual problem is coming from the noise in the region. As shown in the IGV screenshot example below, there are multiple mismatches in the region around the variant (green line).

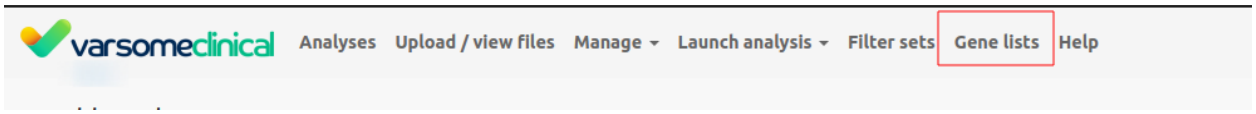


Every color indicates a mismatch and colored reads indicate paired reads whose pair was aligned to a different chromosome. Such "noisy" regions confuse the Variant Callers because when calling, they don't only look at the target position but attempt to call by region haplotype. So they use the region around the variant and build a haplotype based on the region and can miss the variant of interest. Another such example could be the presence of segmental duplications, which makes each read align almost perfectly to separate locations in the genome, leading to low alignment scores.

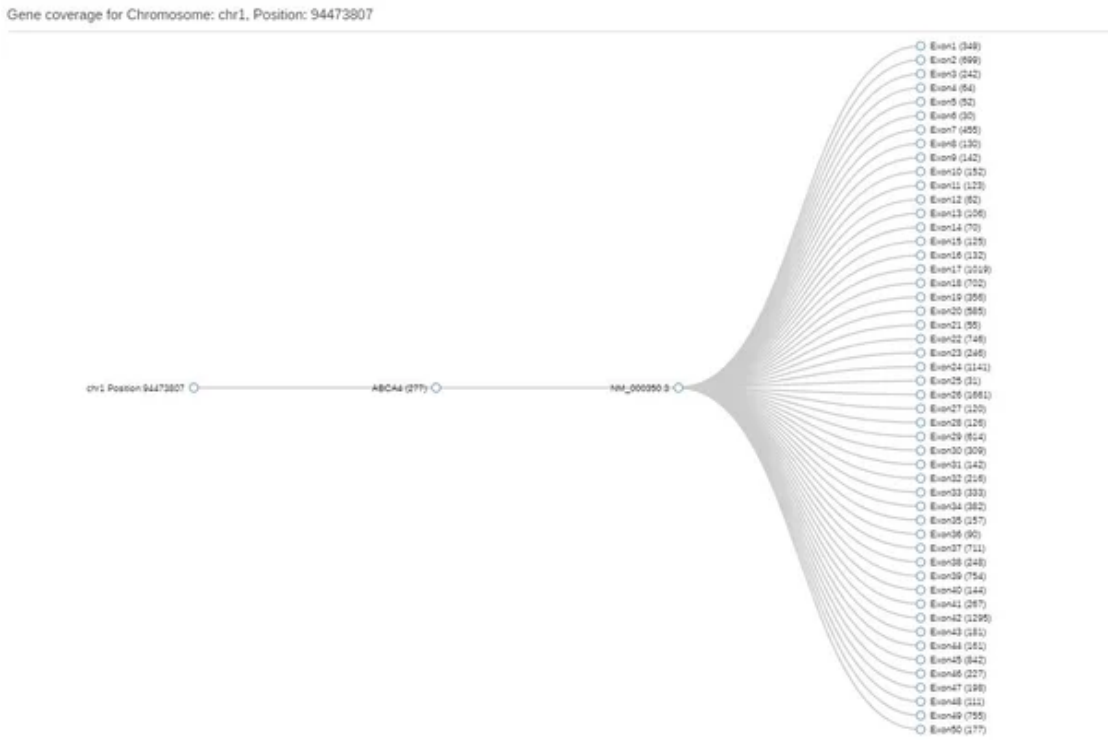
12.10 Investigating exon coverage

There's no direct way of seeing gaps in an exon's coverage, but you can get part way there by using the Coding coverage report as opposed to the Region list coverage report.

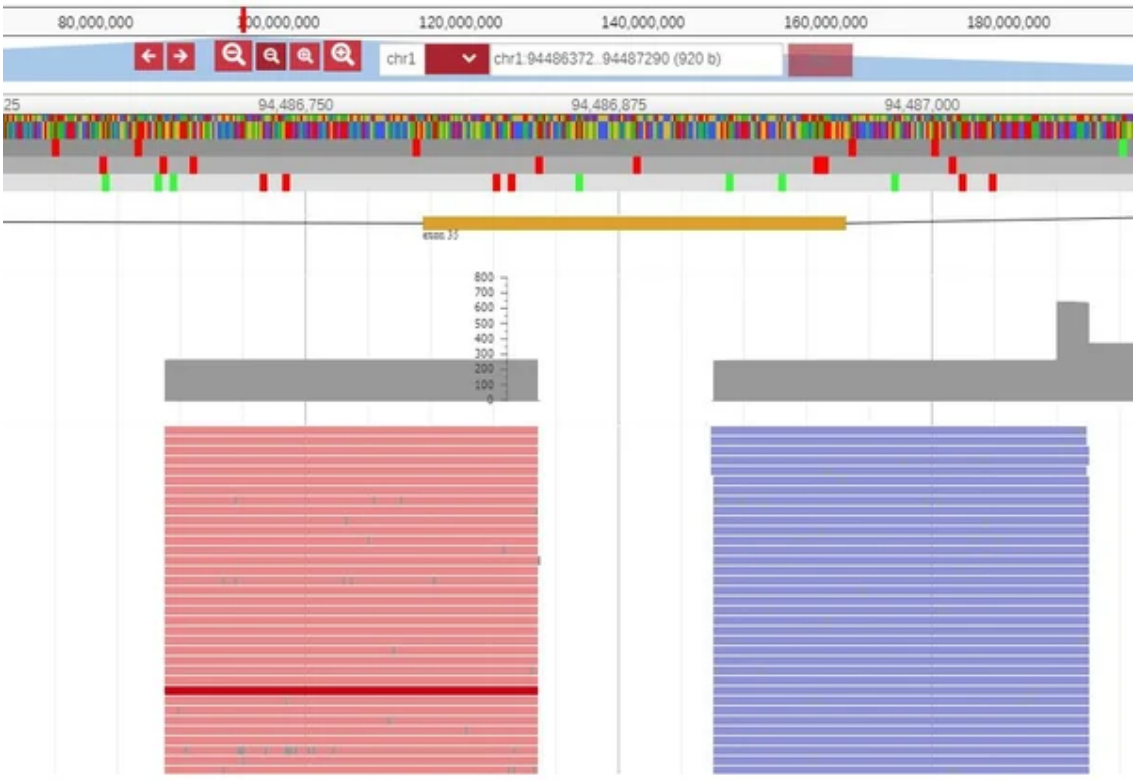
Clicking on that option will bring up a popup where you can choose a gene list and then download a coverage report for the genes in that list. Please note that you will need to create the gene list before clicking on this option. To do so, click on Gene Lists in the top bar:



Then, click on Create new gene list and enter your genes of interest and save it with a name of your choosing. You will now be able to generate coding coverage reports using that list. This file has a breakdown of coverage by exon. Although it won't directly pinpoint gaps, it will at least show the minimum coverage per exon, so if that is 0 you know there's a gap somewhere and can then use the Gene Coverage function to locate the target exon:



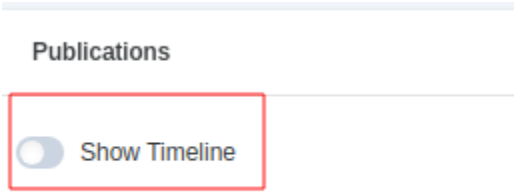
Subsequently, click on the exon and so open the JBrowse window that shows the actual reads at that position:



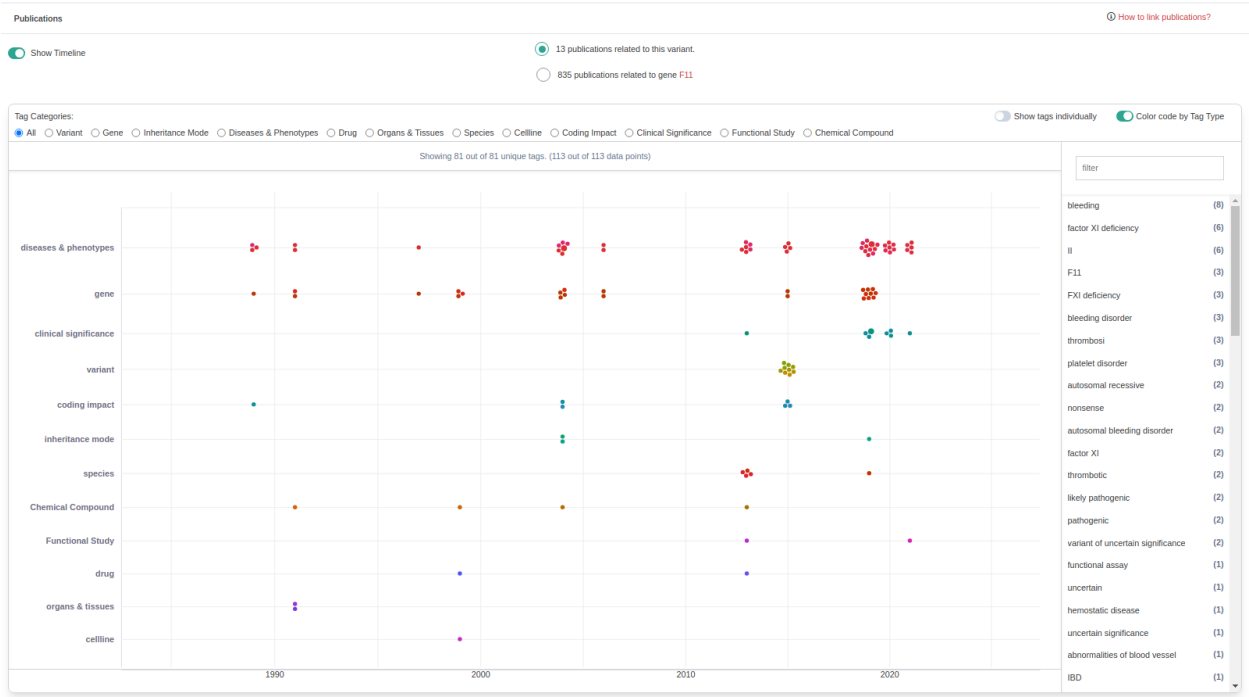
12.11 Publications Timeline Visualization

This interactive viewer allows you to quickly see the number of publications by year, broken down by gene, disease, variants etc. with the ability to then further drill-down and identify specific publications related to the topic you are researching.It is possible to drill-down into sub-categories of terms, for example to see what diseases are associated with a particular gene:

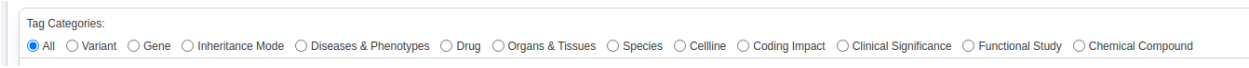
To use the publications Tag Timeline. Click on the link on the top left hand side of the Publications section.



This opens the new Publications Timeline feature.

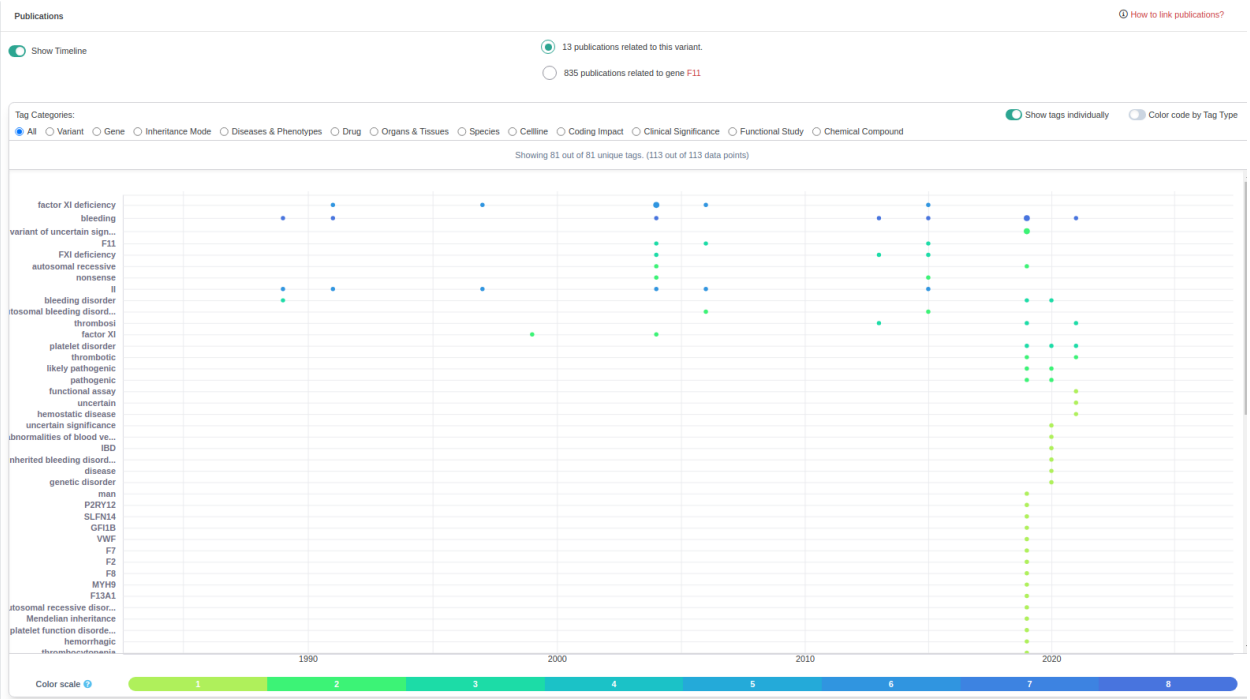


On the top bar there are a number of tools you can use to amend the data presentation of the viewer.



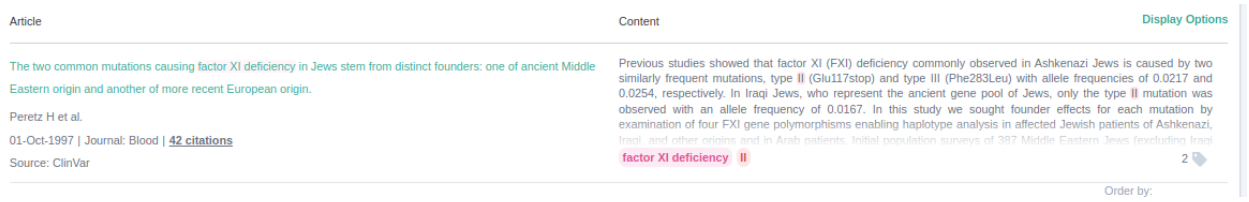
The viewer defaults to showing all tags. It is possible to select only required tag categories by selecting the relevant tag radio buttons.

You can also select to show tags individually. In this way you can clearly visualize the number of publications associated with each tag.



Hovering over a bubble tells you how many publications are linked

Clicking on a “bubble” will then display the list of associated publications.

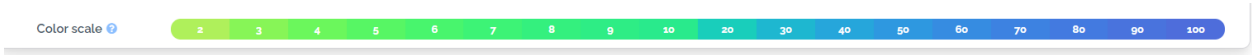


You can use the “color code by tag type” in the following way:

In the initial setting the “bubbles” reflect the same color as the tags in the publications component. For example a green Disease tag in a publication will be represented by a green disease “bubble” and it is clear which disease is mentioned most often.

You can also switch to using a color scale

In this case the color of the “bubbles” represents a scale of the number of publications that mention a tag overall and is depicted at the bottom of the screen.



The size of the “bubbles” represents the number of publications mentioning a certain tag per year. So the larger the bubble the more publications there are in a given year mentioning the specific tag.

12.12 Joint calling

Implementation

When running a multi-sample analysis in VarSome Clinical, we use joint calling for variants. The variant caller will perform a local realignment of all of the samples' reads around the target region being investigated and then will try to identify the haplotypes supported by the resulting pileup. This way the caller uses the information of all samples to boost the evidence of each variant site.

Use an existing analysis for joint calling

When selecting the option "use an existing analysis", if all of the component analyses were initially launched from FASTQ, joint calling is performed as if re-using the FASTQ files from scratch. The only case where joint calling cannot be implemented is if any of the component analyses were launched starting from a VCF file, for which we only merge the VCF results.

12.13 Variant calling and quality filters

Variant calling is the process by which a software program (the variant caller) identifies variants from sequence data. For single and multi germline samples sequenced using hybridization-based capture kits, we use Sentieon's DNA Scope variant caller and for single and tumor/normal somatic samples, we use the Sentieon TNhaplotyper2 algorithm. On the other hand, for amplicon kits on both germline and somatic samples VarSome Clinical uses VarDict.

Depending on the type of assay used to sequence the sample and on the sample itself, VarSome Clinical uses different variant callers.

VarDict is a sensitive variant caller and is especially well-suited to amplicon samples. VarDict implements several novel features such as strand bias aware variant calling from targeted sequencing experiments.

Default workflow

1. Adapter trimming is normally done before the FASTQ file is generated. The FASTQ files VarSome gets are therefore already trimmed.
2. Read mapping
3. Primer clipping (amplicon-based samples)
4. Deduplication
5. Indel realignment
6. Base-quality recalibration
7. Variant calling

Call status

There are different quality metrics associated with each variant which can be used in subsequent steps of the pipeline to assign it a call status. The call status of a variant can be:

- PASS: all the quality metrics are above the thresholds (i.e. the variant has passed all quality filters).
- FAIL: the variant has not passed all quality filters.

Quality filters

The quality filters used for germline and somatic analyses are different since we use different variant callers, which, in turn, use different parametrization. The parameters used in Sentieon have been optimized for the detection of variants from the GIAB set. In VarDict we currently use a set of minimally changed default parameters, adjusted through exchange with VarSome Clinical users.

Quality filters for capture kit samples

- **Germline analyses**

Sentieon's DNAScope

This caller is used for germline capture kit samples and performs an improved version of GATK Haplotype variant calling. We apply the following quality filters after the variant calling step:

- **Coverage**: number of reads aligned against the variant position. The minimum coverage for capture kit samples is 8; all variants with coverage lower than 8 reads will be considered as FAIL.
- **Quality**: the quality score is an internal score calculated by the variant caller algorithm. It can be used to estimate how confident we are that the variant caller has correctly identified a variation in a given genomic position.
- **Single sample analyses**: we assign a FAIL call status to the variants having a QUAL lower than 100 in single sample analyses. The QUAL is the Phred-scaled probability that a REF/ALT polymorphism exists at this site given the sequencing data.
- **Multisample analyses** (couple, family trio or generic multisample): we use the GQ (genotype quality) which represents the Phred-scaled confidence that the genotype assignment (GT) is correct. All variants with a GQ lower than 20 will be marked as FAIL. Please bear in mind that the GQ is associated with each sample. For example, a variant called in a trio analysis will have three different GQs, one per each sample. The variant might have a GQ below the threshold in one of the samples while having a GQ above of it in the other samples. In that case, the variant will be marked as "Failed/Not genotyped" in the sample where it had a low GQ and PASS in the others.

- **Somatic analyses**

Sentieon's Tnhaplotyper2

Tnhaplotyper2, which is used for somatic capture kit samples, is designed to behave like GATK's Mutect2. Tnhaplotyper2, like mutect2, has associated filtering tools which are applied to the

variants found by the caller. These filters can then be used to decide whether a variant should be marked as PASS or FAIL. If a variant fails any of the filters present in the “FAIL” column of the table below, it will be marked as FAIL. Failing to pass a filter in the “PASS” column will not cause the variant to be marked as FAIL.

PASS	**FAIL**
clustered_events	map_qual
duplicate	base_qual
fragment	contamination
multiallelic	weak_evidence
n_ratio	low_allele_frac
orientation	normal_artifact
position	panel_of_normals
slippage	strand_bias
haplotype	
germline	
strict_strand	

Somatic VCF filters that do not mark a variant as FAIL:

- clustered_events: multiple events are present on the same haplotype as the variant which is indicative of a false-positive call.
- duplicate: the alternate allele is overrepresented by apparent sequencing duplicates.
- fragment: a large difference is observed in the median fragment length for reads supporting the reference and alternate alleles.
- multiallelic: the mutation occurs at a multiallelic site.
- n_ratio: too many 'N' bases at the site.
- orientation: the variant is likely an artifact due to orientation bias.
- position: the allele is close to the ends of the reads.
- slippage: the variant is likely an artifact due to polymerase slippage.
- haplotype: variant is on the same haplotype as other filtered variants.
- germline: there is evidence that the variant is germline.
- strict_strand: evidence for the alternate allele is not significant in both directions.

Somatic VCF filters that mark a variant as FAIL:

- map_qual: the median mapping quality of reads supporting the alternate allele is too low.

- base_qual: the median base quality of bases supporting the alternate allele is too low.
- contamination: the alternate allele is present due to contamination.
- weak_evidence: the mutation does not have significant support above noise.
- low_allele_frac: the variant allele fraction is below the threshold.
- normal_artifact: the variant is likely an artifact in the normal sample.
- panel_of_normals: the site is present in the panel of normals.
- strand_bias: evidence for the alternate allele comes from only one read direction.

Quality filters for amplicon kit samples
Both somatic and germline amplicon kit samples are analyzed using VarDict with different quality thresholds.

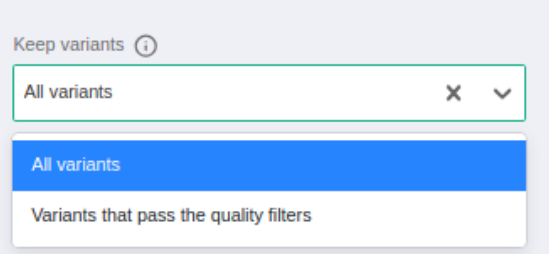
- **Germline analyses**

The Allelic Balance (AB) cutoff is set to 0.2. However this rule applies only for positions covered by more than 100 reads. Otherwise the variant is reported only if the AB is < 20/coverage depth. (I.e. the call will not be made if the variant is supported by less than 20 reads.)

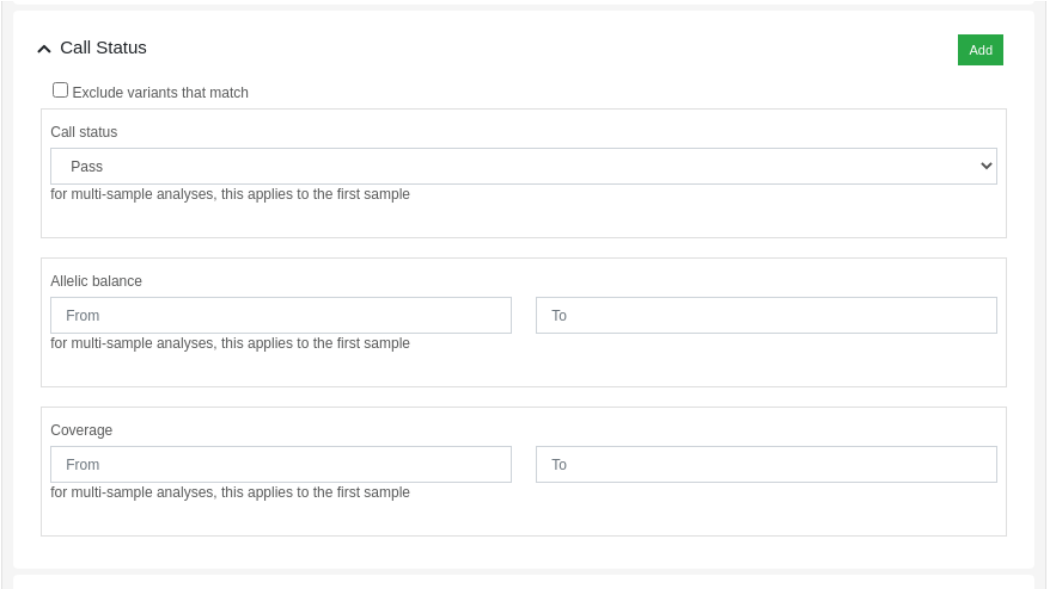
- **Somatic analyses**

The Allelic Balance cutoff is 0.005. However this rule applies only for positions covered by more than 400 reads. Otherwise the variant is reported only if the AB is < 20/coverage depth. (I.e. the call will not be made if the variant is supported by less than 20 reads.)

Call status variant filtering
When launching a germline/somatic analysis from FASTQ, the user will have two options:
- **All variants**: the variant table will contain all variants called by the variant caller including both variants with PASS and FAIL call status.
- **Variants that pass the quality filters**: the variant table will contain only variants having a call status of PASS.



Variants can be filtered by its call status using the dynamic filters feature. The "Call Status" filter allows the user to filter variants based on the following criteria:
- Call Status: PASS, FAIL or anything.
- Allelic balance: proportion of reads supporting the alternative allele.
- Coverage: number of aligned reads against the variant position.



12.13.1 Variant calling local reassembly

Sometimes, the variants reported in the variant table are not visible in the alignment displayed through IGV/JBrowse. The reason lies in an intermediate step of the variant calling process called local reassembly.

Variant calling algorithms

As we mentioned above, we use Sentieon variant caller algorithms to perform the variant calling procedure. Sentieon is a toolkit analogous to GATK (The Genome Analysis Toolkit) but built on a highly optimized backend (Kendig et al., 2019). For calling germline variants we use Sentieon DNAscope and for calling somatic variants we use Sentieon TNHaplotyper2.

Sentieon & GATK local reassembly

Sentieon and GATK use a procedure to re-assemble read data and determine candidate haplotypes as a prelude to variant calling. This procedure, named local reassembly, is a middle step where the program first builds an intermediate alignment and the reads are locally aligned. The algorithm calls variants during this middle step and assigns the variants found values such as allele frequency or read-depth.

⚠️Please note that sometimes, for some particular genomic regions, the whole sequence alignment does not match with the intermediate alignment constructed during the local reassembly. This is the reason why some variants reported in the variant table are not visible when opening the alignment using IGV or JBrowse.

Purpose of performing local reassembly during the variant calling

Local reassembly based methods are less dependent on prior mapping of sequence reads for variant calling and as a result have higher sensitivity and specificity in indel calling (Rimmer et al., 2014).

Can I access the intermediate alignment?

No, the intermediate alignment file is not available for download or visualization. This file is temporarily created during the variant calling process by the algorithm. This is the reason why it is not displayed together with the other alignment in the results.

Should I consider the variant a true positive or an artifact?

A variant found under these circumstances should be treated like any other variant called by the variant calling pipeline. All of them are candidates of being true positives and other factors should be taken into consideration to evaluate whether a variant is real (e.g. quality of the call, read depth, allele frequency...). Our validated variant calling pipeline has demonstrated to have a high sensitivity and specificity, however, as any other NGS pipeline, there might be false positives. Even if the variant is not a variant calling or mapping artifact, there might be other type of artifacts (sequencing errors) that should not be overlooked.

12.14 CNVs

12.14.1 Length of CNVs

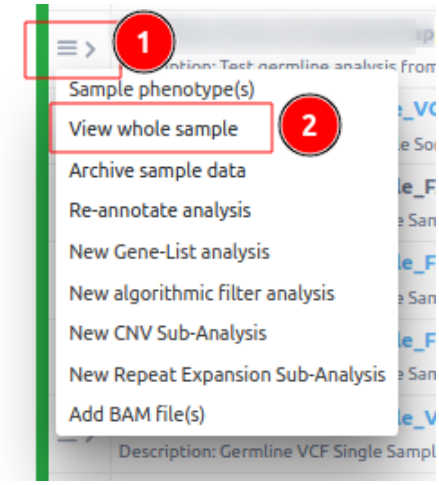
There is no specific minimum size of CNV that can be detected by VarSome Clinical. The length of CNV calls depends on the length of the target genomic intervals of the assay used to sequence the sample. Such intervals are encoded in the form of a .bed file and are assay specific. In whole-exome sequencing (WES) assays, typically each interval is an exon, however assays may also target additional, much shorter regions that are a few bases long. In gene panel assays, the length of regions also varies.

VarSome Clinical's CNV calling pipeline is read depth based, so higher read counts generate more robust CNV signals. On the other hand, low read counts may result in untrustworthy calls. VarSome Clinical checks the coverage and informs the user if it is too low in the call region (Call Quality column). User can also visually inspect the relevant region using the provided links to IGV/JBrowse. Detailed sample coverage analysis is available in the Coverage Report (available in the Analysis Actions menu).

12.14.2 Gene list analysis for CNV calling

In gene list analyzes you can only access by default the results for the targeted genes and not the full parent analysis that is needed for the cohort of a CNV/SV analysis.
The primary reason we don't show the full results when a user has run a gene list analysis and we only show the results of the gene list, is to ensure we don't show incidental findings. In a clinical setting, it is sometimes required that there be no way of seeing any information other than what the target regions include. So, reporting a variant that falls outside the target genes might reveal something that wasn't intended to be known for legal and/or ethical reasons. This is why we don't allow such analyses to be included in CNV calling since that might find CNVs outside the region of interest and can cause issues.

If you know this is not a problem, you can choose to show the variants of the gene list's parent analysis:



Once you have done that, the analyses will be available for CNV calling. To avoid this, you can choose not to run a gene list analysis in the beginning and, instead, run a full analysis and then, once that is finished, run a gene list analysis separately. This will keep both analyses (full and gene list) visible to the user, making the full analysis available for CNV calling by default.

12.14.3 Can I reuse a standard set of control samples to call CNVs?

It is recommended that all samples processed in VarSome Clinical CNV analyses have been sequenced by the same instrument, experimental methodology, in the same batch. However, for practical and cost-related reasons, it may not be possible to repeatedly sequence control samples in every batch, or each time one or more test samples require sequencing. In that case, it is possible to re-use a set of control samples that were sequenced previously. The resulting CNV calls are likely to be less accurate.

In-house benchmarking studies have attempted to quantify this loss in performance using a dataset of 96 samples (66 samples with CNV calls, 30 control). Comparison of a complete CNV run to multiple runs using a fixed limited set of 10 control samples as a pool to select references, showed a single digit drop in sensitivity and a small number of changes in CNV limits/breakpoints.

It is important to always check which samples have been selected as reference in the analysis QC report, and their count, available from Analysis Actions Menu. This is because samples of the reference set should not contain the same CNV as the queried test sample, otherwise such CNV will not be called. Ideally, there should be 5-10 reference samples of high correlation (>0.97 for panels, >0.98 for exomes) for each test sample analyzed. For further detail in assessing CNV run and call quality, please see the sections CNV variant table and Quality Control report for CNV analyses.

12.14.4 How is frequency calculated for CNV calls?

We use the gnomAD database to get the population frequencies for a given CNV. Depending on the type of variant, the frequencies are calculated as follows:

- Deletions: we use gnomAD variants if they fully overlap with the given variants.
- Duplications in coding regions: we compare at the gene level and we use those gnomAD variants that encompass the same coding genes as the given variant.
- Duplications in non-coding regions: we use gnomAD variants if they cover at least 85% of the variant region.

Why are frequencies calculated differently for gains vs losses?

GnomAD reports common structural variants; currently, we retrieve information only for deletions and duplications.

- For deletions, we consider the gnomAD CNVs when they fully encompass a sample's annotated CNV. Even if the gnomAD reported deletion is larger than the sample's CNV, we can assume that the sample's CNV is contained in the gnomAD population. This way, if the gnomAD CNV is reported as a benign loss, then the sample's CNV will be most likely benign as it is contained in that region. Conversely, if the sample's CNV overlaps with a gnomAD pathogenic CNV, then the sample's CNV will be most likely pathogenic as it contains a region whose loss results in pathogenicity.
- For duplications, we differentiate our frequency calculation approach based on the genomic location of a CNV. In case of duplications found at coding regions, we compare the sample's CNVs and gnomAD CNVs at gene level. We will consider that both CNVs may have an equivalent effect only if they encompass the same coding genes.

12.14.5 CNV calling with WES or targeted panel data

VarSome Clinical currently offers Copy Number Variation (CNV) analysis for both Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) or targeted panel data. For the non-WGS analyses we use ExomeDepth, a CNV caller based on a read depth approach. To accurately detect CNVs, ExomeDepth requires at least five samples (ideally around ten) that will be run as a cohort with each sample analyzed, using the rest as a pool to select reference samples.

Why are reference samples necessary in CNV calling for non-WGS samples?
The read depth approaches used for CNV calling in WGS usually assume that reads distribute in a more or less uniform way across the genome and, therefore, the differences in read depth are used to identify CNVs. However, this assumption fails in the context of WES and targeted sequencing. One of the main reasons is that the probes used for capturing the different targeted regions have variable specificity and efficiency depending on the region. This fact introduces strong biases in the number of mapped reads per region that hamper the CNV detection. ExomeDepth requires multiple samples because it uses them to control the biases given by the extensive variability in capture efficiency across exons and/or target areas.

What characteristics must the reference samples have?

For optimal results, the reference set of samples must have the following characteristics in common with the test sample (sample of interest):

- Samples should be prepared with the same library protocol and sequenced by the same sequencing platform.
- All samples (the test one and the reference) should all have been generated as part of the same sequencing batch. It is possible to use samples generated in different batches but the resulting CNV calls are likely to be much less accurate.
- Samples should originate from individuals unrelated to each other. For example, if samples come from the same family, related individuals should be excluded.
- For CNV calls in sex chromosomes, all samples should be of the same sex (either all male or all female). If they are not all of the same sex, calls on those chromosomes will not be reliable.

How are reference samples used by ExomeDepth

Each sample given as input for ExomeDepth analysis will be taken to call CNVs on it by using a selection of the remaining as reference samples. This means that, when running a CNV analysis with ExomeDepth, you will get calls for all of the input samples, no matter if you consider the sample as a test or reference.

Another important key to bear in mind is that every input sample might not be compared to all other samples. Each input sample is compared against an optimized set of reference samples that are well correlated with it. The first step of the CNV calling process is to construct the reference set of samples. To do this, ExomeDepth takes one of the input samples and ranks the remaining by order of coverage correlation with the first sample. Then, the remaining samples are sequentially added to the reference set. After the addition of one sample to the reference set, a statistical calculation is performed to see how good the current reference set is to predict CNVs on the test sample. The addition of samples to the reference set stops when it is unable to improve the reference set power to predict CNVs. Therefore, using a high number of samples for CNV analysis does not necessarily increase the accuracy of the results because:

- Not all available samples are included in the reference set, which means that not all the samples are used as reference for calling CNVs in the test sample.
- Some of the CNVs present in the test sample can be missed if they are shared between the test and the reference samples.

The reference selection process is automated in the analytical pipeline implemented in VarSome Clinical and does not require any additional steps by the user. VarSome Clinical SV pipeline will analyze all samples of the cohort successively and generate CNV calls for all.

⚠️ Please note that ExomeDepth authors estimate that the optimum size of the reference set is ~10. Adding further samples in the reference set actually might decrease the power.

12.15 Why does the position of some indels reported not match the position of the indels reported by ExAC/Gnomad?

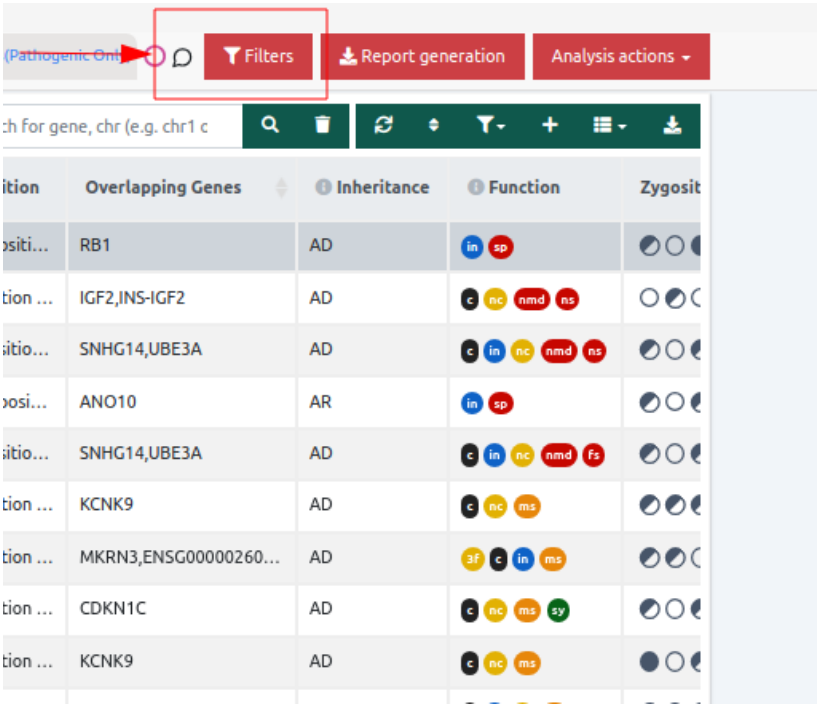
The same sequence can be the result of multiple insertions or deletions. In cases of such equivalent indels, one version may have been called by the variant caller but a different one found ExAC or other databases.

For example, given a reference sequence ATATATATATATATATATATATATATATATATAT. Now suppose an insertion of an additional AT is found in the beginning of the sequence while ExAC has an insertion of an extra AT at the end of that sequence with a frequency of 10%. The two insertions are in fact equivalent because the resulting sequence is the same. Therefore, our system will match the insertion it found to the one in ExAC.

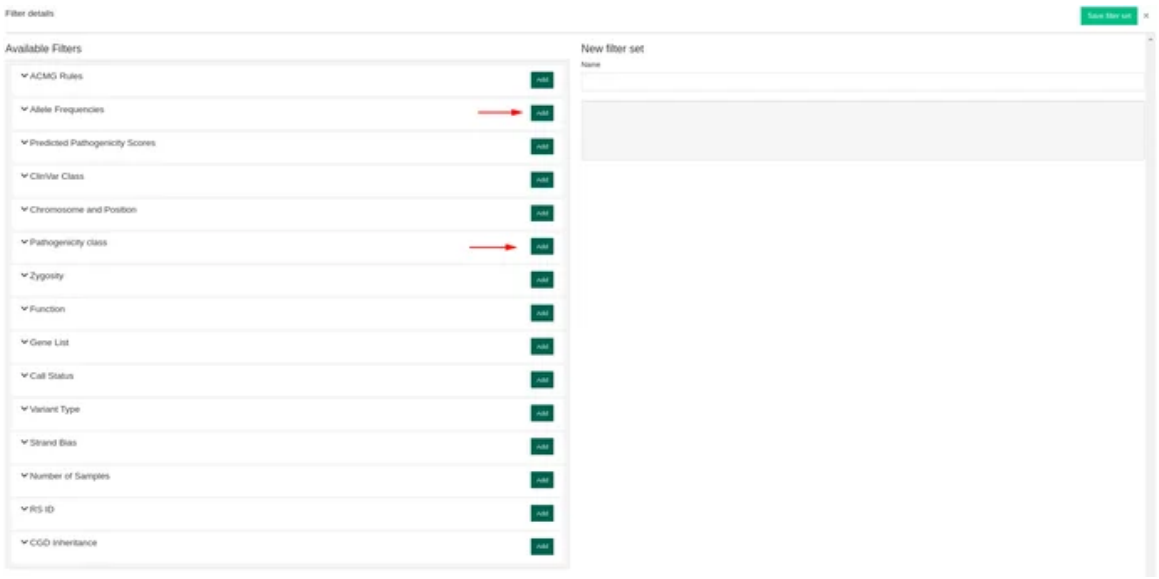
When we inspect the read alignments however, the insertion may not be seen in the same place as shown on our system. Therefore, our algorithm carefully and correctly matches indels to the frequency of their equivalent found in Gnomad and other databases.

12.16 Filter by Frequency or Pathogenicity

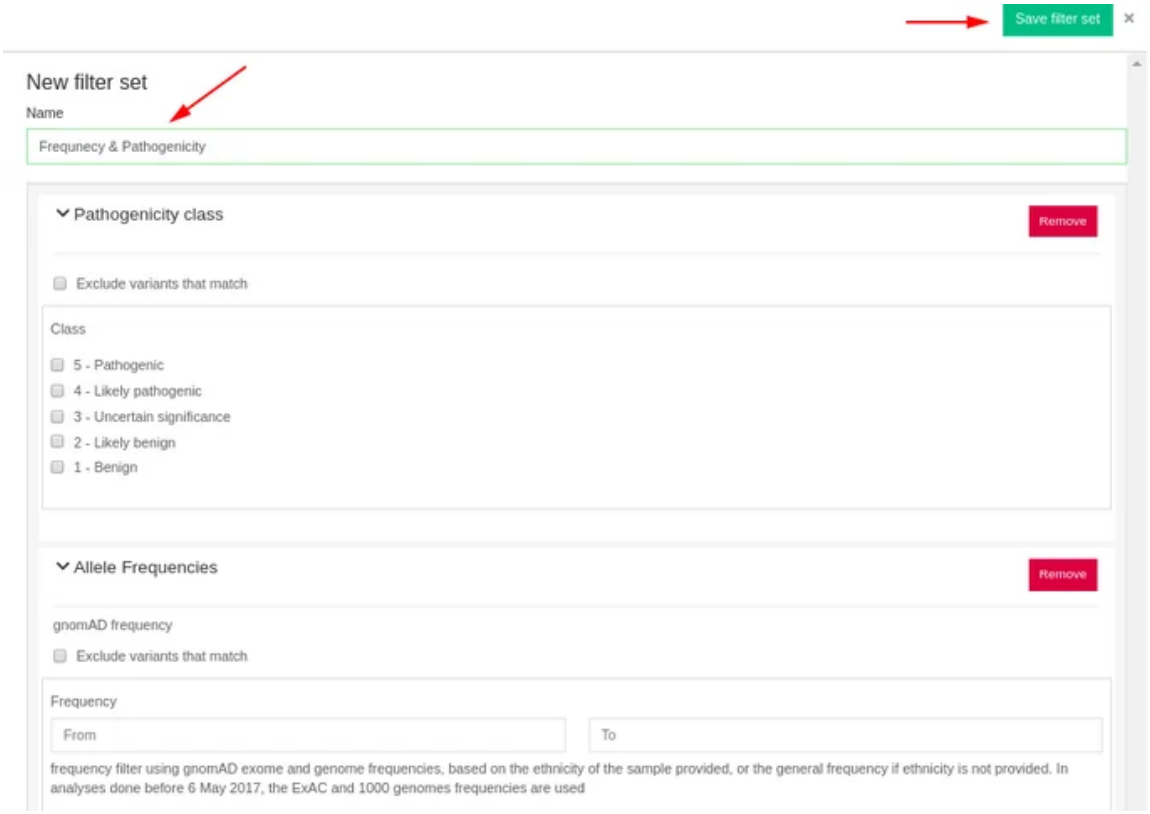
After an analysis is complete, your results are available at the Dashboard. By clicking on the name of your analysis, you will see the Variant Table:



Once you click on the Filters red button, you will get a screen with dynamic filters:



In order to filter for specific categories (e.g. Allele Frequency and Pathogenicity), you could click on the filter icon and then “Create” and a pop-up screen will appear.



After you have completed a name for the Filter Set you have just created, clicking on “save” will apply the filter set to your results.

12.17 Hemizygous Variants

Hemizygous variants are the variants that fall within regions without allelic counterparts and

pertaining to a diploid cell with only one copy of a gene instead of the usual two copies.

Are those that are X-linked and are not in the pseudoautosomal regions according to this ref https://en.wikipedia.org/wiki/Pseudoautosomal_region:

- The pseudoautosomal regions get their name because any genes within them (so far at least 29 have been found) [ref] are inherited just like any autosomal genes. PAR1 comprises 2.6 Mbp of the short-arm tips of both X and Y chromosomes in humans and great apes (X and Y are 155 Mbp and 59 Mbp in total). PAR2 is at the tips of the long arms, spanning 320 kbp [ref].

An example of an X-linked variant within a pseudoautosomal region can be found on https://varsome.com/variant/hg19/chrX%3A38145358%3AT%3A (gene RPGR).
VarSome Clinical would call the variant above as heterozygous. In this case, VarSome Clinical's Sentieon/GATK variant caller does not distinguish between heterozygous and hemizygous terms. However, since the difference between the terms hemizygous and heterozygous is largely semantic, we feel the variant is still called correctly.

For more Frequently Asked Questions about the platform please see also the following doc file https://docs.varsome.com/en/faq.

13. Troubleshooting

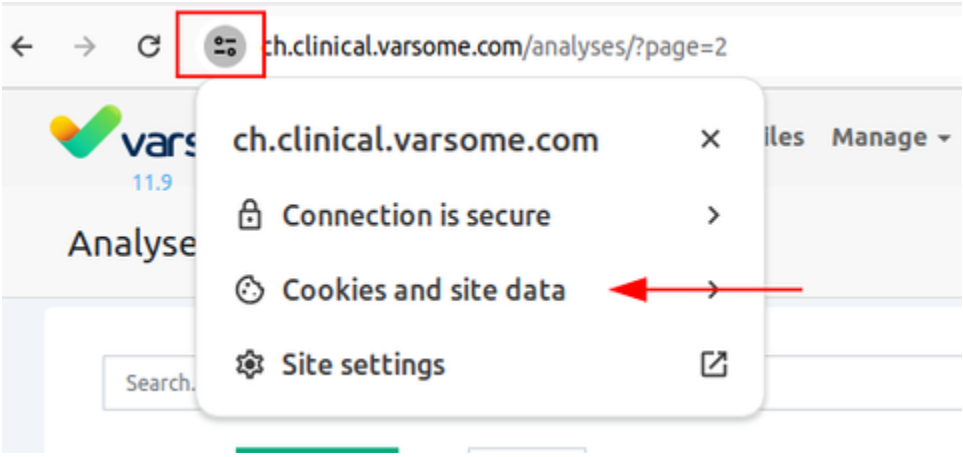
13.1 Clear VarSome Cookies from the browser

The procedure for Firefox is the same as described below. For Microsoft Edge, please see here and for Safari, please see here.

Sometimes there is a need to clear the cookies and the cache from the browser in order to visualize the variant table.

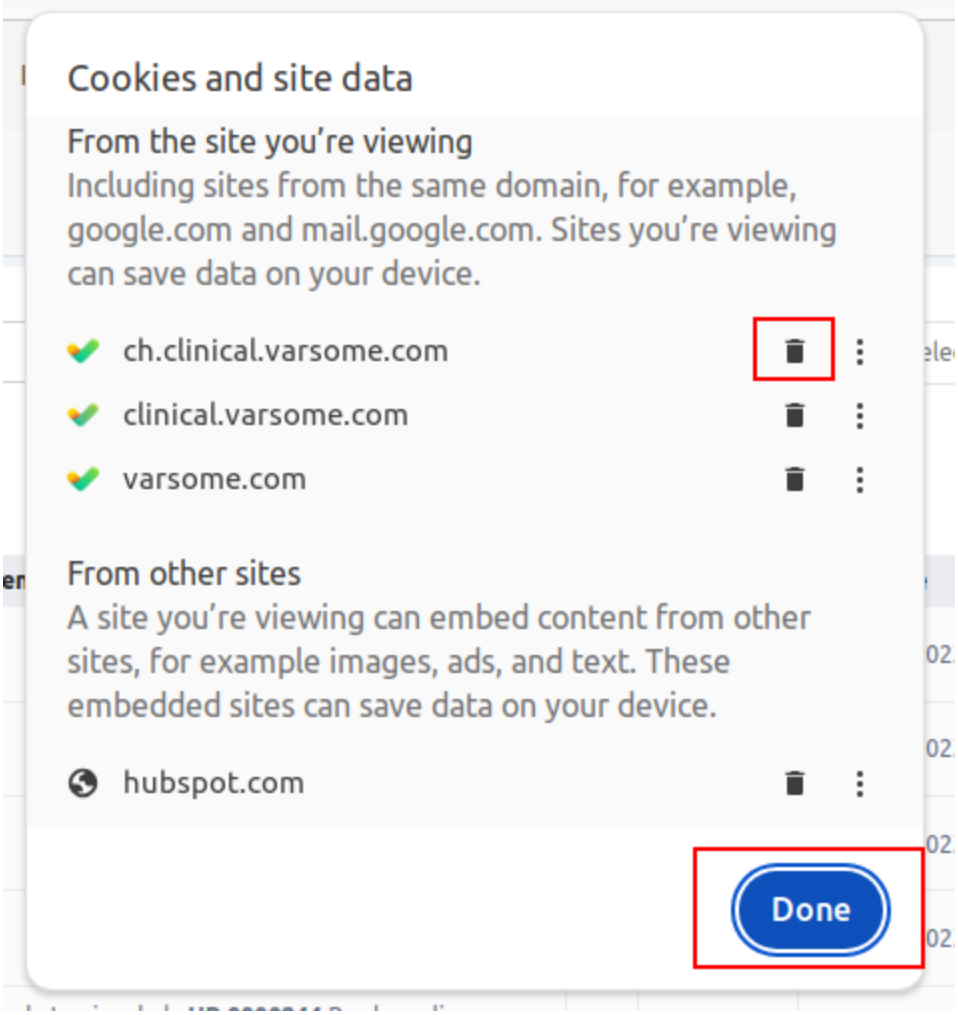
Google Chrome

On the VarSome Clinical webpage, there is a ***lock icon*** on the left of the URL. Clicking on it displays a drop-down menu. After clicking on Cookies, a new window will be displayed.



It is possible to select which cookies to remove from **VarSome.com** and/or from **VarSome Clinical**.

We recommend removing both, and then click done.



Once finished, there is the need to refresh the page which was not possible to visualize the data properly.