SeqOne platform

RUO product

For Research Use Only.
Not for use in diagnostic procedures

Valid from SeqOne Platform v2.1

1. Introduction

You are advised to read the instructions carefully before using the software. A paper version of these instructions can be provided upon request to the SeqOne support team (support@seqone.com). To achieve the best results when using the platform, carefully read all the precautions highlighted with the ⚠ symbol. The various regulatory symbols and other symbols used throughout this guide and the software are as follows:

Symbol	Meaning
	Manufacturer
	Manufacturing date (Software Release Date)
	Consult the instructions for use
	"Warning" This safety symbol indicates that special care is required when using the device.
	"Research use only" means the device is intended for research purposes only, not for use in diagnostic procedures Consequently, the use of this functionality as a diagnostic aid is the responsibility of the healthcare professional.

The label (including SeqOne software version information and contact details) and this document describing instructions for use are available from the "About" section.

2. Contact manufacturer

SeqOne S.A.S
Web: https://seqone.com/
22 rue Durand
34000 Montpellier
FRANCE

User support :
support@seqone.com

3. Description of the platform

3.1. Introduction

SeqOne Platform is a research use product that integrates a suite of applications working in concert to provide expert users with a powerful toolkit for interpreting complex genomic and sequencing data.

SeqOne Platform integrates established information from a wide range of **external sources**, including public databases, scientific literature, and proprietary knowledge bases. These external data sources provide context and meaning to the identified genomic variants.

Input files for various applications within the platform are processed using a tailored combination of Bioinformatic and Annotation modules, ensuring that each analysis is optimized for its specific requirements.

Bioinformatic Modules conduct secondary analysis using computational methods such as alignment and variant calling to identify genetic variants in DNA and RNA sequencing data by comparing them to a reference genome (hg19 or hg38). The data, which can be used for both germline and somatic analysis, can be generated from whole genome sequencing, exome sequencing, or targeted gene panels.

Annotation Modules conduct tertiary analysis to annotate variants with details such as the gene and its known function, the location of the variant within the gene, and any publications or clinical reports that mention the gene or variant. This annotation will determine which ACMG and AMP/CAP/ASCO classification rules apply to each variant, and from there, a pathogenicity verdict can be reached. The annotation is also performed through DiagAI, a suite of AI-powered tools that helps analyze genetic variants to quickly identify potential causative variants.

This modular architecture allows SeqOne to be highly adaptable to different types of analyses

and research questions. Input files for various applications within the platform are processed using a tailored combination of these Bioinformatic and Annotation modules, ensuring that each analysis is optimized for its specific requirements. By integrating these diverse data sources and analysis tools, SeqOne empowers expert users to gain deep insights into the complex landscape of genomic data and its implications for human health and disease.

The platform operates as a cloud-based SaaS solution, with all features accessible through a web browser. It includes SeqOne Flow, an optional tool that can be deployed either in the AWS cloud or on customer servers to facilitate data transfer and workflow integration.

For security, all data is encrypted at rest using AES-256 encryption, and traffic between users and the platform is secured using SSL certificates.

3.2. Characteristics

Each Application (workset) consists of multiple **Bioinformatics and annotation modules**, each responsible for specific analytical tasks. In this chapter, the performances are evaluated at the module level to ensure a detailed assessment of each component's accuracy and reliability.

The analytes detected by the SeqOne Platform fall into two categories of genomic alterations:

- Genetic variants such as SNVs, indels, MNVs, CNVs, UPD, gene fusions, structural variants, splicing variants, Alu insertions, and FLT3-ITD (Internal Tandem Duplication).
- Signatures or complex biomarkers such as IGHV status, Microsatellite Site Instability (MSI), Tumor Mutational Burden (TMB), methylation or double strand break repair Defects via Homologous Recombination (HRD).

Bioinformatics module	Workset	Analyte type
CNV Detection Exome Germline	CNVExome, GermVar, GermVar Family	Genetic variant
CNV Detection Panel Germline	CNVCapture	Genetic variant
SNV/Indel Germline	GermlineVar, GermlineFamily	Genetic variant
SV Germline	GermlineVar, GermlineFamily	Genetic variant
UPD	GermVar, GermVar Family, GermVar Tertiary, GermVar Tertiary Family	Genetic variant
Alu Insertions	GermlineVar, GermlineFamily, SomaVar, Somavar LF, SomaCGP, SomaHemato, GermVar, GermVar Family	Genetic variant
MNV	GermlineVar, GermlineFamily, GermVar, GermVar Family, GermVar Tertiary, GermVar Family Tertiary, SomaVar, SomaVar LF, SomaCGP, SomaHemato, SomaLBx, SomaVar Tertiary	Genetic variant
SNV/Indel Somatic	SomaVar, SomaHemato, SomaRNA, SomaCGP	Genetic variant
SNV/Indel Somatic LF	SomaVar LF	Genetic variant
CNV Large Somatic Panel	SomaCGP	Genetic variant
Methylation Index	SomaMethyl	Signature
Somatic DNA fusion	SomaCGP, SomaVar, SomaVar LF, SomaHemato	Genetic variant
Somatic LBx DNA fusion	SomaLBx	Genetic variant
TMB	SomaCGP	Signature
MSI	SomaCGP, SomaMSI	Signature
GI probability (GiP)	SomaHRD	Signature
Sentieon SNV/Indel Germline	GermVar, GermVar Family	Genetic variant
Mosaic variants Germline	GermVar, GermVar Family	Genetic variant
CNV WGS	GermVar, GermVar Family	Genetic variant
Sentieon SNV/Indel Somatic	SomaLBX	Genetic variant
Fusion RNA	SomaRNA	Genetic variant
Splicing RNA	SomaRNA	Genetic variant
CNV Small Somatic Panel	SomaCNVcapture	Genetic variant
IGHV status	SomaHemato	Signature
IGHV subsets	SomaHemato	Signature
CFH	GermlineVarLR	Genetic variant
FLT3	SomaVar, SomaHemato	Genetic variant

3.2.1. Analytical performance

The performance claimed in this IFU is the performance of the latest IVD versions of the worksets, which are listed in the table below:

Workset Name	Version
CNVCapture	2.6
CNVExome	2.2
GermlineVar	2.6
GermlineFamily	2.6
GermlineVar LR	1.2
GermVar	3.12
GermVar Family	3.12
GermVar Tertiary	1.16
GermVar Tertiary Family	1.14
SomaCGP	1.6
SomaCNVcapture	2.4
SomaHemato	0.6
SomaHRD	1.5
SomaLBx	0.7
SomaMethyl	0.3
SomaMSI	1.1
SomaRNA	3.3
SomaVar	2.9
SomaVar LF	2.1
SomaVar Tertiary	2.4

3.2.1.1. Bioinformatic modules

- **CNV Detection Exome Germline and Panel Germline**

Variant type	Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
CNV Detection **Exome** Germline (GermlineVar)	GRCh37	8	8	0	100%	**100%**
CNV Detection **Exome** Germline (GermVar)	GRCh37	8	8	0	100%	**100%**
CNV Detection **Panel** Germline	GRCh37	15	15	0	100%	**100%**
CNV Detection **Panel** Germline	GRCh38	11	11	0	100%	**100%**

- **SNV / Indel Germline - Panel data**

Genome Version	Variant type	Acceptance Criteria	Sensitivity before filter	Sensitivity after filter
GRCh37	SNV	100%	98%	**100%**
GRCh38	SNV	100%	98%	**100%**

Genome Version	Variant type	Acceptance Criteria	Precision before filter	Precision after filter
GRCh37	SNV	100%	90%	**100%**
GRCh38	SNV	100%	90%	**100%**

- **SNV / Indel Germline - Exome data**

Genome Version	Variant type	Acceptance Criteria*	Sensitivity before filter	Sensitivity after filter
GRCh37	**SNV**	**> 98%**	**98.7%**	**98.5%**
	Indel	**> 80%**	**88.0%**	**87.5%**
GRCh38	**SNV**	**> 98%**	**98.5%**	**98.3%**
	Indel	**> 80%**	**84.9%**	**84.3%**

**for the detection of variants located in the high confidence regions of NA12878 and NA24385.*

Genome Version	Variant type	Acceptance Criteria*	Precision before filter	Precision after filter
GRCh37	**SNV**	**> 90%**	**92.1%**	**94.4%**
	Indel	**> 80%**	**84.7%**	**89.0%**
GRCh38	**SNV**	**> 90%**	**94.2%**	**95.4%**
	Indel	**> 80%**	**82.2%**	**87.6%**

**for the detection of variants located in the high confidence regions of NA12878 and NA24385.*

- **SNV / Indel Germline - Repeatability and reproducibility**
    - **Repeatability**

Variant Type	Sensitivity	Precision
SNV	98,59% ± 0%	95,92% ± 0%
indel	87,75% ± 0%	81,63% ± 0%
all	**98,38% ± 0%**	**95,63 ± 0%**

    - **Reproducibility**

Variant type	Sensitivity	Precision
SNV	98,57% ± 0,03%	95,93% ± 0,08%
indel	86,58% ± 0,42%	78,97% ± 0,96%
all	**98,34% ± 0,03%**	**95,58 ± 0,09%**

- **Mosaic variants Germline**
    - **Sensitivity SNV**

Genome Version	VAF bins	# true variants	Acceptance Criteria	Sensitivity 100x
GRCh38	[0.05,0.1[	3704	40%	47.6%
GRCh38	[0.1,0.2[	747	70%	73.1%
GRCh38	[0.2,0.3[	685	90%	90.7%
GRCh38	[0.3,1.0[	76	90%	93.4%

    - **Sensitivity Indels**

Genome Version	VAF bins	# true variants	Acceptance Criteria	Sensitivity 100x
GRCh38	[0.05,0.1[	87	40%	43.7%
GRCh38	[0.1,0.2[	14	70%	78.6%
GRCh38	[0.2,0.3[	15	90%	100%
GRCh38	[0.3,1.0[	0	N/A	N/A

- **SV Germline**

Variant type	Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
SV Germline	GRCh37	2	2	0	100%	**100%**

- **Alu insertions**

Genome Version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	1	1	0	100%	**100%**
GRCh38	1	1	0	100%	**100%**

- **MNV**

Genome version	Variant type	Acceptance criteria	Precision Exome	Precision Genome
GRCh37	MNV	> 85%	**87.0%**	**86,9%**
GRCh38	MNV	> 85%	**93.0%**	**91,5%**

- **SNV / Indel Somatic**

  - **Sensitivity**

Genome version	Variant Type	Positive controls	True positives	False negatives	Acceptance Criteria*	Sensitivity
GRCh37	SNV (>= 5%)	121	121	0	100%	**100%**
	Indel (>= 5%)	21	21	0	100%	**100%**
GRCh38	SNV (>= 5%)	121	121	0	100%	**100%**
	Indel (>= 5%)	21	21	0	100%	**100%**

**The acceptance criterion for variant detection is defined as 100% sensitivity for SNVs and indels with an expected VAF above 5%.*

  - **Trueness**

			Trueness			
			GRCh37		GRCh38	
Gene	Variant	Expected frequency	Average frequency*	Bias	Average frequency*	Bias
KIT	p.D816V	10%	9,07% +/- 0,79%	0,97%	9,07% +/- 0,79%	0,97%
PIK3CA	p.E545K	8,80%	7,88% +/- 0,84%	1,04%	7,87% +/- 0,83%	1,05%
EGFR	p.E746_A750del	1,90%	1,53% +/- 0,5%	0,40%	1,53% +/- 0,5%	0,40%
KRAS	p.G12D	6,30%	6,27% +/- 0,76%	0,57%	6,27% +/- 0,76%	0,57%
KRAS	p.G13D	15%	14,59% +/- 0,86%	0,69%	14,59% +/- 0,86%	0,69%
EGFR	p.G719S	24,50%	24,5% +/- 1,39%	1,15%	24,5% +/- 1,39%	1,15%
PIK3CA	p.H1047R	17,50%	17,47% +/- 1,04%	0,77%	17,47% +/- 1,04%	0,77%
EGFR	p.L858R	2,80%	3,58% +/- 0,51%	0,83%	3,58% +/- 0,51%	0,83%
NRAS	p.Q61K	12,50%	11,27% +/- 1,19%	1,27%	11,27% +/- 1,19%	1,27%
EGFR	p.T790M	0,90%	1,33% +/- 0,15%	0,43%	1,33% +/- 0,15%	0,58%
BRAF	p.V600E	10,70%	11,68% +/- 0,62%	0,98%	11,68% +/- 0,62%	0,98%

**The average of the frequencies is obtained over 12 replicates of the HD200 sample.*

- **SNV/Indel Somatic LF**

Genome version	Positive controls	True positives	False negatives	Acceptance Criteria*	Sensitivity
GRCh37	42	42	0	100%	**100%**
GRCh38	42	42	0	100%	**100%**

**The acceptance criterion for variant detection is defined as 100% sensitivity for SNVs and indels with an expected VAF above 0.1%.*

- **CNV Somatic Large Panels**

Genome version	Variant type	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	Gain	27	26	1	90%	**96,3%**
	Loss	20	19	1	90%	**95.0%**
	LOH	11	11	0	90%	**100.0%**
GRCh38	Gain	27	26	1	90%	**96.3%**
	Loss	20	19	1	90%	**95.0%**
	LOH	11	10	1	90%	**90.9%**

- **Methylation Index**

Genome version	Acceptance Criteria	Weighted agreement
GRCh37	>99%	**99.88%**

- **Somatic DNA fusion**

Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	9	9	0	100%	**100%**
GRCh38	9	9	0	100%	**100%**

- **Somatic LBx DNA fusion**

Genome version	Tumor fraction	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh38	1%	22	21	1	85%	**95.5%**
	0.5%	22	19	3	85%	**86.4%**
	0.1%	22	21	1	85%	**95.5%**

- **TMB**

Genome version	Acceptance Criteria	Sensitivity	Precision
GRCh37	>80%	**83%**	**85%**

- **MSI**

Genome version	Positive controls	Negative controls	True positives	False negatives	Acceptance Criteria	Sensitivity	Precision
GRCh37	22	24	22	0	100%	**100%**	**100%**
GRCh38	6	17	6	0	100%	**100%**	**100%**

- **GiP**

Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	171	165	6	90%	**96,5%**

Genome version	Positive controls	True positives	False positives	Acceptance Criteria	Precision
GRCh37	171	165	7	90%	**95,9%**

- **Sentieon SNV / Indel Germline**

Genome Version	Variant Type	Acceptance Criteria*	Sensitivity Exome	Sensitivity Genome
GRCh37	**SNV**	**> 90%**	**98.9%**	**98.5%**
	Indel	**> 80%**	**96.5%**	**98.5%**
GRCh38	**SNV**	**> 90%**	**98.5%**	**99.3%**
	Indel	**> 80%**	**95.5%**	**99.1%**

**for the detection of variants located in the high confidence regions of NA12878 and NA24385*

Genome Version	Variant Type	Acceptance Criteria*	Precision Exome	Precision Genome
GRCh37	**SNV**	**> 90%**	**98.7%**	**98.5%**
	Indel	**> 80%**	**96.5%**	**98.6%**
GRCh38	**SNV**	**> 90%**	**99.6%**	**99.6%**
	Indel	**> 80%**	**97.2%**	**99.5%**

**for the detection of variants located in the high confidence regions of NA12878 and NA24385.*

- **CNV WGS**

Genome version	Positive controls	True positives	False positives	Acceptance Criteria	Sensitivity
GRCh38	**Duplication**	**1kb - 5 kb**	**8**	**10%**	**37,5%**
	Duplication	**≥ 5kb**	**3**	**10%**	**100%**
	Deletion	**1kb - 5 kb**	**321**	**80%**	**61,7%**
	Deletion	**≥ 5kb**	**132**	**80%**	**92,4%**

Although the small deletion sensitivity (1-5kb) falls below the acceptance criteria, this reduction is an acceptable compromise in order to maximise the detection of more impactful large events and duplications.

Genome version	Positive controls	True positives	False positives	Acceptance Criteria	Precision
GRCh38	**Duplication**	**1kb - 5 kb**	**8**	**NA**	**30%**
	Duplication	**≥ 5kb**	**3**	**NA**	**75%**
	Deletion	**1kb - 5 kb**	**321**	**NA**	**94,3%**
	Deletion	**≥ 5kb**	**132**	**NA**	**92,4%**

- **Sentieon SNV/Indel Somatic**

Genome version	Variant type	Positive controls	True positives	False positives	Acceptance Criteria	Sensitivity
GRCh38	SNV	39	37	2	85%	**95%**
	Indel	21	16	5	75%	**76%**

- **Fusion RNA**

Genome Version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	20	20	0	100%	**100%**
GRCh38	20	19	1	95%	**95%**

- **Splicing RNA**

Genome Version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	3	3	0	100%	**100%**
GRCh38	3	3	0	100%	**100%**

- **CNV Small Panel Somatic**

Genome Version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	4	4	0	100%	**100%**
GRCh38	4	4	0	100%	**100%**

- **IGHV status IMGT**

Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	122	116	6	95%	**95,9%**

- **IGHV status igblast**

Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	122	117	5	95%	**95,1%**

- **IGHV subsets**

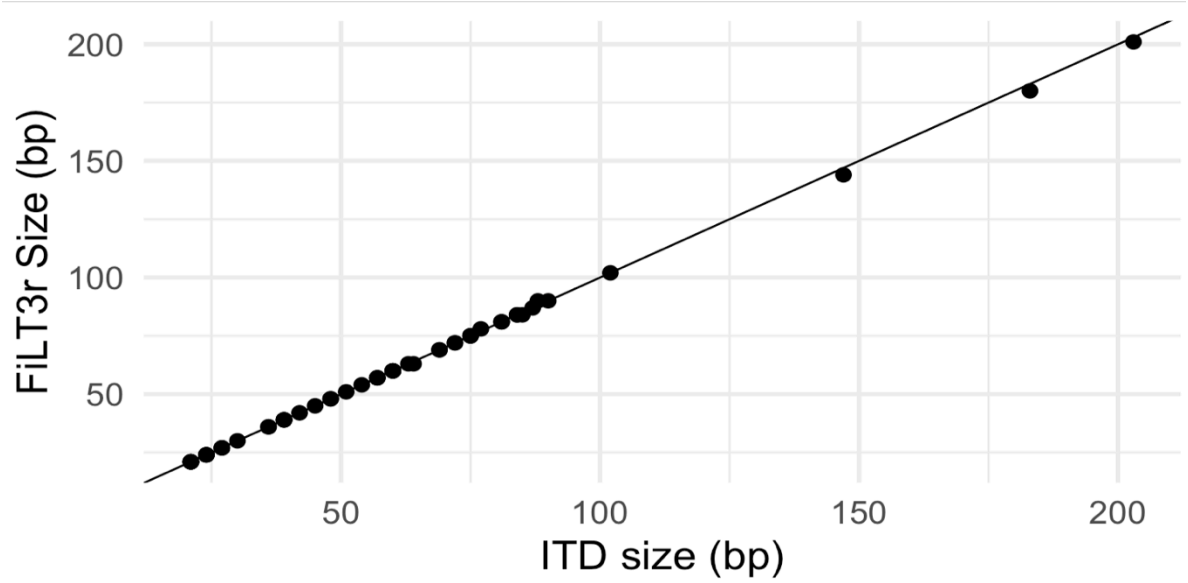
Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh37	61	10	0	95%	**100%**

Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Precision
GRCh37	61	10	0	95%	**100%**

- **CFH**

Genome version	Positive controls	True positives	False negatives	Acceptance Criteria	Sensitivity
GRCh38	4	4	0	100%	**100%**

- **FLT3**



Observed vs Theoretical FLT3-ITD Event Size. **Pearson coefficient $R^2 > 0.99$**

The results show 100% sensitivity for detecting FLT3-ITD events and an almost perfect correlation between the expected and observed event sizes from the analysis.

- **UPD**

    - **Sensitivity**

Analysis Type	Genome Version	Positive controls	Negative controls	True positives	False negatives	Acceptance Criteria	Sensitivity
Singleton	GRCh37/38	7	23	7	0	80%	**100%**
Trio	GRCh37/38	7	2	6	1	80%	**86%**

    - **Precision**

Analysis Type	Genome Version	Positive controls	Negative controls	True positives	False positives	Acceptance Criteria	Precision
Singleton	GRCh37/38	7	23	7	0	80%	**100%**
Trio	GRCh37/38	7	2	6	0	80%	**100%**

3.2.1.2. Annotation Modules

- **DiagAI score**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Rare disease positive exomes	Germline	Median rank	= 1	**1**
Rare disease positive exomes	Germline	top_10	> 90%	**95%**
Rare disease CNV exomes	Germline	Median rank	= 1	**1**
Rare disease positive CNV exomes	Germline	top_1	> 80%	**83.3%**

- **DiagAI somatic score**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Cancer positive TSO500	Somatic	top_10	> 90%	**100%**

- **DiagAI shortlists**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Rare disease positive exomes	Germline	sensitivity	> 90	**95%**
Rare disease positive exomes	Germline	Average length	< 25	**18.475**
Rare disease positive exomes	Germline	Median length	< 25	**17**

- **DiagAI smartpicks**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Rare disease positive exomes	Germline	sensitivity	> 40%	**45%**
Rare disease positive exomes	Germline	precision	> 80%	**78%**

- **ACMG CNV**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
ACMG expert annotated CNV dataset	Germline	sensitivity	> 70%	**76%**

- **ACMG SNV**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
ACMG expert annotated dataset	Germline	accuracy	> 50%	**56%**

- **DiagAI UP2**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Disease causing variants	Germline	sensitivity	90%	**92%**

- **DiagAI UP2 SV**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Disease causing variants	Germline	sensitivity	90%	**98%**

- **DiagAI Phenogenius**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Gene symptoms associations	Germline	top_100	50%	**49.8%**
Gene symptoms associations	Germline	top_200	60%	**60.7%**
Gene symptoms associations	Germline	top_500	75%	**77.0%**
Gene symptoms associations	Germline	Median in top100	<20	**15**
Gene symptoms associations	Germline	Mean in top100	<30	**25**

- **Compermed**

Dataset	Type analysis	Metric evaluated	Acceptance Criteria	Metric value
Cancer positive TSO500	Somatic	sensitivity	90%	**91%**

3.2.2. Clinical performance

The SeqOne Platform is not intended to be used for the assessment of a particular clinical condition or physiological state. While it provides information on a patient's physiological state, such as the identification of certain variants, it is not intended for assessing specific clinical conditions or physiological states. Therefore, the SeqOne Platform itself does not necessitate a demonstration of clinical performance.

4. Precautions

4.1. Compliant use

Any inappropriate use of the platform is forbidden.

Users are responsible for checking that the platform is operational at all times and ensuring the necessary conditions for the platform's proper performance.

⚠ SeqOne Platform is a research use product in your country and SeqOne is not responsible for the use of the device in diagnostic procedures. Consequently, users must exercise due caution when using the software and its results.

4.2. Malfunctions

In the event of a malfunction:

- Immediately stop using the software
- Try to identify or eliminate the causes by following the troubleshooting procedure described in this guide (see section **Troubleshooting and error messages**)
- If you are unable to identify or eliminate the cause using this guide, please email the support team at **support@seqone.com**.

5. Minimum requirements

5.1. Hardware

The SeqOne platform software modules are installed as part of a client-server arrangement. The service provider must be a Linux server to ensure compatibility with most of the tools commonly used in bioinformatics. Each system item is encapsulated in a runtime environment such as the Docker® container, which can be deployed on any conforming Docker® platform, irrespective of the underlying hardware architecture. Users must have a computer with an internet connection and a web browser (see below).

The minimum system requirements for a client computer are as follows:

- 4 GB RAM
- Dual-core processor
- Internet access
- Chrome browser (version 36.0.1985 onwards) or Firefox (version 31 onwards) or Microsoft Edge (all Chromium-based versions).

5.2. Network

The minimum recommended upload/download speed for using the platform over the internet is 100 Mbps. The HTTPS protocol is used to connect to the platform. To access the platform correctly, we recommend that you ask your IT department to authorize the following addresses :

*.eu.seqone.com
*.common.seqone.com

6. Authentication

6.1. Login

For security reasons, users can only access the platform after signing in on the home screen.

The home screen is available via the following URL:

- For Europe and Africa: https://auth.eu.seqone.com/
- For Switzerland: https://auth.ch.seqone.com/
- For Asia and Oceania: https://auth.au.seqone.com/
- For Americas: https://auth.us.seqone.com/

When signing in on the home screen, you must enter your:

- Login or email address in the "*Username" or "email"* field
- Password in the "*Password"* field

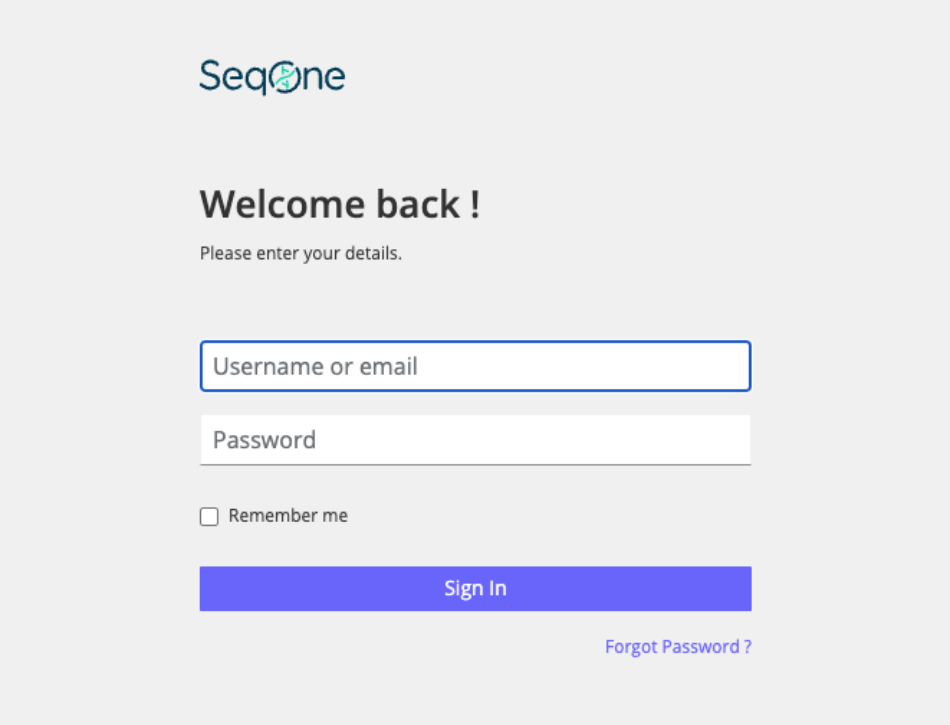


Figure 1: Login page

If two-factor authentication is activated for your account, you will need to enter a 6-digit verification code sent by email to the address associated with your SeqOne account. This code is valid for 30 minutes.

If you have access to several workspaces or entities, you will be asked to select the entity to which you wish to connect via a drop-down menu under "Choose a workspace".

You can change your password by clicking on "Forgot Password". Follow the instructions by entering the email address associated with your SeqOne account, and you will receive an email allowing you to reset your password.

The password policy follows the following rules:

- Minimum length: 12 characters.
- Complexity: passwords must include at least one upper case letter, one lower case letter, one number and one special character. The larger the character set, the stronger the password.
- Password history: no re-use of previous passwords.
- No maximum length: to allow the use of secret phrases.

It is not possible for users to change their own user ID.

6.2. Logout

Users can sign out by clicking on the user account logo in the top-right corner of the page and then on "*Logout*".

6.3. Home page

After authentication, the user accesses the dashboard page, which is divided into two parts:

- a blue search and action banner, accessible throughout the navigation, comprising from left to right :
    - the SeqOne icon for returning to the dashboard at any time,
    - a search bar for direct access to a given analysis,
    - a VKB search icon  for searching for variants evaluated in the VKB database (see section **Evaluating variants**),
    - an icon  opening the user's "To do list", from which they can access the analyses assigned to them (see section **Assigning analyses for interpretation**),
    - a menu bar, available via the icon  , which gives access to the user parameters, the administrator panel, the user manual, the terms of use, the support email address and which allows the user to log out (see section **User account management**).
- the main screen for managing projects and analyses.

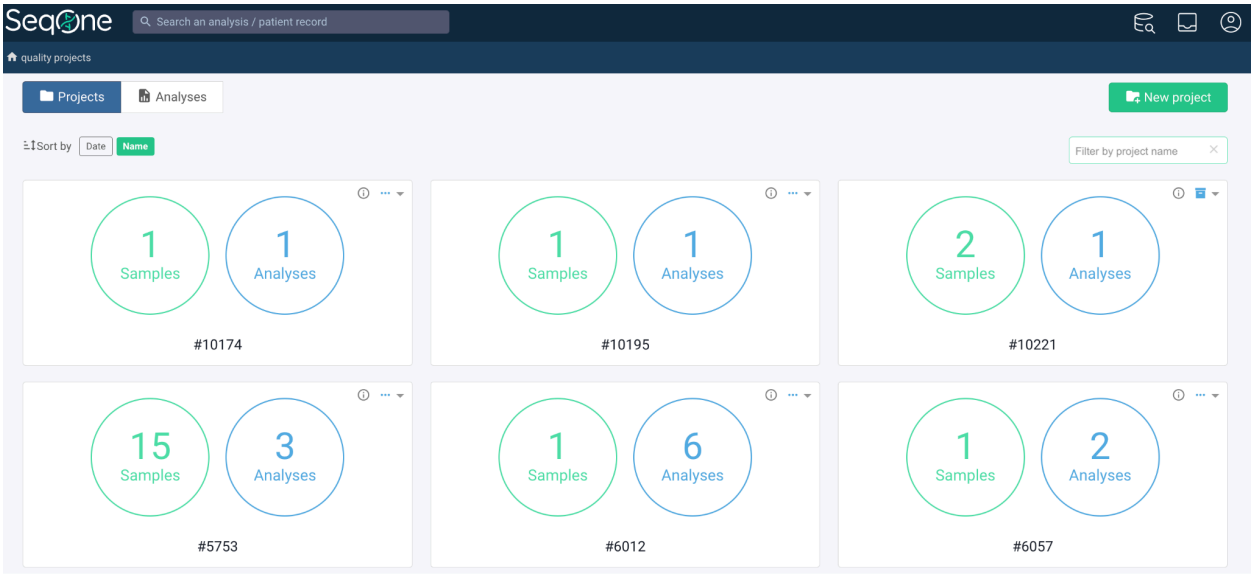


Figure 2: User dashboard

You can access the user dashboard at any time by clicking on the SeqOne logo.

6.4. User account management

You can manage your user account by clicking on the logo in the top-right corner of the taskbar.

In this section, the user has the possibility to:

- access account settings by clicking on “My account”
- access user parameters by clicking on "Entity Settings"
- change the entity to which the user is connected by clicking on "Entity", and selecting the entity of choice under "Choose your workspace",
- access the administrator panel, by clicking on "Admin", allowing access to the list of users with access to the entity, and the creation, modification or deletion of user accounts.
- click on "About" to access the user manual in various languages, SeqOne software version information and contact details,
- access the terms of your SeqOne license by clicking on "Terms of Use",
- click on "Support" to access the support email address for any questions or problems encountered on the platform,
- access the shortcut by clicking “Keyboard Shortcuts”
- switch between the standard version and the preview version of the application by clicking on "SeqOne Mode", and selecting either "SeqOne" (standard mode) or "SeqOne Preview" (preview mode). The currently active mode is indicated by a checkmark.
- log out by clicking on "Logout".

6.4.1. Managing My Account

In the “My account” section, the user can:
- upload their signature to include it in the clinical report,
- change their password by clicking on "Change your password" (opens a new tab) in the "Manage your credentials" section.

6.4.2. Managing the user settings - definitions

In user settings the user can manage the settings, files, and configurations described below.

Entity

The entity refers to the facility or team that has taken out a SeqOne platform license. The platform is designed to be used by several members within the same group. From the user settings, the *Account* tab is used to manage information relating to the entity.

Users

For this tab please refer to the Admin management - interface presentation

Worksets

This tab lists the different analysis worksets available on the entity to which the user is connected. Because several versions of certain pipelines can coexist on the SeqOne platform, the user can define from this menu which version should be pre-selected when launching the analysis.

The modifications made to this list are tracked, so that the user who defined the version to launch by default of a workset is automatically displayed, as well as the date of the modification.

Manifest

The manifest displays the genomic regions of interest (ROI) as a BED file, which may be accompanied by their associated annotations. You can upload a new manifest via the *Manifest* tab in the user settings by clicking on "Add manifest" and selecting the manifest file that you wish to upload. Once uploaded, this manifest will be available to all users when creating a new project and when applying an in silico panel in an analysis.

Panels - Transcript

Users can upload a new list of reference transcripts (RefSeq) via the *Transcript* tab.

Panels - Genes

This tab allows you to import or generate gene panels based on the NCBI GeneID by clicking on "Add gene list".
The expected file format is a tabular file (e.g. a .tsv file) without column headers, which is detailed on the platform. The first column of the expected file is dedicated to gene identifiers (NCBI Entrez gene ID), the second column (optional) contains the transcript of interest (RefSeq ID) and the third column (optional) contains the gene's HGNC symbol. This file can also be generated by inserting the desired list of genes, separated by a comma, in the "Gene symbols" field.

Marker

This tab is used to import a list of genetic identity markers by clicking on "Add marker config". The expected file format is a tabular file (.tsv or .txt file) with a column header containing 5 columns detailing: chromosome (CHR), variant identifier or rsID (ID), genomic position (POS), reference allele (REF) and alternative allele (ALT). The file must not contain more than 100 lines, including the header.

Protocols

An analysis protocol is a predefined combination of tasks designed to standardize analysis interpretation. The corresponding tab allows users to view the list of protocols available on the entity, their versions, and to create them.

To create a protocol, click "New protocol", then add and order the tasks that match your lab's SOPs. To edit an existing protocol, click the pencil icon next to it and save the new version of the protocol.

Select a protocol at analysis launch, or attach it later from the Workflow tab in the Analysis drawer. Mark tasks as complete directly from the Analysis drawer. Each completed task automatically records the user and date for full traceability.

Reporting - Global settings

The Global settings section of the Reporting tab allows users to enable or disable automatic analysis locking when the status changes to "Reporting Validating", and to customize the classification and reporting sections. To do so, under Germline / Somatic, click on New section, enter a section name, and select a tag to associate with the desired classification. Variants that match one of the selected tags or classifications are automatically added to this section. You must first create tags in the corresponding tab to use this feature.

Reporting - Template list

The Reporting tab under "Template list" allows users to configure the templates used when exporting analysis results. From this tab, you can customize the information included in generated reports.

Tags

The "Tags" tab lets users manage custom labels for organizing and filtering variants within the VKB and the report. Users can create tags by clicking on "Add tag", or rename and delete tags that are available to all users within the entity using the corresponding pen and bin icons.

API Keys

An API key is a unique authentication token allowing external applications or scripts to interact with the SeqOne platform programmatically. From the API Keys tab, you can view all existing keys along with their associated information.
Create a new key by clicking on "+ New API Key". Expired keys can be shown or hidden using the "Show expired key" toggle. Each key can be deleted individually. API keys must be kept confidential, as they grant access to the platform with their associated permissions.

6.4.3. Admin management - interface presentation

This interface is dedicated to account administrators and enables them to :

- access the list of users in the entity,
- create new users,
- modify user rights: create projects and upload samples, launch analyses, delete samples, analyses or projects, evaluate variants in VKB, create reports, filter profile management, manage assignments.
- activate dual authentication,
- deactivate a user account.

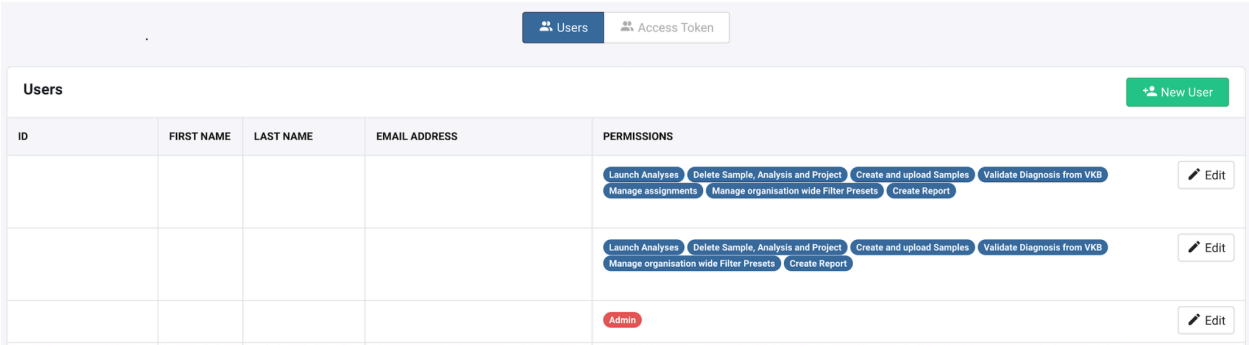


Figure 3: Administrator panel

The ID corresponds to the user's username. It must be in lower case and contain no accents or spaces; special characters such as dot and underscore are accepted. It is advisable to follow the nomenclature firstname.lastname when creating a new user.

Regarding the user's email address, we strongly advise against using a generic email address for security reasons.

As permissions are deactivated by default, it is up to the administrator to activate them manually.

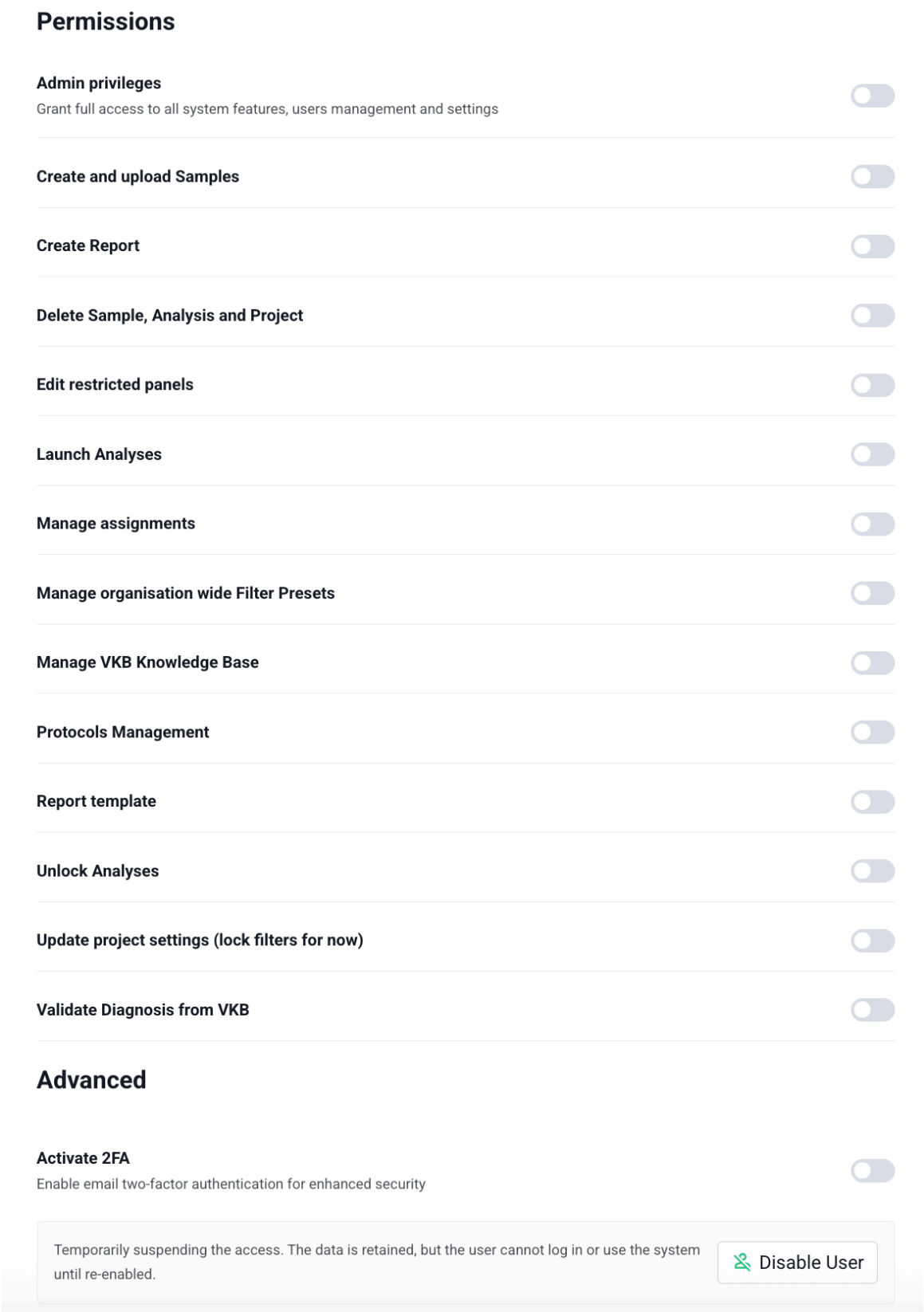


Figure 4: User permissions

You can also enhance security by activating dual authentication. This option sends a unique code to the user each time they connect to the platform.

The "Manage organization wide Filter Presets" role enables the user to modify public filter profiles. If you don't have this role, you can only modify private filter profiles.

Permissions are managed by an entity administrator from the Admin panel. When a user is created, all permissions are disabled by default, and the administrator must manually enable the rights required for that user. The rights assigned to a user apply to all entities (workspaces) the user has access to on the SeqOne platform (for example, if the user can access multiple entities, the same permission set applies across them).

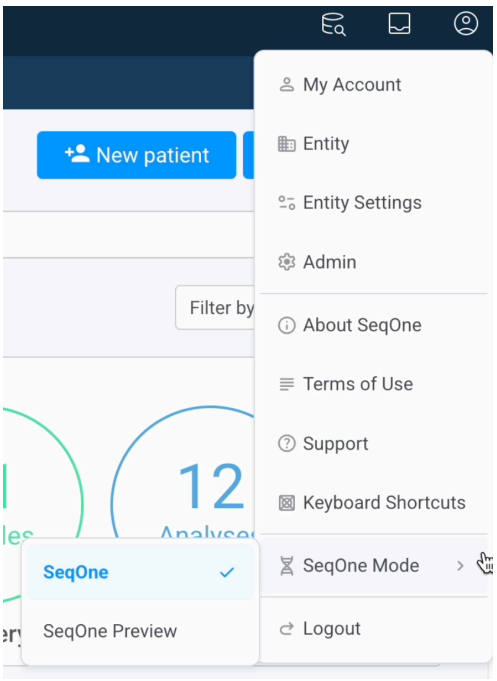
If you receive an error message, please contact customer support at **support@seqone.com**.

6.4.4. Switching between Main and Preview modes

Connecting to SeqOne through the default login URL takes users to their entity's "Main" mode. This default mode hosts CE-marked worksets and other local certifications.

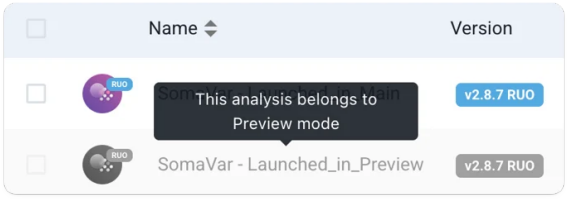
For early access to new features and workset versions, switch to "Preview" mode (RUO: Research Use Only) by going to the bottom of the settings menu and, under SeqOne One, selecting "Preview".

"Main" and "Preview" modes share the same core data (projects, samples, analyses), permissions, resources (files, templates, manifests, gene lists), knowledge (VKB), and frequencies (sample frequency data, occurrences).



When navigating in "Preview" mode, the "SeqOne Preview RUO" logo appears as a reminder that this mode contains RUO features and worksets.

In "Main" mode, analyses launched in "Preview" appear greyed out. If you click one of these greyed-out analyses, the platform prompts you to switch to Preview mode to view it.



7. Description of the operational steps

Step	Description
1	**Project Creation** The analyses conducted on the SeqOne platform are organized within projects, which define the characteristics of the samples analyzed through user-entered metadata (sequencing type, preparation method, genes of interest).
2	**Sample Upload** Samples are uploaded as raw sequencing files (FASTQ), which should be placed in the designated project. Once the files are uploaded, a verification step ensures the integrity of the submitted files before any analysis.
3	**Analysis Launch** Once the data is uploaded to the SeqOne servers, the user can choose from a selection of applications or "worksets," depending on the type of sample analyzed and the types of genomic alterations sought (SNV, CNV, gene fusions, splicing variants, etc.). These applications perform bioinformatics analysis.
4	**Data Quality Check** The applications available on the SeqOne platform include several modules dedicated to data quality analysis, using various sequencing metrics, alignment, and coverage depth.
5	**Interpretation of analysis results** The genomic variations detected in the sample are displayed in various tabs or "viewers," where the user can search for the diagnostic variant responsible for the patient's pathology.
6	Obtaining Therapeutic recommendations and clinical trials associated with the patient's pathology.
7	**Issuance of a report** The analysis results can be exported as a PDF report, which can then be downloaded.

8. Project Creation

The analyses conducted on the SeqOne platform are organized within projects, which define the characteristics of the samples analyzed through user-entered metadata (sequencing type, preparation method, genes of interest).

8.1. Creating a project

After clicking on "*New project*", users are prompted to enter the values for several types of information, or metadata, which are used to describe the project and the associated samples.

- **Project description**
- **Sequencing:**
  - Assay type: type of analysis ("Germline" or "Somatic"),
    - Germline projects (*Assay type* "*Germline*") feature a UMI Trimming checkbox, enabling samples containing UMIs to be processed. By selecting this option, UMIs will be trimmed from FASTQ sequences, as in the UMI disabled mode for somatic worksets (see section **Choosing the UMI processing mode**).
  - Sequencing platform: the sequencing platform
  - Library type: type of read sequencing performed (*Paired-end* or *Single-end*),
  - Data type: DNA, RNA, ctDNA or Methylation Sequencing
    - If DNA is selected for the data type, the sequencing method must be specified using the "DNA Sequencing method" drop-down menu, as well as the amplification method (Capture or Amplicon) via "Amplification method".
  - File input type: *Fastq files, VCF files*
  - Genome: the reference genome version (GRCh37 or GRCh38),
  - Low-Quality Mapping Region: the variant calling in regions of low quality alignment by considering MAPQ0 reads,
  - Sequencing method: Gene Panel, Whole exome sequencing, Comprehensive Genomic Profiling, Shallow WGS, Whole genome sequencing, CGH MicroArrays.
  - Secondary analysis provider: AmoyDX, Pillar, Thermofisher,
  - Manifest: the manifest in the case of a gene panel,
  - ID marker (optional): the identity marker file.
- **Samples:**
  - *(Optional)* New sample metadata can be added by clicking "*You can also create your own metadata*". When prompted, enter the metadata name and value.

Once the different metadata have been entered for the project, users are invited to create their project by clicking on the "*Create project*" button.

8.2. Managing projects

Via the dashboard, users can manage the different projects that have been created on the platform. In particular, they can:

- Access the projects via the *Projects* tab, as well as the samples and analyses for each project by clicking on the "*Samples*" and "*Analysis*" logos respectively.
- Access the analyses directly via the *Analysis* tab.
- Create a new project by clicking on "*New project*" (see the section entitled **Creating a project**).

8.3. Deleting a project

To delete a project, users must:

- Select the project
- Click on the *"Settings"* tab and then on "*Delete project*"
- Confirm by checking "*Yes*" in the message displayed. This action is irreversible.

8.4. DNA data (DNA Sequencing, ctDNA Sequencing)

Low-Quality Mapping Region. When sequencing gene panels, exomes or genomes, the user can include regions of low quality alignment when calling variants. These regions correspond to so-called MAPQ0 reads, whose alignment quality is equal to 0, such as pseudogenes and homologous regions of the genome.

The option is applicable to specific target genes or regions. To use this option, the user is invited to load the manifest targeting the desired regions, or to select it via a drop-down menu.

This option may reduce the quality and accuracy of results. It may introduce background noise in allele frequency and have a negative impact on variant prioritization. It may also increase analysis time due to a higher computational load.

Exome. In the case of exome sequencing, users do not need to specify the manifest to be used; they are invited to select the manufacturer of the kit used by clicking on one of the icons in the Exome kit provider category. A drop-down menu then allows the user to select the desired manifest. If the desired manifest is not available on the platform, the user is invited to contact user support (**support@seqone.com**).

Gene panel. In case of a gene panel, users are prompted to select the *Amplification method* as well as the manifest (in BED format) corresponding to the genes of interest in a drop-down menu. If the required manifest is not available on the platform, it can be uploaded by selecting "*Upload new manifest*" and following the instructions on the screen.

Shallow WGS, CGH MicroArrays, Whole genome sequencing. In the case of genome sequencing, the user does not need to specify which manifest to use.

Capture. By choosing the *capture* amplification method, a UMI icon (*Unique Molecular Identifiers*) is available. Unique molecular identifiers are indices that are added before amplification to help eliminate PCR duplicates and reduce background noise after analysis.

Amplicon. Specific attention must be paid to amplification primers when analysing amplicons. When creating a project for amplicon data, users are advised to select a manifest in which the regions corresponding to the targets do not include the primer sequences. If such a BED is not available, an option can be selected to *clip* a fixed length at the end of the sequences.

Identity monitoring. For sequencing methods corresponding to gene panel, exome, genomic profiling (CGP) or genome (shallow WGS) from FASTQ files, the user is invited to select the identity monitoring markers file from a drop-down menu. If the desired file is not available on the platform, it can be imported by selecting 'Upload new markers' and following the on-screen instructions (see also the **Marker** paragraph in the **Managing the user settings** section).

8.5. RNA data (RNA Sequencing)

If RNA is selected for the data type, users must specify whether their protocol retains the original strand information via the *Stranded protocol* drop-down menu. If the protocol is strand-specific, the order of synthesis for the strands (*Forward-reverse* or *reverse-forward*) must be specified using the *Read orientation* option. Lastly, the sequencing method must be specified (gene panel).

Read orientation. Two options are available for specifying the sequence orientation:

- **Forward-Reverse :** the first read (R1) for each pair corresponds to the sequence of the transcribed messenger RNA (sense strand).

- **Reverse-Forward :** the first read (R1) for each pair corresponds to the complementary sequence of the transcribed messenger RNA (antisense strand).

In general, the orientation of the reads is as follows for the kits mentioned below:

- Agilent: Reverse-Forward
- Illumina: Unstranded
- Roche Kapa Hypercap: Reverse-Forward
- Twist: Reverse-Forward

In case of doubt, please contact support (support@seqone.com).

Once the data type has been specified and the project properly configured, users are prompted to validate by clicking on the "*Next step*" button.

8.6. Project settings

From the project “*Settings”* tab, you can :

- modify the project name and description in the "*Project name*" and "*Project description*" fields,
- modify the quality control (QC) profile,
- change the default filter profile in the variants table,
- permanently delete the project.

8.6.1. Configure default project QC metrics

The user can choose to use the default QC metrics or define customized threshold values from the project's “*Settings”* tab. The user is invited to contact user support (support@seqone.com) to define the desired values.

Custom QC metrics will be applied :

- *QC checks* in the project's *Analyses* view,
- *QC checks* in the *QC auto-validation* section of the analysis QC tab,
- to the *GeneCov* tab of the analysis.

8.6.2. Set the project's default filter profile

The user can choose to set the predefined filter profile applied to the default variant table from the project's “*Settings”* tab (see also section **Using a predefined filter profile**).

A filter profile can be selected for each workset via a drop-down menu, provided it is public. The filter profile will then be applied by default when viewing the variants table, for all project analyses, for all users in the organization.

If this parameter is not modified by the user, when the project is created, the default value will be the profile defined by SeqOne.

The default filter profiles can be locked by clicking the “*Lock Variants Filters*” toggle.

9. Sample Upload

9.1. Uploading samples

Users can import their samples via the ‘Add samples’ icon in each project. Samples can be imported into the SeqOne platform in two ways:

From a PC
Users can upload their FastQ files directly from a folder on their PC by clicking on the "*From your computer*" tab from the "*Add files*" icon. Users can also drag and drop the files from their PC to the dedicated area in the interface.

Via SDS (SeqOne Data Sync)
SDS is an optional premium module for automatically creating projects, transferring data, starting analyses and recovering structured data (TSV) for integration into the laboratory information management system (LIMS).

This module compatible with Linux, MacOS and Windows is an executable that allows users to authenticate and interact with SeqOne APIs in a predefined and highly secure environment.

Minimum requirements:

- Linux 1.19 minimum
- MacOS 10.15 with Intel CPU minimum
- MacOS with Apple Silicon CPU (M1 chip) minimum
- Windows 10 64bits minimum
- 4 GB RAM (8 GB for Windows)
- Quad-core processor
- Internet and SeqOne platform access (app.eu.seqone.com)
- Access to the data storage system (either directly or over the network)

If you would like to use this solution or for further information, contact our support team (support@seqone.com).

Sample upload status
The SeqOne platform uses different statuses to allow users to track the progress of their sample uploads:

- *Upload pending*: import in progress,
- *Verifying sample…*: sample conformity being verified,
- *Mapping…*: sample alignment step,
- *Verified sample*: verified sample,
- *Upload failed:* import failed.

Upload capacity
The file upload capacity (especially for fastq.gz files) depends directly on the user's internet connection. Therefore, the speed of the internet connection should be taken into account when uploading samples. Refer to the following table for a clearer idea of the estimated upload times:

Estimation du temps d'attente					
		ADSL		Fibre optique (Mbits/s)	
		Ethernet : 10 Mbits/s	Wifi : 7,5 Mbits/s	Ethernet : 100 Mbits/s	Wifi : 90 Mbits/s
Poids du fichier fastq	***100Mo***	1min 20s	1min 45s	8s	9s
	1Go	13min 20s	17min 45s	1min 20s	1 min 30s
	10Go	2h 13min 20s	2h 57mn	13min 20s	15 min
	100Go	22h 22min	30h	2h 13min 20s	2h 30min

Figure 5: Estimated upload time according to the file size and internet connection speed.

9.2. Input file format

9.2.1. Primary analysis files

The format of the file to be uploaded for each sample depends on the sequencing platform used. File names must conform to each sequencer's file naming convention.

Files need to be compressed, using the .gz (gzip) extension, before they can be imported into the platform. In the case of paired-end sequencing, two files are expected, corresponding to reads from both ends of each molecule.

When a sample has been correctly detected by the platform at the time of import, the full name of the file is displayed. It is then possible to rename the file by clicking on it. Paired-end files are automatically combined into a single sample.

The following sequencers are supported by the platform:

- *Illumina*: fastq.gz format, paired-end only,
- *Element Biosciences*: fastq.gz format, paired-end only,
- *MGI*: format fastq.gz, paired-end only. The expected file names are: sample_1.fq.gz **and** sample_2.fq.gz **or** flowcellID_L01_sample_1.fq.gz **and** flowcellID_L01_sample_2.fq.gz

9.2.2. Secondary analysis files

The format of the file to be uploaded for each sample depends on the sequencing platform and application used for secondary analysis. Files must be compressed with the .gz (gzip) extension before platform import, unless otherwise specified.

The import format of secondary analysis files generated using Illumina, Oxford Nanopore Technologies, Agilent, AmoyDx, Pillar, Thermofisher applications depends on the workset used (see section **Analysis worksets available from secondary analyses files**).

When the platform correctly detects a sample during import, it displays the complete filename. You can rename the file by clicking on it. Files that share the same base name are automatically combined into a single sample.

DRAGEN files. When creating a project, if VCF files (Secondary Analysis data) and Illumina metadata are chosen, the expected files and formats for each workset are as follows:

GermVar Tertiary, GermVar Tertiary Family	sample.hard-filtered.vcf.gz sample.cnv.vcf.gz sample.sv.vcf.gz sample.repeats.vcf.gz

Oxford Nanopore Technologies files. When creating a project, users can import VCF files (Secondary Analysis data) and Oxford Nanopore Technologies metadata. The expected files and formats allow for importing a standalone variant (SNP) file or combining it with supplementary files for CNV, STR, or SV analysis:

GermVar Tertiary, GermVar Tertiary Family	sample.wf_snp.vcf.gz sample.wf_cnv.vcf.gz sample.wf_str.vcf.gz sample.wf_sv.vcf.gz

Thermofisher/Affymetrix files. The import format for CGH secondary analysis files is the .txt format generated by Thermofisher's Affymetrix application. This format is supported only when both Thermofisher/Affymetrix and CGH MicroArrays metadata are selected during project creation. The files and formats expected for each workset are as follows:

GermVar Tertiary	ARRAY_SampleID.txt

CGH MicroArrays files. The import format for CGH secondary analysis files is the .xls format generated by Agilent's Cytogenomics application. This format is supported only when both

Agilent and CGH MicroArrays metadata are selected during project creation. The files and formats expected for each workset are as follows:

GermVar Tertiary	IntervalBasedReport_sample.xls

Alissa Reporter files. The import format for Alissa Reporter secondary analysis files is the .xls format generated by Agilent's Alissa Reporter application. This format is supported only when both Agilent and CGH MicroArrays metadata are selected during project creation. The files and formats expected for each workset are as follows:

GermVar Tertiary	IntervalBasedReport_sample.xls ABERRATION_&_LOH_INTERVAL_sample.xls

PacBio-HiFi files. When creating a project, if VCF files (Secondary Analysis data) and PacBio-HiFi metadata are chosen, the expected files and formats for each workset are as follows:

GermVar Tertiary, GermVar Tertiary Family	sample.hard-filtered.vcf.gz sample.cnv_sv.vcf.gz sample.repeats.vcf.gz

Somatic gene panel VCF files. The import format for SNV, CNV, Fusions secondary analysis files is the .vcf.gz format generated by Illumina, AmoyDx, Pillar, Thermofisher applications. Original file names must be converted to match the required naming format described below.

The expected files and formats are as follows when selecting VCF files - Secondary Analysis data, and Gene Panel metadata during somatic project creation. Users can import either a variant (SNV) file by itself or combine it with supplementary files for CNV or fusions analysis:

SomaVar Tertiary	sample.snv.vcf.gz sample.cnv.vcf.gz sample.fus.vcf.gz

The expected files and formats are as follows when selecting "VCF files - Secondary Analysis data" and "Comprehensive Genomic Profiling" metadata during somatic project creation. Users can import a multi-variant VCF file such as the "All Variants" VCF file from the Oncomine™ Comprehensive Assay V3/Plus.

SomaVar Tertiary	Oncomine_<LibPrepID>_<analysisID>.vcf

9.3. Sample metadata

For each uploaded sample, users can assign values to the metadata that they selected when creating the project or define new values by clicking on the icon in the top-right corner of the samples list .

Some sample metadata can be uploaded directly as a tab-separated values file by selecting the icon.

These criteria play an important role in ensuring data traceability and for cross-referencing different analyses performed on the platform. Some metadata, such as gender, disease and HPO IDs, are involved in the process of identifying variants.

When the Assay Type field is filled with "Germline", the metadata associated with the sample by default are "HPO", "Sex" and "GnomAD" group. When the "Assay Type" field is filled with "Somatic", the metadata associated with the sample by default are "Disease" and "Sex".

To add or modify HPOs in a sample, the user can click on '+ Add' in the HPO column after selecting a sample. This opens a pop-up window allowing HPOs to be added one by one, by searching for a phenotype or HPO code in the search bar. It is also possible to import a text file containing HPO codes by clicking on the icon . A summary of the terms entered is then available, with the option of deleting terms if required.

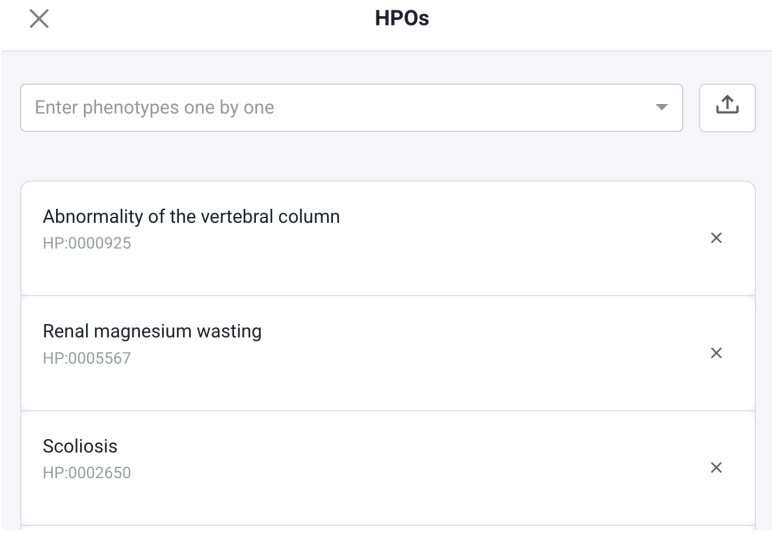


Figure 6: HPO import pop-up window

9.4. Edit a sample

The name, as well as the metadata associated with a sample can be edited by the user at any time, by selecting the sample in the "Samples" list, then clicking on the  icon. Modifications must then be validated to be effective.

9.5. Delete a sample

Users can delete a sample's FastQ files from the platform by clicking on the  icon after selecting the required sample. These files will free up storage space, and the results will still be available on SeqOne.

To delete a sample, click on the  icon in the "Samples" tab after selecting the required sample. The sample can be permanently removed only after deleting all analyses associated with it.

10. Analysis

10.1. Launching a new analysis

Users can launch a new analysis by clicking on "*New analysis*" in the *Overview* screen for each project.

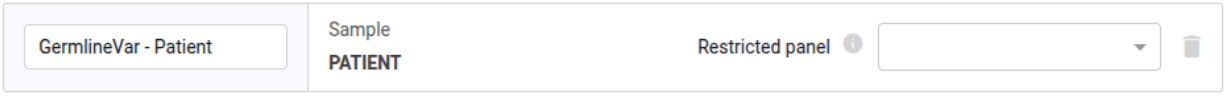
Analysis launch is performed in three distinct steps, described below.

The "*1. Workset*" step allows you to choose the appropriate analysis workset (see section **Analysis Workset Selection**).

The "*2. Select Samples*" step is divided into two parts:

- *2.1. Set options*:
    - describes the selected workset and reference genome
    - allows adding *padding* through the "Extra calling padding" functionality if applicable (see section **Choosing the analysis workset**)
    - allows choosing the UMI processing mode (see section **UMI Processing Mode Selection**) if applicable.
    - allows choosing the CNV baseline, MSI baseline, or adjusting the minimum VAF threshold if applicable.
    - allows choosing the protocol
- *2.2. Select sample(s)*: allows selecting samples for analysis. Users can specifically select new samples, which are marked as 'New', using the 'select all new' function.

The "*3. Confirm Analysis*" step summarizes the selected samples. For each analysis to be launched, it is possible to apply a restricted panel by selecting one or more gene panels via the "*Restricted panel*" option (see also section **Panel Modal View**).



The "*Restricted panel*" option will apply a filter to the analysis in the form of one or more gene panels. This option does not affect variant calling, which is determined by the project manifest and added *padding*. The choice of restricted panels is reversible; the user can modify these panels from the *Panel* tab of the analysis management side panel after analysis (see **Panel modal view** section).

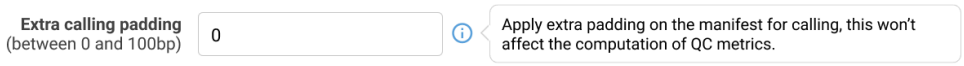
10.1.1. Choosing the analysis workset

Several analysis worksets are available for answering various biological questions. The worksets proposed when creating a new analysis depend on the metadata that were defined when creating the project and may therefore vary according to the data.

Analysis worksets available from FASTQ files

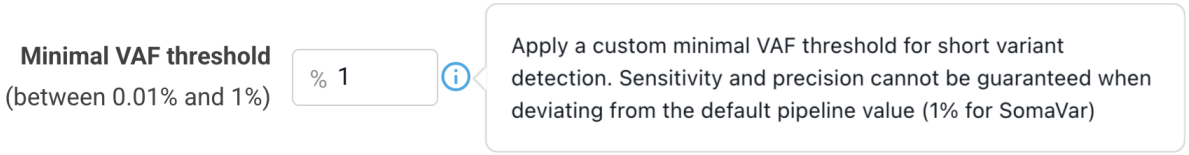
The following worksets are available when the metadata *Fastq files - Primary Analysis data* is selected during project creation.

Excluding projects where the amplification method is amplicon, some worksets (i.e. GermlineVar, GermlineFamily, SomaVar, SomaVar LF *(versions ≥ 2.0),* SomaCGP *(version ≥0.6),* GermVar, GermVar Family, SomaLBx, SomaMethyl) allow you to extend the search area for variant calling and add between 0 and 100 bp of padding to the initial manifest.



The option is available in step 2 (Select Sample) of the analysis launch via the "Extra calling padding" feature. By default, the value is set to 0 and must be a natural integer.

Some worksets (i.e. SomaVar *(versions ≥ 2.6)* and SomaHemato *(versions ≥ 0.2)*) offer an option to adjust the minimum VAF threshold for variant detection. This threshold can be set between 0.01% and 1%, with 1% being the default value.



For more information about this option, contact the user support at support@seqone.com.

GermlineVar

GermlineVar is used to detect germline mutations based on the sequencing of a gene panel or patient's exome.

When running a GermlineVar analysis, users are invited to:

- Select the version of workset to execute and click on "Select",
- Select the samples to be analysed, and then click on "*Next step*",
- Launch the analysis by clicking on "Start analysis".

GermVar

GermVar enables the detection of germline mutations, based on the sequencing of a gene panel, exome, or genome of a patient.

When running a GermVar analysis, users are invited to:

- Select the workset version they wish to execute, then click on "*Select*",
- Select the model to be used,
- Select the sample(s) to analyze then click on "*Next step*",
- Launch the analysis by clicking on "*Start analysis*".

GermlineVarLR

GermlineVarLR is used to detect germline mutations in long-read sequencing data from gene panel sequencing.

When running a GermlineVar LR analysis, users are invited to:

- Select the version of workset to execute and click on "Select",
- Select the model to be used and the sample(s) to be analysed, and then click on "*Next step*",
- Launch the analysis by clicking on "Start analysis".

GermlineFamily

GermlineFamily is used to detect hereditary mutations based on the sequencing of a gene panel, the exome or the genome of a patient and the members of his/her family. This workset can also be used for carrier screening.

When running a GermlineFamily analysis, users are invited to:

- Select the version of the workset to execute, then click on "Select",
- Select at least 2 samples,
- In the "Set conditions" section, indicate whether the sample is the patient's or a family member's by defining the family link under "Family link". The patient is automatically marked as "affected". In the case of a carrier screening, indicate for each sample the family relationship by "Mother" or "Father" without checking the "Is affected?" box.
- Check that the parameters match, then click on "Next step",
- Launch the analysis by clicking on "Start analysis".

GermVar Family

GermVar Family enables the detection of germline mutations, based on the sequencing of a gene panel, exome, or genome of a patient.

When running a GermVar Family analysis, users are invited to:

- Select the version of the workset to execute, then click on "Select",
- Select the model to be used,
- Select at least 2 samples,
- In the "Set conditions" section, indicate whether the sample is the patient's or a family member's by defining the family link under "Family link". The patient is automatically marked as "affected". In the case of a carrier screening, indicate for each sample the family relationship by "Mother" or "Father" without checking the "Is affected?" box.
- Check that the parameters match, then click on "Next step",
- Launch the analysis by clicking on "Start analysis".

CNVCapture

CNVCapture is used to detect germline copy number variations (CNVs) from gene panel capture data.

When running a CNVCapture analysis, users are invited to:

- Select the version of the workset to execute, then click on "Select",
- Select a minimum number of 8 samples to analyse, then click on "Next step",
- Run the analysis by clicking on "*Start analysis*".

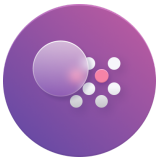
CNVExome

CNVExome is used to detect copy number variations (CNVs) from exome capture data.

When running a CNVExome analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- Select the model to be used and the sample(s) to be analysed, and then click on "*Next step*",
- Launch the analysis by clicking on "Start analysis".

SomaVar



SomaVar detects somatic mutations from tumour samples. This workset can be used with capture and amplicon data.

When running a SomaVar analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- Select the sample(s) to be analysed and then click on "*Next step*",
- Launch the analysis by clicking on "Start analysis".

SomaCNVCapture

SomaCNVCapture is used to detect somatic copy number variations (CNVs) from gene panel capture data.

When running a SomaCNVCapture analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- Select a minimum number of eight samples to be analysed, and then click on "*Next step*",
- Run the analysis by clicking on "*Start analysis*".

SomaRNA

SomaRNA can be used to detect gene fusions from RNA-seq capture data.

When running a SomaRNA analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- select the sample(s) to be analyzed, then click on "Next step",
- Run the analysis by clicking on "*Start analysis*".

SomaMSI

SomaMSI detects the state of the microsatellite regions (MSS or MSI).

When running a SomaMSI analysis, users are invited to:

- Select a minimum number of eight samples to be analysed, and then click on "*Next step*",
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- select a minimum number of 8 samples to analyze then click on "Next step",
- Run the analysis by clicking on "*Start analysis*".

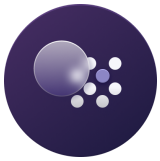
SomaHRD

SomaHRD allows the detection of HRD status from shallow WGS data alone or accompanied by a panel of genes including BRCA1 and BRCA2.

When running a SomaHRD analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- Enter the cellularity metadata (% of tumoral cells) for each sample (see section **Sample metadata**),
- Select the sample(s) to be analysed and then click on "*Next step*",
- Select the appropriate SomaVar or SomaCGP (GRCh37 or GRCh38) analysis for each sample using the 'Select panel analysis (Optional)' option,
- Run the analysis by clicking on "*Start analysis*".

SomaVar LF

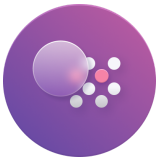


Detection of low frequency short variants and structural variants, as well as fusion genes, from tumour samples (solid or liquid biopsies, circulating tumoural DNA) enriched by capture on a panel of genes.

When running a SomaVar LF analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- Select the sample(s) to be analysed and then click on "*Next step*",
- Run the analysis by clicking on "*Start analysis*".

SomaCGP



SomaCGP enables the detection of somatic mutations, structural variants, gene fusions, as well as MSI and TMB status from tumour samples. This workset is dedicated to the analysis of FFPE DNA prepared with Agilent's SureSelect Cancer CGP Assay panel, and is only available when the Comprehensive Genomic Profiling metadata is selected at project creation. This workset can be used with capture data.

When running a SomaCGP analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- If requested, select a "CNV baseline" and/or "MSI baseline"(if no baseline is selected, the analysis will run without CNV/MSI detection),
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- Select the sample(s) to be analysed and then click on "*Next step*",
- Launch the analysis by clicking on "Start analysis".

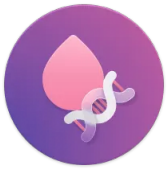
SomaHemato

SomaHemato enables the detection of somatic mutations from samples with hematology specificities. This workset can be used with capture data.

When running a SomaHemato analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- Select the sample(s) to be analysed and then click on "*Next step*",
- Launch the analysis by clicking on "Start analysis".

SomaLBx



SomaLBx enables the detection of low frequency somatic mutations, structural variants, gene fusions, from ctDNA. This workset is dedicated to the analysis of ctDNA and can be used with capture data.

When running a SomaLBx analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- Select a "CNV baseline" (if no baseline is selected, the analysis will run without CNV detection),
- In the options, select the applicable UMI processing mode from the eligible modes for the analysis (see Section **Choosing the UMI processing mode**),
- Select the sample(s) to be analysed and then click on "Next step",
- Launch the analysis by clicking on "Start analysis".

SomaMethyl

SomaMethyl enables the detection of methylated CpG islands, from Circulating tumor DNA samples. This workset is dedicated to the analysis of methylation in ctDNA and can be used with capture bisulfate data.

When running a SomaMethyl analysis, users are invited to:

- Select the version of the workset to execute and click on "Select",
- Select a "Methyl baseline" corresponding to the control cohort (if no cohort is selected, the analysis will run without threshold detection and differential methylation),
- Select the sample(s) to be analysed and then click on "Next step",
- Launch the analysis by clicking on "Start analysis".

Analysis worksets available from secondary analyses files

The following worksets are available when the *VCF files - Secondary Analysis data* is selected during project creation or when both Agilent and CGH MicroArrays metadata are selected (see section **Secondary analysis files)**. In this case, the platform receives output data from the selected provider's application, proceeds to annotate and inject variants into their respective visualization interfaces.

GermVar Tertiary

Single sample analysis pipeline of variants detected by the Illumina Germline application or the EPI2ME application from Oxford Nanopore Technologies, from WGS and WES data or variants detected by Agilent's Cytogenomics application from MicroArray CGH data.

When running a GermVar Tertiary analysis, users are invited to:

- Select the workset version to execute, then click on "Select",
- Select the sample(s) to be analyzed, then click on "Next Step".
- Launch the analysis by clicking on "Start analysis".

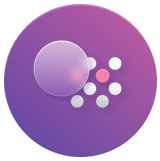
GermVar Tertiary Family

Multi-sample analysis pipeline of variants detected by the Illumina Germline application or the EPI2ME application from Oxford Nanopore Technologies, from WGS and WES data.

When running a GermVar Tertiary Family analysis, users are invited to:

- Select the workset version to execute, then click on "Select",
- Select at least 2 samples,
- In the "Set conditions" section, indicate whether the sample is the patient's or a family member's by defining the family link under "Family link". The patient is automatically marked as "affected".
- Check that the settings match.
- Run the analysis by clicking on "Start analysis".

SomaVar Tertiary



SomaVar Tertiary is a VCF-based tertiary analysis pipeline that processes both Illumina WES VCF files and gene panel VCF files from various kit providers (see section **Input file format)**.

When running a SomaVar Tertiary analysis, users are invited to:

- Select the workset version to execute, then click on "Select",
- Select the sample(s) to be analyzed, then click on "Next Step".
- Launch the analysis by clicking on "Start analysis".

10.1.2. Choosing the UMI processing mode

Presentation of the UMI processing modes

In practice, the platform proposes one or more methods for processing UMIs when running an analysis:

1. ***Standard mode*** (recommended): consensus are generated from the PCR duplicates featuring the same UMI when their number is greater than or equal to 2. The UMIs represented by a single sequence (singletons) are also retained. The minimum allelic frequency of the variants reported and the minimum number of reads supporting the mutation (AO, Alternate Observation) differ according to the workset used:
    - In SomaVar, SomaHemato, SomaRNA, SomaCNVCapture, SomaCGP and SomaMSI, the minimum allelic frequency is 1% and the AO is a minimum of 3 reads.
    - In SomaVar LF and SomaLBx, the minimum allelic frequency is 0.1% and the minimum AO is 2 reads. This workset enables the detection of variants with an allele frequency of less than 1%. It is recommended when the sequencing depth is greater than 5000x, and for such applications as ctDNA sequencing.

2. ***UMI disabled*** (available in SomaVar, SomaRNA, SomaHemato, SomaLBx, SomaCNVCapture, SomaCGP and SomaMSI only): UMIs are not used for deduplication, and they are clipped from the end of the sequences before the analysis. Therefore, the data generated by this analysis pipeline are raw (non-deduplicated). The minimum allelic frequency of the variants reported is 1% and the minimum number of reads supporting the mutation (AO) is 3.

The differences in settings between these modes are summarized in the following table:

	UMI disabled	UMI standard
Minimum median/average quality of the mutated base (Phred score)*	30	30
Minimum number of reads by consensus	N/A	2
Minimum quality of each base in the consensus (Phred score)	N/A	30
Reads without filtered consensuses	N/A	no

Table 1: *Settings of the different UMI processing configurations*

** These filters only apply to variants that are not selected by the SVM model after variant calling (SomaVar and SomaRNA pipelines).*

UMI treatment process

The structure of the UMIs depends on each supplier kit and the following kits are currently available on the SeqOne platform:

- QIAGEN QIAseq
- Agilent XTHS/Low input
- Agilent XTHS V2
- Illumina TruSight Oncology 500
- Twist UMI Adapter System
- Roche KAPA HyperCap
- QIAseq targeted DNA Pro

⚠ Warning: The UMI configuration "QIAseq Targeted DNA Pro" should only be used for SomaVar, with a read length of at least 150 bp. For more information, contact support.

UMI processing is based on the use of the fgbio tool suite, which can handle UMIs and reads tagged with UMIs. Our UMI analysis pipeline consists of the following steps:

1. If necessary, trim UMIs from the end of sequences and generate an unaligned BAM from the FASTQ file pair,
2. Align sequences without UMIs with BWA-MEM2 (DNA/ctDNA) or STAR (RNA)
3. Merge the BAM files containing the UMIs with the aligned sequences using ZipperBam,
4. Group reads together if they have the same position and share the same UMI,
5. Generate the consensus sequence for each group of UMIs,
6. Filter each base according to its quality: confidence in the consensus sequence increases with the number of reads per group,
7. Filter consensus sequences: if more than 20% of the bases in the consensus sequence are filtered, the consensus sequence is not retained,

8. Conversion of these filtered consensus sequences into FASTQ and final realignment on the reference genome with BWA-MEM2 (DNA/ctDNA) or STAR (RNA).

Compatibility with the analysis worksets

The SomaVar, SomaHemato, SomaCGP, SomaLBx, SomaMethyl, SomaVar LF, SomaCNVCapture, SomaMSI and SomaRNA worksets are compatible with the UMI data. The processing options that are available when launching the different worksets are described in the table below:

	UMI disabled	UMI standard
SomaVar	Yes	Yes
SomaCNVCapture	Yes	Yes
SomaMSI	Yes	Yes
SomaRNA	Yes	Yes
SomaVar LF	No	Yes
SomaCGP	Yes	Yes
SomaLBx	Yes	Yes
SomaHemato	Yes	Yes
SomaMethyl	No	Yes

Table 2: *UMI processing options available when launching the compatible worksets.*

10.2. Analysis status

After launching an analysis, the SeqOne platform uses different statuses to allow users to track the progress of their analysis:

- ***Bioinformatics Pending*:** the platform is waiting for another analysis to end
- ***Bioinformatics Running***: the analysis is in progress; this status also shows which stage of the analysis is being performed
- ***Bioinformatics Finished*:** the analysis has been completed
- ***Analysis failed*:** an error occurred during the analysis; in this case, users are prompted to contact the user support team
- ***Interpretation - new*:** the analysis is finished and ready for interpretation, no user has opened it yet
- ***Analysis failed - error*:** The analysis encountered an error. In this case, the user is invited to contact the user support.
- ***Analysis failed - canceled*:** the analysis did not start because the sample attached to it encountered an error during loading or verification.

10.3. Sample quality

The project QC tab is divided into 2 sub-tabs: ‘General statistics’ and ‘ID check summary’.

General statistics. This tab displays a table representing the overall quality of alignment and coverage for each sample of the project. The user can export the table as a TSV file by clicking on  .

The calculation is based on the SeQC tool, a program developed by SeqOne to calculate alignment and coverage statistics from all reads aligned to the genome.

The user can choose to use the default QC metrics or define custom threshold values from the project's "Settings" tab (see section **Configure default project QC metrics**).

The proposed metrics are as follows:

- total number of reads
- average length of reads
- median size of DNA fragments in the library (inserts)
- percentage of quality reads >30
- percentage of reads aligned with the reference genome
- percentage of reads on target
- percentage of reads duplicated

- transition/transversion rate
- Average coverage
- Fraction of targets covered at different thresholds depending on the context (amplicon, capture).

The panel or exome targets are determined from the BED file selected when the project is created.

ID check summary. This tab details the genotypes of all the markers included in its analyses for each sample (see **Marker** paragraph in the **Managing user parameters** section). A row corresponds to a sample, and each column represents a marker with its associated ID. The user can export this table as a TSV file by clicking on .

11. Data Quality check

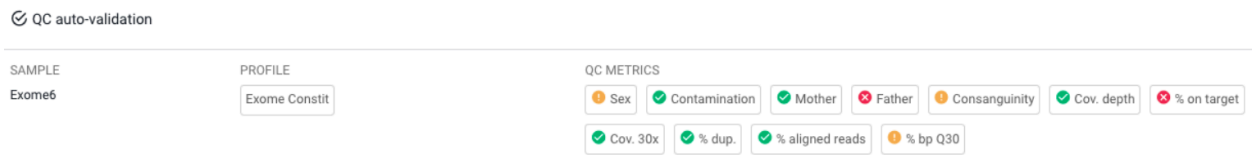
11.1. QC metrics and report

11.1.1. QC metrics and QC checks

From the *QC* tab, the user accesses the quality controls applied to the analysis, in the form of sub-tabs described in the sections below: *QC report*, *ID check* (if applicable) and *Technical report*.

In the case of a MicroArray CGH analysis, the QC tab is not available. However, the technical report is available from the *Files* tab of the analysis in *report.html* format.

In the upper part of the QC tab, a "*QC auto-validation*" section allows the user to color-code QC metrics as *Passed*, *Warning* or *Failed*. By clicking on the associated icon, the user accesses the thresholds applied and the QC details: sex, contamination, parentage and inbreeding, depth of coverage, % on target, 30X or 100X coverage, % duplicates, % reads aligned with the reference genome or % bp Q30.



11.1.2. QC report

The global *QC report* contains various metrics describing the quality and alignment of the sequences, the coverage of the regions in the manifest, and some variant-related statistics. This report is composed of the following sections:

Alignment details (SomaMethyl). Total reads processed, uniquely and non-uniquely aligned reads, and mismatch rates.

Deduplication (SomaMethyl). Duplication metrics for aligned reads, highlighting potential issues with library preparation.

Cytosine coverage & methylation (SomaMethyl). Cytosine-to-uracil conversion rates for each context (CpG, CHG, CHH).

UMI groups statistics (*SomaVar, SomaVar LF, SomaMethyl and SomaRNA UMI only*). See paragraph **Choosing the UMI processing mode**.

The QC report of SomaVar, SomaVar LF, SomaMethyl and SomaRNA analyzes offers a more detailed view of the composition of the sample, in particular the distribution of UMIs according to the number of sequences that harbour them.

Statistics describe in particular:

- The fraction of singletons (UMI carried by a unique sequence)
- The mean and median size of the UMI groups, as well as the standard deviation
- The size of the largest group, and the fraction of total sequences it includes.

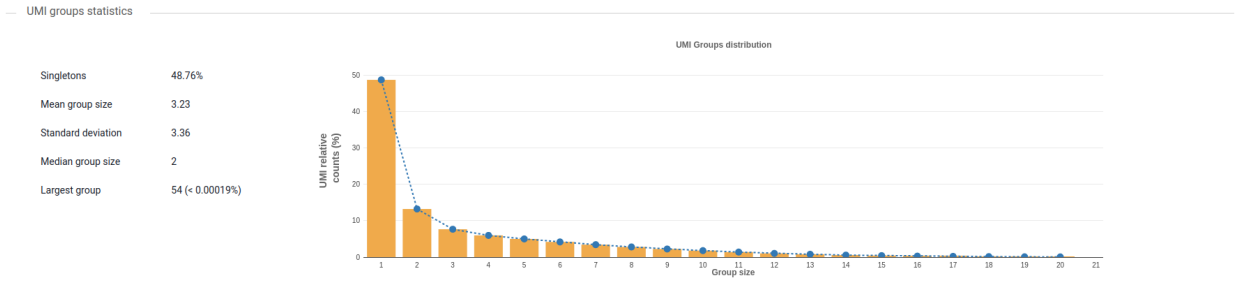


Figure 7: Statistics relative to UMI distribution in terms of the number of sequences within a sample.

The global QC metrics for the project are also more detailed and divided into three sections:

- QC metrics after UMI processing, which indicate the trend in each metric (as a percentage) based on the raw data.
- Raw QC metrics calculated from all the sequences (before UMI processing).
- UMI distribution in terms of the number of sequences within each sample.

Cross-contamination estimate (*SomaVar LF only*). Cross-contamination is estimated individually using the GATK CalculateContamination tool. The latter uses the signal of the reference alleles detected at homozygous loci in order to calculate the fraction of sequences in the individual originating from cross-contamination.

Sample integrity and identity (Germline analyses only). This section lists metrics related to sample integrity and identity:

- ***Family link*** : user-defined metadata
- ***Original sex***: sex-related metadata entered by the user (Male/Female/Unknown)
- ***Detected sex***: sex detected (Male/Female/Unknown)
- ***Relationship coefficient*** : relationship coefficient calculated using the *Somalier* tool for each sample in the family analysis, this metric displays N/A for the sample itself.
- ***Contamination***: estimated percentage of sample contamination by foreign DNA, using the *verifyBamID2* tool.

Also listed are the number of homozygous and heterozygous variants, the ratio of these two values (Hom/Het ratio), the number of homozygous and heterozygous variants on chromosome X, the equivalent ratio on chromosome X (Het/Hom ratio on chr X), and the number of variants on chromosome Y.

The *Family link* and *Relationship coefficient* metrics, used to check the relationship between individuals in a family, are only displayed in trio family analyses (GermlineVar Family or GermVar Tertiary Family).

Gender metrics are used to check the correspondence between the gender entered and the gender detected in the data.

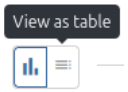
The contamination metric is not calculated for GermVar Tertiary and GermVar Tertiary Family worksets.

QC checks on gender matching, contamination and relationship are available at project or analysis level to easily identify any matches or mismatches.

Alignment metrics. This section lists the main metrics calculated at the end of the alignment, such as:

- The total number of sequences in the sample,
- The median size of the inserts, calculated from the distance between the two reads of the same pair, whose respective length is included,
- The average sequence length,
- The fraction of sequences successfully aligned to the reference genome,
- The duplicate rate (capture only),
- The fraction of bases whose quality is greater than 30,

- The average coverage of the targeted regions, determined from the analysis manifest.

Coverage. A bar graph representing the fraction of target regions covered at different thresholds, shown on the x-axis. Clicking the designated button 

allows users to view coverage values in a table format, which can then be copied and pasted.

QC graphs. Graphs from the FastQC quality analysis, illustrating in particular the GC content of the sequences as well as the quality of the bases according to their position in the sequences.

Variant statistics. A list of statistics around the variants detected in the analysis, which include in particular:
- The number of transitions (Ts) and transversions (Tv) among the detected substitutions,
- The ratio Ts / Tv,
- The total number of mutations detected,
- The fraction of SNVs, MNVs and indels in this total.

11.1.3. 'ID-Check' identity monitoring report

The 'ID check' tab is dedicated to the results of the identity monitoring in the QC tab of worksets calling variants.

The results are displayed in table form with the following information:
- A description of each marker based on genomic coordinates, for example: chr-start-ref-alt,
- a label (rsID, custom label) when given by the user,
- observed genotype (GT) or allele frequency (VAF),
- the depth of coverage at the marker location (COV),
- the quality of the database (QUAL).

When several samples are included in the analysis, the sample-specific columns (GT/VAF, QUAL, COV) are repeated for each sample individually.

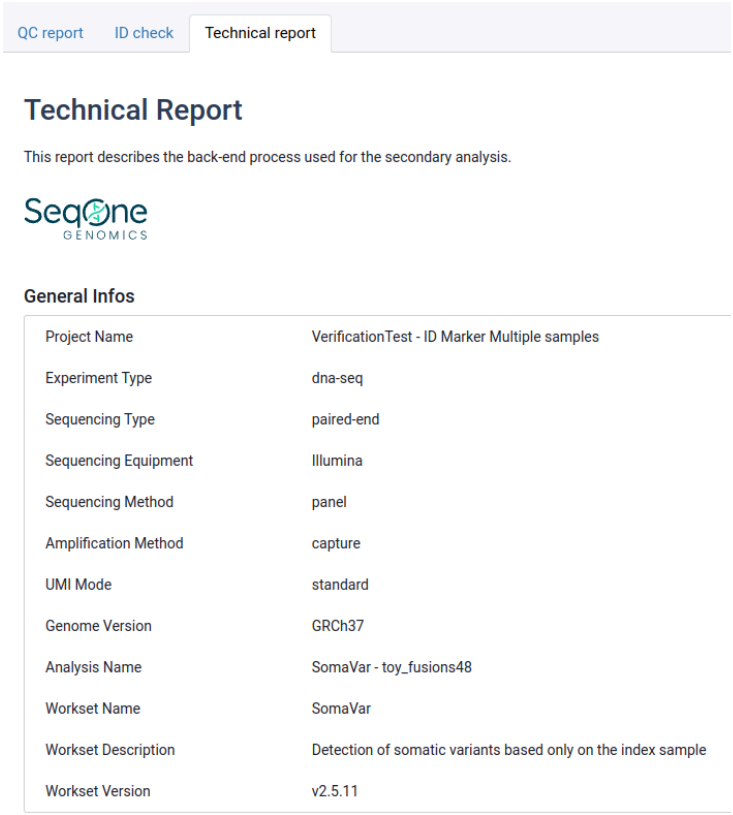
Selecting a variant in the table opens an IGV browser in a new tab, centred on the genomic coordinates of the variant.

11.1.4. Technical report

The technical report lists the tools used during the analysis and their version. It can be accessed from the QC tab of an analysis, then under 'Technical report'.

In particular, it describes :

- general information on the configuration and metadata of the project and the analysis, such as the type of sequencing, the genome version and the workset version used;



- details of the manifest used, such as name, size, length (bp) and padding added to the analysis;
- details of the databases and bioinformatics tools used.

11.1.5. UPD report

The UPD Report is available in the QC section of compatible germline analyses (e.g., GermVar and related workflows). It summarizes the UPD detection results and provides visual elements to support interpretation.

The UPD Report typically includes:

- Result overview: indication of whether the analysis detected a UPD-like event and the type of event (e.g., complete or segmental; in trio analyses, the report can also support parental origin and event subtypes such as UPD | Iso / UPD | Het).
- Event list: a list of detected events with their chromosomal location and size, and the associated score(s) used by the algorithm.
- Visualization plots: plots based on SNP genotypes (e.g., B-allele frequency (BAF) and homozygosity/ROH signal) that help confirm the presence of long homozygous regions and boundaries.
- Quality control information: metrics related to the reliability of the detection (e.g., SNP density / number of SNPs) and warnings when the data are not sufficient for confident calling.

Detected UPD events are also displayed in the Large Variant Viewer (LVV) with specific labels (e.g., UPD | Iso and UPD | Het).

UPD Score

The UPD score is a numerical indicator used to quantify how strongly the sample shows a UPD-like signal along the genome. UPD detection in SeqOne relies on SNP genotype patterns:

- In a *singleton* (solo) analysis, the algorithm searches for large regions of homozygosity (ROH) and for abrupt changes in SNP distribution along chromosomes.
- In a *trio* analysis, the SNP pattern is additionally used to determine the parental origin of the event and to distinguish isodisomy (UPD | Iso) from heterodisomy (UPD | Het).

In practice, the UPD score reflects the strength and consistency of the signal supporting a UPD event (e.g., clear ROH/BAF pattern and well-delimited boundaries). A higher score indicates that the event is more likely to be a true UPD event, while lower values indicate weaker or noisier signals.

The UPD score must always be interpreted together with the elements provided in the UPD Report tab (QC section), including:

- the UPD visualization plots (e.g., B-allele frequency / homozygosity signal),
- the size and chromosomal location of the event,

- the analysis quality metrics (e.g., SNP density / number of SNPs).

B-allele frequency (BAF) plot from UPD report

The B-allele frequency (BAF) plot is a visualization of SNP genotypes along the genome and is used to assess allelic imbalance and loss of heterozygosity (LOH).

For each SNP position, the BAF corresponds to the fraction of sequencing reads supporting the "B" allele and ranges from 0 to 1:

- Values close to 0 correspond to a homozygous genotype (AA).
- Values close to 0.5 correspond to a heterozygous genotype (AB).
- Values close to 1 correspond to a homozygous genotype (BB).

In a normal diploid region, the BAF plot typically shows three bands around 0, 0.5, and 1. In long regions of homozygosity (ROH) or UPD isodisomy, the heterozygous band (around 0.5) is reduced or absent, which supports the presence of a UPD-like signal.

Important limitations:

- A UPD-like signal can also be observed in the context of parental consanguinity (multiple ROH regions across the genome).
- UPD calling is primarily based on SNP distribution and may not be sufficient alone to distinguish some UPD patterns from copy-number related events; when needed, cross-check with other available evidence (e.g., coverage / CNV results).

12. Interpretation of analysis results

12.1. Analysis results

Users can access a project's analyses by clicking on the "*Analyses*" icon in the associated box or by clicking on the icon of the same name in the top-left corner of the user dashboard.

When an analysis has been completed and is available for viewing, it is shown with the "*Interpretation - New*" status.

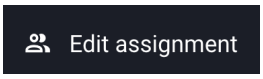
The analysis results are then available as several reports, which can be opened directly by clicking on them.

There are several variations of the "*Interpretation*" status, depending on the progress achieved in interpreting the results:

- ***New*** - the variants table has not yet been viewed
- ***In progress*** - the variants table has been viewed, but no variants have yet been validated
- ***Evaluating*** - at least one variant has been validated
- ***Reporting validated*** - a report has been generated for the analysis

12.1.1. Managing Analysis Interpretation Workflow

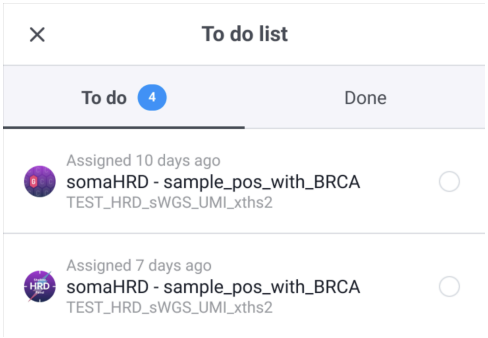
12.1.1.1. Assigning analyses for interpretation

Entity administrators can choose to assign one or more analyses to be interpreted by different entity users from the Analyses tab of the project. To do this, the administrator can select one or more analyses, then click on the button 

at the bottom of the page.

An "Assignment" window allows you to assign one or more users to the selected analyses, by searching for the user's identifier (username) via the dedicated search bar. It is also possible to modify the assignment by adding or deleting users. Users to whom an analysis is assigned are listed in the "Assigned to" column. Users who have never logged into SeqOne cannot be assigned until they create their credentials.

Users can access the analyses assigned to them (To do tab) and the validated analyses (Done tab) in their 'To do list', accessible via the icon at the top right of the page. A notification appears for each new assignment.

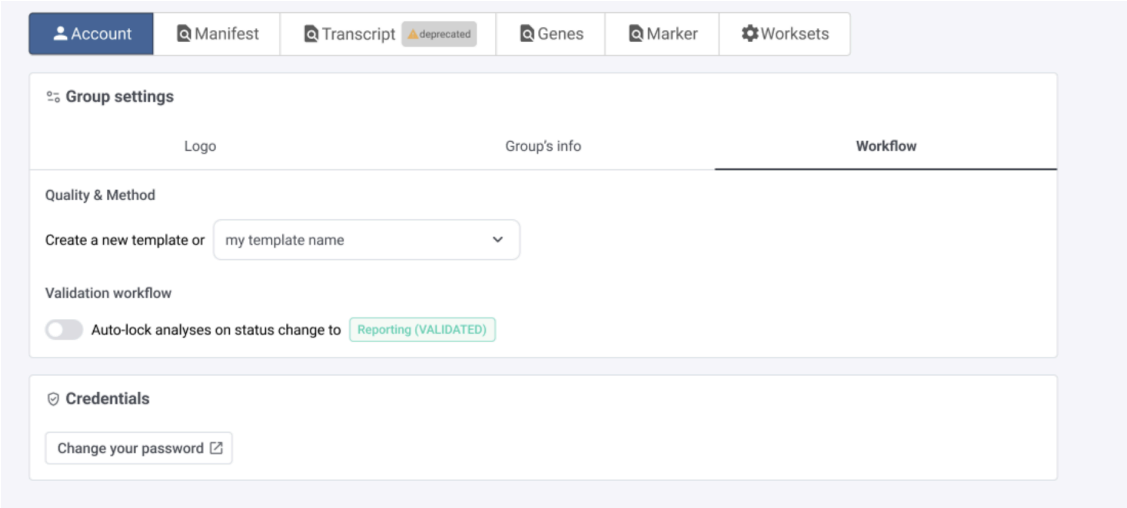


Users can access assigned analyses directly by clicking on them. They can validate their task by clicking on the corresponding checkbox in the 'To do' tab to move it to the 'Done' tab. This action is reversible.

Analyses are displayed in the following order according to their status: completed, running or failed. For each analysis, the assignment date, analysis name and project name are detailed.

12.1.1.2. Locking analysis

The administrator of an entity can lock or unlock an analysis. In the entity's *Settings*, from the "*Account*" tab then "*Workflow*", they can activate automatic locking of validated analyses using the "*Auto-lock analyses on status change to Reporting (VALIDATED)*" feature:



At the end of his interpretation work, when a user changes the status of the analysis under "*Reporting - validated*", the analysis is automatically locked. Locked analyses display a "Reporting - Validated & Locked" status, visible with the icon Reporting VALIDATED & LOCKED .

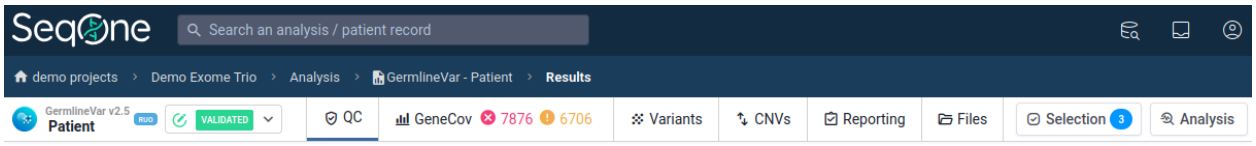
When an analysis is locked, users cannot perform the following actions from the interface :

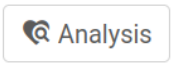
- add or modify variant evaluations in VKB,
- add or modify comments associated with variants,
- generate a new report from the *Selection* window,
- generate a new HRD report,
- manually modify analysis status,
- perform one of the following actions in the analysis panel:
    - modify assay assignment (under "Assignment")
    - modify patient gender
    - add or modify HPOs
    - modify analysis panels
    - delete clinical report

If a user attempts any of these actions, he/she will be prompted to unlock the analysis, if he/she has permission, or to contact his/her administrator if he/she does not. A user wishing to unlock an analysis must have "Unlock analyses" permission.

12.1.1.3. Side panel for analysis management

The Analysis side panel a.k.a "Analysis Drawer" can be accessed from any analysis tab.



By clicking on the button 

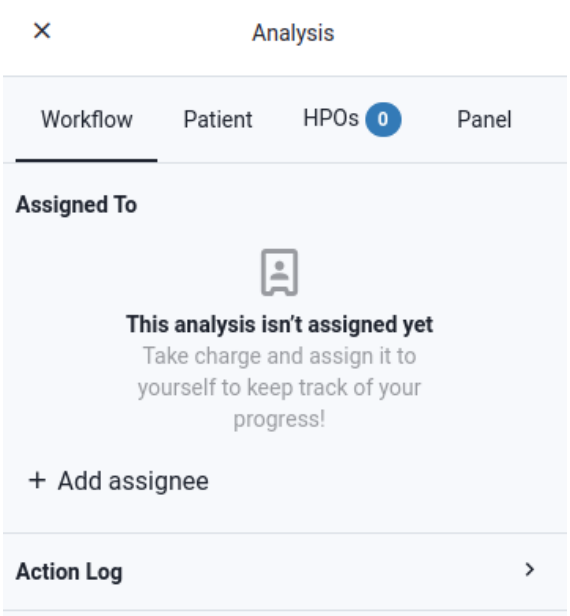
in the top right-hand corner in the analysis tabs banner, the user can directly view and manage analysis data and patient HPOs without leaving the current screen.

The Analysis Drawer contains 4 views:

- **Workflow**: allows you to organize and trace your workflow for analysis
- **Patient**: allows you to view and modify patient metadata
- **HPOs**: allows you to add or modify HPO terms in the analysis, only available for analyses in constitutional context.
- **Panel**: allows you to apply a list of genes to the analysis and filter the results of the analysis on the basis of this list.

Workflow modal view

This view allows you to organize and track user workflows for the analysis being interpreted.



Assigned To. This view displays the tasks assigned for the current analysis. Only administrators or users with 'Manage assignments' permission can assign tasks (see **Admin management - interface presentation** section).

When tasks are assigned to the current analysis, a list is displayed of all assigned users, together with the date and time of completion. This list is updated automatically as soon as a user completes a task.

Users can only edit their own tasks by checking or unchecking the corresponding box. When hovering over a read-only task, a tooltip states 'You can only complete yours'.

If no tasks have been assigned, the view displays 'This analysis isn't assigned yet'.

If an unauthorized user attempts to assign a task by clicking on the '+ Add assignee' button, the error message indicates "You must have administrator privileges to assign analysis. If needed, please contact your platform administrator to request the appropriate access."

Action Log. The "Analysis Action Log (Audit Trail)" banner displays the history of all actions performed on HRD status in SomaHRD analysis, filters, filter profiles and restricted gene panels in the variant table, detailing the action type, user, timestamp and a description of the action.

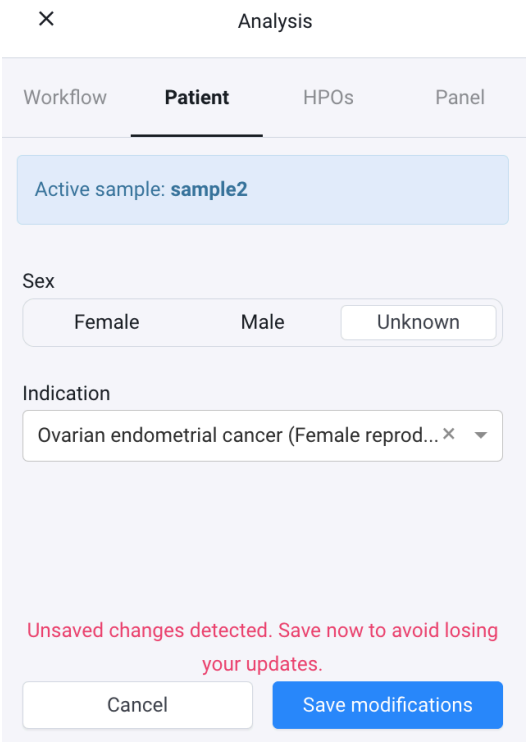
The most recent actions appear at the top of the list. Pagination allows users to efficiently navigate through the action history. The entire history can be downloaded in .csv format via the "Download Log" icon.

Patient modal view

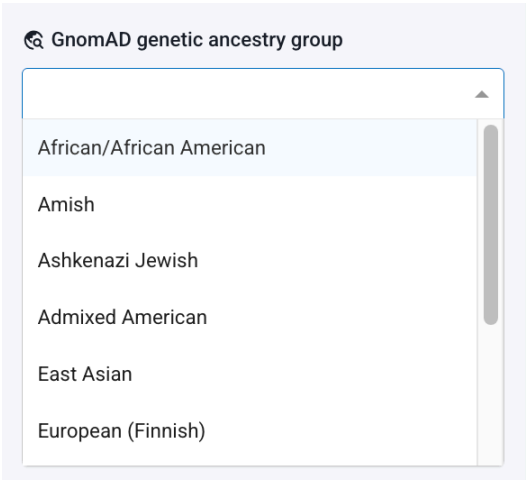
This view allows you to view and modify patient metadata such as gender (Female, Male or Unknown), the GnomAD genetic ancestry group (Germline only) or the Indication (Somatic only).

Displaying and editing the indication is only available for **somatic** analyses:

- If the indication was entered prior to analysis (see **Sample metadata**), it will be automatically defined for the sample. It will then be marked with a star and appear first in the list.
- If the indication is changed during interpretation, a warning message indicates that the indication selected is different from the original indication and the pill colouring in the 'A' sub-column in the *Clinical Evidences* column will be updated (see **Selecting an indication**).
- If no indication is available for analysis, the message 'No available indication' will be displayed.



Displaying and editing the "GnomAD genetic ancestry group" is only available for **germline** analyses. The default value is empty and is treated as the combined GnomAD population.



Modifications made in this view must be saved or cancelled by clicking on the corresponding "Save modifications/Cancel" icons at the bottom of the pane, in order to change tabs.

HPOs modal view

This view, available only in Germline projects, allows you to view, edit and assign HPO terms relating to the sample, either manually or from a clinical description of the patient.

The user can visualize the list of HPOs terms already associated with the sample prior to analysis. In the absence of HPOs codes, the message "No HPOs yet" is displayed.

To add terms manually, the user can enter them from the "Enter phenotypes one by one" search bar, by searching for a phenotype or HPO code. They can also import a text file containing HPO codes by clicking on the icon  (see also **Sample metadata** section).

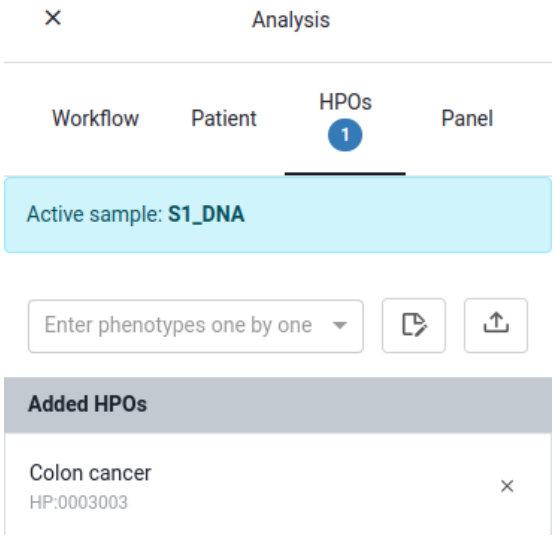


Figure 8: HPOs view of the analysis panel

The user can generate HPO terms based on the patient's clinical description using DiagAI. By clicking on the icon  , a dialog window appears, prompting the user to describe the patient's clinical profile (this description is limited to 4000 characters) and then click on "Extract HPOs with DiagAI" to generate a list of HPOs found. HPO terms are highlighted in color and the user can associate them with the patient by hovering the mouse over them and clicking on "Add", or add all HPO terms found by clicking on "Add all" (see also section **Edit by sample metadata with analysis panel**).

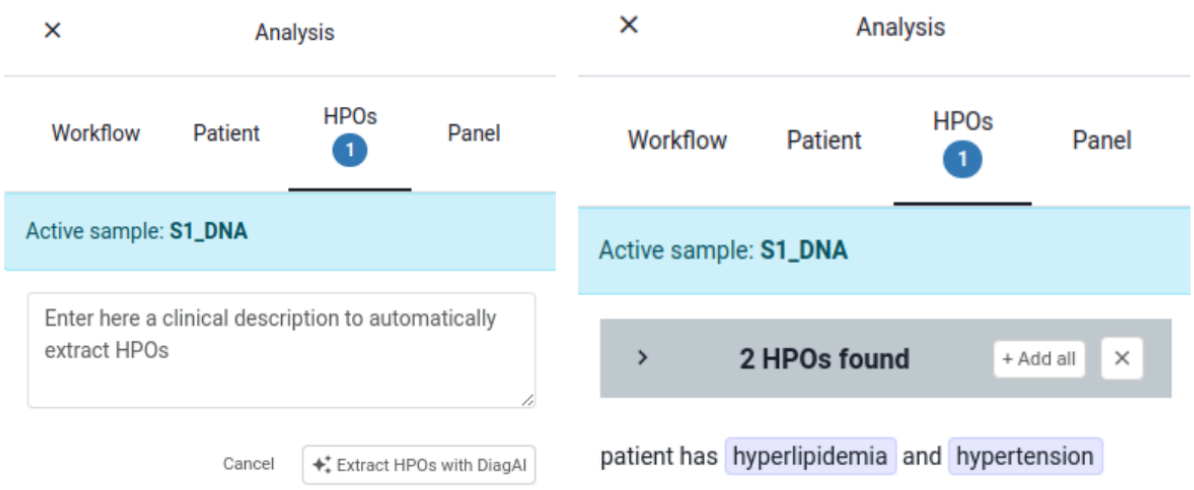
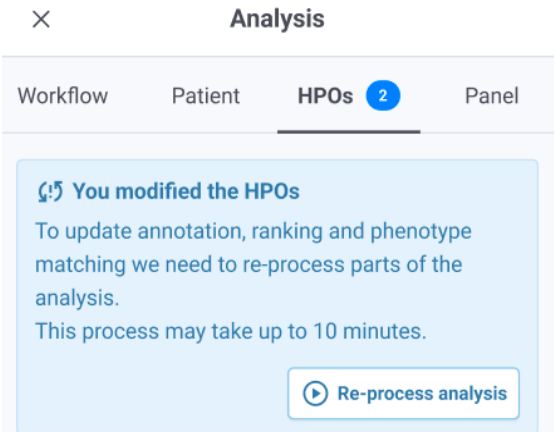


Figure 9: Extracting HPOs terms with DiagAI from the HPOs view in the Analysis Drawer

When HPOs are added or modified from this view, a blue banner appears, suggesting that the analysis be reprocessed with the new HPOs.

By selecting "Re-process analysis", the user can update the DiagAI score and the classification of variants (Short list or SmartPick) according to the new HPOs entered.

This re-processing step freezes the variant table for several minutes in order to update the variant classification and DiagAI column.

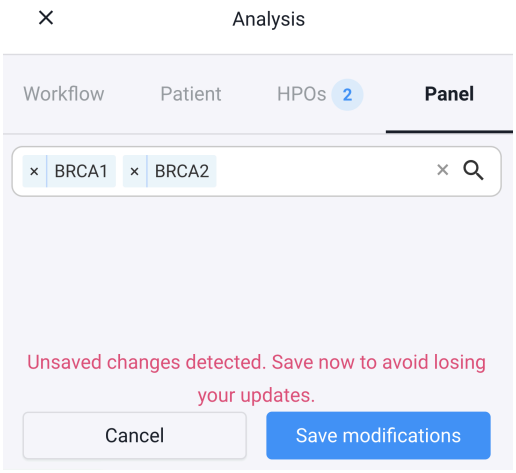


Panel modal view

This view is available for all worksets except CNV analysis worksets (CNVExome, CNVCapture and SomaCNVCapture).

It gives access to the "*Restricted panels"* feature, which allows you to restrict and lock the analysis to selected genes, in the manner of an *in silico* panel.

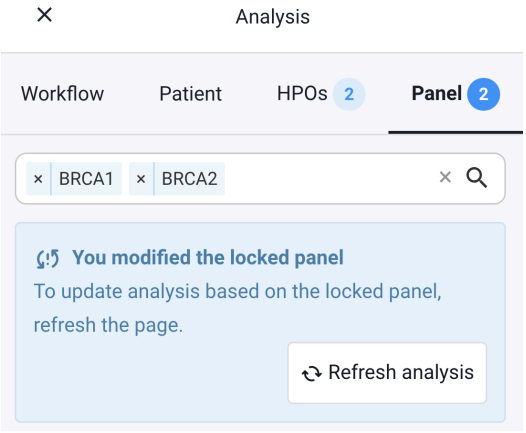
To do this, the user can select one or more lists of genes to apply, via a drop-down menu, from those previously imported (see **Genes** paragraph **,** in the **User parameters management** section)**.** By saving the selected panels via 'Save modifications', a filter is applied to the entire analysis according to the gene panel(s) selected.



Restricted panels can be selected as soon as the analysis is launched (see section **Launching a new analysis**) or from the Panel view after analysis.

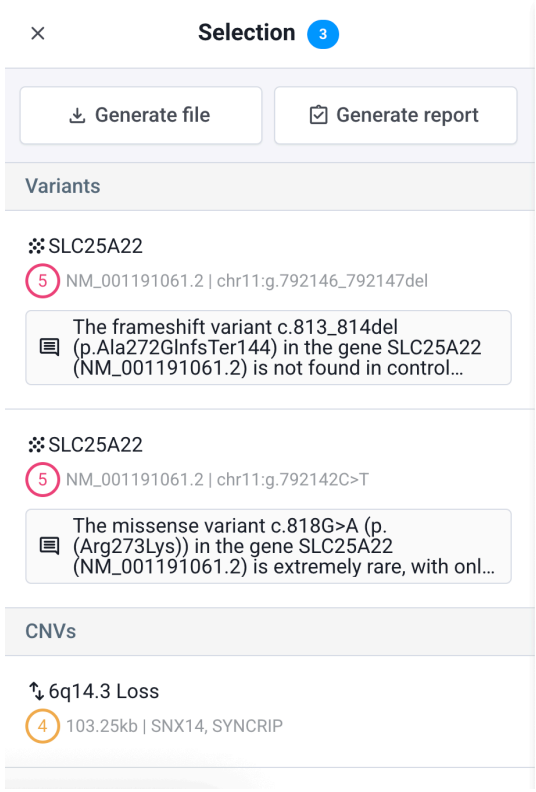
When gene panels are added or modified, a blue banner warns the user of the change ("You modified the locked panel") and suggests refreshing the analysis with the new panels via the "*Refresh analysis*" function.

By clicking on "*Refresh analysis*", the user chooses to update the variants table, the CNVs tab (if applicable) and the GeneCov tab according to the new panel.



12.1.1.4. Side panel for managing assessments

The panel Selection can be accessed from any tab within an analysis. It allows the user to view the small variants and CNVs annotated in VKB for the current analysis, to generate a JSON file ("Generate file") and to edit a clinical report ("Generate report").



CNVs validated in VKB are also available in this section, only in the case of CNVs from exome, genome and CGH Microarray analyses.

Small variants validated in VKB are differentiated by the logo , while CNVs are displayed with the logo and the words ‘gain’ or ‘loss’.

12.1.2. Overview of analysis results in the Dashboard tab

For some worksets, a "Dashboard" tab that summarizes all results and QC metrics is accessible to users and appears by default when opening an analysis. The "Dashboard" tab provides an overview of analysis results across multiple sections. The sections available vary depending on the type of workset. This tab is available for SomaCGP, SomaVar Tertiary, and SomaLBx.

Patient. This section summarizes the sample metadata: sample name (*Sample ID*), sex and pathology indication (Indication).

Quality Controls. This section provides information on several key metrics related to sample and sequencing quality:

- Tumor purity,
- Sample ploidy,
- the percentage of bases whose quality is greater than or equal to 30 (% Q30 bases),
- average coverage of regions of interest (ROI average coverage) before and after UMI processing,
- percentage of regions of interest with coverage greater than 200X (ROI > 200X) before and after UMI processing.

Signatures. The section on signatures shows the key figures for TMB, MSI and HRD analyses:

- Calculation of TMB (Tumor mutational burden): number of mutations/Mb.
- Microsatellite instability: percentage of unstable sites.
- HRD status (Homologous Recombination Deficiency).

TMB and MSI scores statuses can be selected by the user via the dropdown menus located under each score.

Selection. This section displays the variants evaluated in the VKB database for the current analysis.

Suggested by SeqOne. The genomic alterations detected in the analysis are presented as a reduced list of clinically relevant variants. Variants are selected if their ComPerMed class is 5, or if they have a Tier I assertion corresponding to the patient's indication, or a Tier II assertion for experimental therapies related to this indication, or a Tier I assertion for another indication. The user can add variants from this category to the Selection section by clicking on "Add".

The JSON file can be generated by clicking on "Generate JSON file" and contains all the information from the "Dashboard" page.

12.1.3. Detailed coverage by exon

In the *GeneCov* tab, users can view the coverage rate of the genes included in the analysis manifest for each exon, whether a gene panel or exome.

Methodology

This module intersects the user manifest with RefSeq annotations for all genes included in the panel or the exome, keeping 10bp on each side in order to cover splice junctions, then proceeds to calculate coverage depth for each base included in this intersection using the mosdepth tool.

Different thresholds are set in order to assign each exon a status (covered, warning, failed) according to the percentage of bases below these coverage thresholds.

The status is defined from the base coverage in the exon including the adjacent 10bp.
If UMIs are used, coverage is calculated after deduplication.

The thresholds that apply for the different types of analyses are detailed in the table below*:

	GermlineVar /Germline Family/ GermVar / GermVar Family				SomaVar		SomaRNA		SomaVar LF		SomaCGP	
	Panel		Exome		Panel		Panel		Panel		Panel	
	Warning	Failed	Warning	Failed	Warning	Failed	Warning	Failed	Warning	Failed	Warning	Failed
Capture	**Cov. 100x** < 100%	**Cov. 30x** < 100%	**Cov. 30x** < 100%	**Cov. 10x** < 100%	**Cov. 500x** <100%	**Cov. 200x** <100%	**Cov. 200x** <100%	**Cov. 50x** <100%	**Cov. 5000x** <100%	**Cov. 2500x** <100%	**Cov. 100x** < 100%	**Cov. 50x** < 100%
Amplicon	**Cov. 100x** < 100%	**Cov. 30x** < 100%			**Cov. 1000x** < 100%	**Cov. 500x** < 100%	**Cov. 200x** <100%	**Cov. 50x** <100%				

Table 3: *Coverage thresholds for the Gene Coverage module according to the analysis type.*

If the user has predefined custom QC metrics from the project parameters, then these will be applied in the GeneCov tab (see section **Configure default project QC metrics).*

View and interpretation

Each gene can be visualized on a graphic representation, and results from Gene Coverage can be filtered based on genes, in silico panels and transcript lists. Targeted exons are represented in either green (*Covered*), orange (*Warning*) or red (*Failed*) according to the thresholds described above.

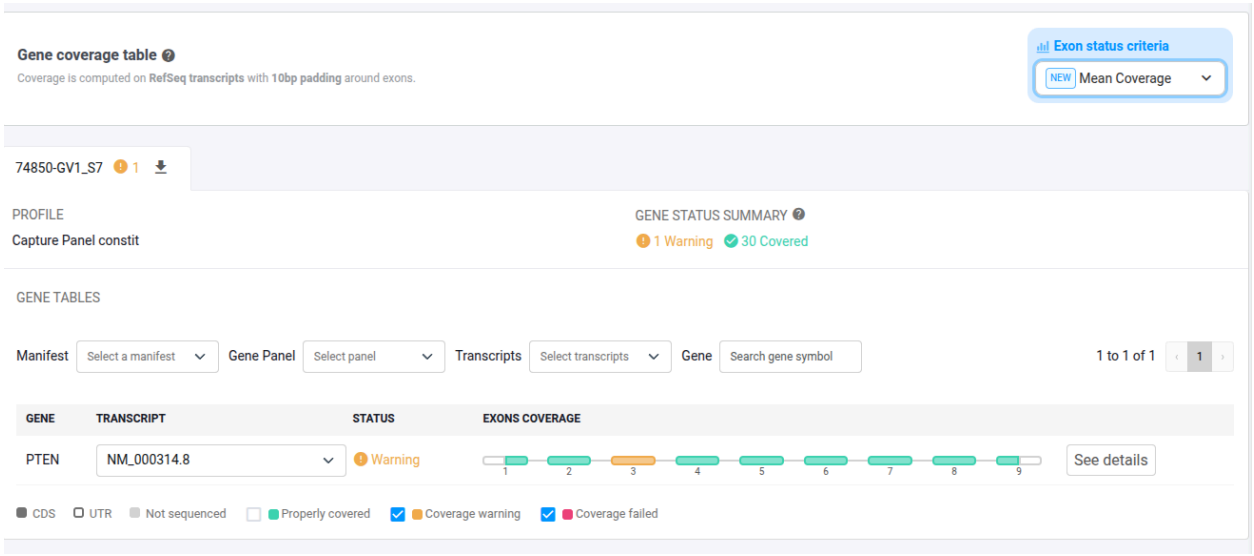


Figure 10: Details of the GeneCov tab

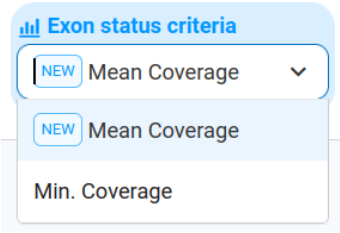
Filter options

A menu of filters allows you to:

- restrict the display to a subset of regions using a .bed file previously imported into the interface, via the "Manifest" filter,
- apply a list of genes via the "Gene Panel" filter,
- apply a list of transcripts via the "Transcripts" filter,
- search for a specific gene via the "Gene" filter.

Coverage calculation methods

The user can select the coverage calculation method to be applied using the "*Exon status criteria*" drop-down menu at the top right of the page:



1. **Mean Coverage** (default): Uses the average coverage depth over the entire exon to determine coverage status.
2. **Minimum Coverage** : Uses the minimum fraction of bases that meet a specific coverage threshold. This ensures that a certain percentage of bases have at least the minimum required coverage depth.

The **"Gene status criteria"** drop-down menu allows users to apply these calculation methods at the gene level:

- **Mean coverage (default):** Uses the average coverage depth across all exons of the gene to determine coverage status.
- **Worst exon status:** Determines gene coverage status based on the worst-performing exon. If any exon fails to meet the coverage threshold, the gene is considered not covered. This was the only supported method in previous versions.

Data displayed

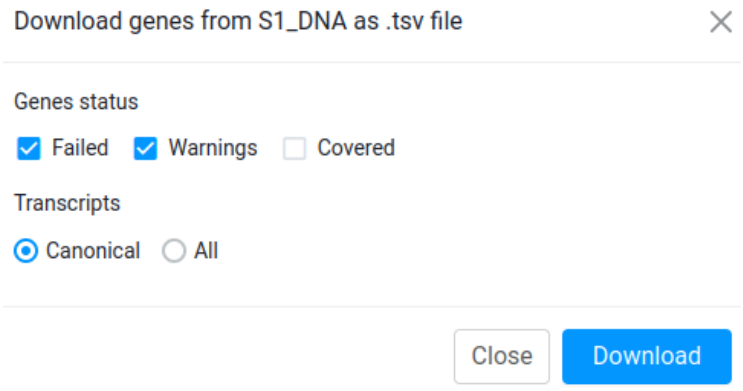
Exons that are not or partially covered by the design (< 80% overlap between manifest interval and exon RefSeq boundaries) are considered non-targeted. These exons are represented in gray (Not sequenced) on the graph.

A detailed view of each gene can be used to display the exon coverage metrics by clicking on "*See details*": average coverage, median, percentage of each region covered at 10x, 30x, 50x, 100x, 200x, 300x and 500x.

For each exon, an IGV window allows users to view the sequence coverage and alignment in greater detail.

GeneCov data export

The complete coverage analysis data is available as a sample_genecov.v2.tsv.gz file, accessible from the Files tab of the analysis.



From the GeneCov tab, the user can export the coverage analysis according to the coverage calculation method (Mean Coverage or Min Coverage), via the icon next to the sample name in the top left-hand corner of the sample tab.

The user can choose to export data according to gene coverage status (Failed, Warning, Covered) and to export the canonical transcript only (Canonical) or all available transcripts (All) before clicking on Download to obtain the desired file.

Data is exported as a .tsv file containing the following columns:

- Gene (gene_symbol),
- Gene number according to HGNC (gene_id),
- Canonical transcript (is_canonical_transcript),
- Transcript status (transcript_status)
- Exon (exon_number)
- CDS or UTR exon type (exon_type)
- Exon covered or uncovered status (exon_status)
- Region covered by manifest or not (target)
- Genomic position (interval_id)
- Chromosome number (chr)
- Start position (start)
- End position (end)
- Forward or reverse strand (strand)
- % coverage at 1, 10, 20, 30, 50, 100, 200, 300, 500 and 1000X (cov_)
- Average coverage (cov_avg)
- Median coverage (cov_med)

12.1.4. Small variants

Methodology

The detection of small variants is handled by a combination of several tools, which may vary from one workset to another.

Whether for exome or gene panel analysis, variant calling is determined by the manifest (in BED format) associated to the project, including the intervals of the regions of interest and the padding applied to these regions (see section **Create a project**).

In addition, variants are called taking into account the alignment quality of reads on the human genome (GRCh37 or GRCh38 version). So-called "MAPQ0" reads with an alignment (mapping) quality equal to 0, because they align at several sites in the genome, are not taken into account when calling up variants for somatic analysis worksets. In "Germline" analyses, MAPQ0 reads are considered for variant calling if a manifest specifying low-quality alignment regions is entered when the project is created under the Low-Quality Mapping Region metadata. Variants detected from MAPQ0 reads are identified in the variants table by a "_mapq0" tooltip visible in the CAL (Caller) column.

SNVs / indels. For the detection of SNVs and indels :

- *Freebayes*[1] is used as the sole variant caller by the GermlineVar (exome), GermlineFamily (exome), SomaRNA and SomaVar LF worksets. The lower detection thresholds are as follows:
    - VAF ≥ 10% in GermlineVar for the exome,
    - VAF ≥ 5% in GermlineVar for gene panels,
    - VAF ≥ 1% in SomaRNA,
    - VAF ≥ 0.1% in SomaVar LF,
- It is accompanied by *GATK HaplotypeCaller*[2] in the GermlineVar workset (panel).
- Its somatic counterpart, *Mutect2*[3], is used alongside Freebayes in the SomaVar, SomaHemato and SomaCGP worksets. The minimum detection threshold is VAF ≥1% for these two callers in SomaVar.
- *Sentieon*[4] is used as a major variant caller by GermVar, GermVar Family and SomaLBx worksets.
- *Clair3*[5] is used as the sole variant caller by GermlineVarLR workset.

[1] https://github.com/freebayes/freebayes
[2] https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller
[3] https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2
[4] https://github.com/Sentieon
[5] https://github.com/HKU-BAL/Clair3

GermVar and GermVar Family analyses rely on Sentieon bioinformatics components for key secondary analysis steps. In particular:

- Alignment is performed using Sentieon's optimized implementation of BWA-MEM2, and duplicate reads are marked.
- Germline calling (SNVs/Indels) is performed using Sentieon DNAScope[6].
- Sentieon TNScope[7] can be used to detect low-VAF mosaic variants (depending on the GermVar configuration).

Note about low-mappability regions (MAPQ0): in highly homologous regions (e.g., SMN1/SMN2), variants may be filtered out by the caller's machine-learning filtering. When required, SeqOne provides an optional MAPQ0 (low-quality regions module) based on Freebayes with adapted thresholds to increase sensitivity in user-defined regions.

SomaLBx analyses rely on Sentieon bioinformatics components for key liquid-biopsy secondary analysis steps. In particular:

- Read alignment is performed with Sentieon (BWA-MEM), and the pipeline includes Sentieon-based UMI processing/consensus to support high-sensitivity ctDNA workflows.
- Somatic calling of short variants (SNVs/Indels) is performed using Sentieon TNScope[7] in tumor-only mode, with very low allele fraction support (VAF ≥ 0.1% depending on configuration).
  Note: variant genotypes are reported as heterozygous by default in SomaLBx SNV/indel results because Sentieon TNscope does not perform genotype prediction.

SVs - Structural Variants. Insertions and deletions whose length is close to or greater than that of a sequencing read, are detected using the SV caller GRIDSS[8] in the GermlineVar (panel/exome), SomaVar, SomaVar LF SomaHemato, and SomaCGP worksets.
These events are detected using the SV caller Delly[9] in the GermVar and GermVar Family worksets. Only insertions and deletions smaller than 300bp are detected in somatic worksets, while a 10kb limit is applied in the GermlineVar workset (for versions ≥ 2.0), GermVar and GermVar Family worksets (versions ≥ 3.5). Variants detected by GRIDSS are detected whatever the VAF.

FLT3-ITD. A specific detection module for internal tandem duplications of the FLT3 gene (FLT3-ITD) is only available in the SomaVar and SomaHemato worksets, from version v2.2. This module is based on the *FiLT3r*[10] tool.

[6] https://github.com/Sentieon/sentieon-dnascope-ml
[7] https://github.com/Genomic-Medicine-Linkoping/TNscope
[8] https://github.com/PapenfussLab/gridss
[9] https://github.com/dellytools/delly
[10] https://gitlab.univ-lille.fr/filt3r/filt3r

Alu Insertions. Finally, a module for detecting insertions of transposable elements (*AluMEI*), such as Alu elements, has also been developed by SeqOne. It is used by the GermlineVar (panel/exome), GermlineFamily (panel/exome), GermVar (panel/exome/genome), GermVar Family (panel/exome/genome), SomaHemato and SomaVar worksets. No lower VAF threshold is applied for variants detected by AluMEI.

The combined *Freebayes*, *HaplotypeCaller* (or *Mutect2*) and *FiLT3r* caller variants are available in the analysis's *final.vcf.gz* file. Variants specific to *GRIDSS*, *FiLT3r* or *AluMEI* are available in their respective VCF files (**gridss.vcf.gz, filt3r.vcf.gz** and **variants.alumei.vcf.gz**), also available for download from the *Files* tab of the analysis.

In the specific case of *FiLT3r* :

- *FiLT3r* variants common to *Freebayes* and *Mutect2* are available in the final VCF of the analysis (**final.vcf.gz**),
- variants unique to FiLT3r as well as variants common to FiLT3r and GRIDSS are listed in the FiLT3r VCF (**filt3r.vcf.gz**).

Two additional modules developed by SeqOne then enable :

- Retention of multinucleotide variants, or MNVs, that are coding and restricted to a single codon during the process of decomposing delins-like variants.
- Recalibration of the allelic frequency of indels and delins variants detected by *Freebayes* and *GATK* (*Kraf* tool).

These two modules only concern the GermlineVar (panel/exome), SomaVar, SomaRNA and SomaVar LF worksets.

Kraf uses a single k-mers strategy, without alignment. The recalibrated allelic frequency of indels and delins is only retained if it is higher than the initial allelic frequency of the variant.

The variants are then mainly annotated using the *Variant Effect Predictor*[11] (VEP) tool, and additional information is added from the following databases:

- ClinVar
- OMIM (commercial version) + PanelApp Australia + PanelApp England
- CKB
- dbSNP
- COSMIC (commercial version)
- GnomAD (exome + genome data)
- Dfam
- DGV
- LOVD

[11] https://www.ensembl.org/info/docs/tools/vep/index.html

In addition, the following *in silico* scores are available:

- Splicing scores (*MaxEntScan, GeneSplicer, dbscSNV*)
- Conservation score (*GERP++, PhastCons100way, PhastCons20way, PhyloP100way Vertebrate, PhyloP20way Mammalian, SiPhy*)
- Functional predictions (*REVEL, DANN, FATHMM, FATHMM-MKL, LRT, MetaLR, MetaSVM, MutationAssessor, MutationTaster 2, PROVEAN, SIFT, PLI*)

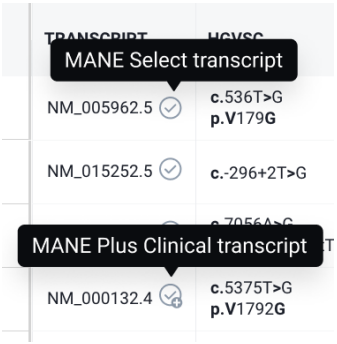
Visualization and interpretation

The Variants tab lets you view, filter and annotate the variants obtained during an analysis. These are displayed in table form, with each row corresponding to a different variant. The number of variants selected after filtering (selected variants) and the total number of variants detected (total number of variants) are displayed in the top right-hand corner of the variants table header using the following icon: 9/22 . A pagination at the bottom of the page allows you to navigate from one page of results to the next. If the analysis uses a restricted panel, the table header will also include the total number of variants in the restricted panel (*total number of variant in the restricted panel*) .

This navigation is facilitated by keyboard shortcuts using the ← and → keys on the keyboard.

By default, several descriptive columns are displayed in the table. Users can export the table as a TSV file by clicking on .

Several transcripts may be available for each variant, and their number is displayed on the left of each row. Variants are annotated on RefSeq transcripts, with the exception of mitochondrial variants which are annotated on *Ensembl*.



MANE Select and MANE Clinical Plus transcripts are identified by an icon, or , respectively, accompanied by a tooltip.

For the same variant, except for mitochondrial variants, the transcripts are sorted according to the order below:

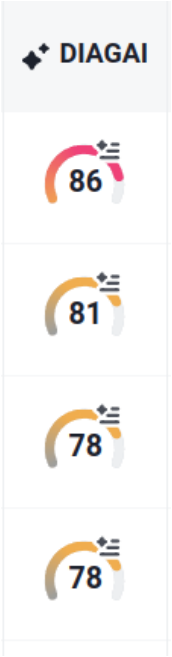
a) the impact ['high' > 'moderate' > 'low' > 'modifier']
b) type of transcripts ['NM', 'XM', 'NR', 'XR'] (for RefSeq transcripts)
c) MANE transcript type [Mane Select > MANE Clinical Plus > non MANE]
d) alphanumeric order

For mitochondrial variants, transcripts are sorted in the following order:

a) impact ['high' > 'moderate' > 'low' > 'modifier']
b) if impact is 'modifier' :
    i) the 'non_coding_transcript_exon_variant' effect
    ii) the [protein coding > Mt_rRNA > Mt_tRNA] biotype
c) alphanumeric order

DiagAI. This column shows the score from 0 to 100 based on the DiagAI suite using machine learning. The DiagAI score is based on four components:

- The UP² algorithm, which predicts the pathogenicity of the variant
- The PhenoGenius matrix, which assesses the match between the gene and the patient's HPOs (if these have been entered during the analysis)
- The concordance between the observed genotype and the mode of transmission of the disease described in the GenCC, OMIM and PanelApp (Australia and England) databases.
- The quality of the variant



The closer the score is to 100, the more the variant is predicted to be pathogenic and of diagnostic interest. The closer it is to 0, the more benign the variant is predicted to be.

Two icons also draw attention to certain variants of interest:

- The icon means that the variant has a higher probability of being the cause of the disease (AI Smart Pick).
- The icon indicates that the variant is present in the short list of variants with a high probability of being the cause of the patient's disease.

HPO terms can be entered before analysis is launched by editing the sample metadata (see **Sample metadata** section). They can also be edited during interpretation using the HPOs modal view in the analysis side panel (see **HPOs modal view** section). In this case, the user can reprocess the DiagAI score and the classification of variants (Short list or SmartPick) according to the new HPOs entered, using the "Re-process with new HPOs" function.

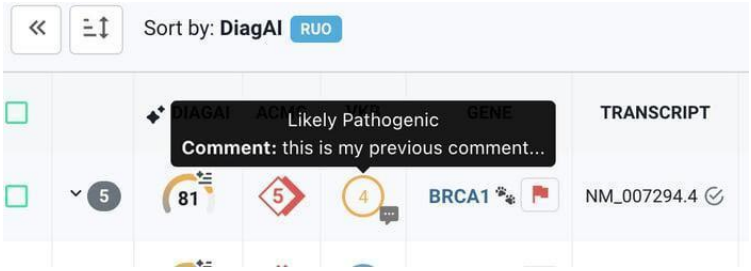
If the user clicks on the DiagAI score icon, the variant page opens directly into the DiagAI tab detailing the score (see DiagAI in the Variant page section below).

ACMG. The pathogenicity classification of each variant is calculated as part of a semi-automated process according to the recommendations issued by the *American College of Medical Genetics and Genomics*. Users can revise the classification in a dedicated tab on the variant page, which is then used to generate the final evaluation for a variant.

ComPerMed. The pathogenicity classification of each variant is calculated as part of a semi-automated process according to the recommendations issued by the Belgian ComPerMed expert committee (Personalised Medicine Commission). Users can access detailed information about the classification in a dedicated tab on the variant page.

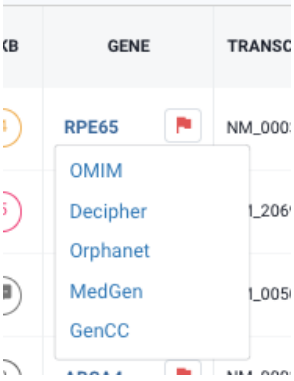
AMP. *(SomaVar analysis version v1.8 or later, SomaVar LF or SomaCGP only)*. This column is displayed by default in analyses that contain CKB annotations and shows the AMP tier of each variant where available (Tier I or II). The AMP tier displayed is based on the best clinical evidence across all types, regardless of the indication selected.

VKB. Evaluations made by the user during previous analyses are stored in the user's knowledge base. This evaluation appears in the VKB column of the table for each new occurrence of the variant. A tooltip quickly displays the class and comment associated with the variant evaluated or previously evaluated.



Gene. Gene naming according to HGNC and associated OMIM flag :

- Red flag: « OMIM gene **with** phenotypes **and** clinical characteristics »,
- Red contour flag: « OMIM gene **with** phenotypes but **no** clinical features »,
- Red dot: « OMIM gene **without** phenotypes or clinical features ».



Genes known to be Imprinting genes according to PanelApp (Australia) are highlighted by the icon 🐾. The associated tooltip indicates whether the gene is maternally expressed, paternally expressed, unknown, or if several imprintings are referenced.

Clicking on the flag next to the gene name takes you directly to the OMIM page. By clicking on

the gene name, a drop-down menu appears, providing links to the OMIM[12], Decipher[13], Orphanet[14] MedGen[15] and GenCC[16] databases. Each link leads directly to the page corresponding to the selected gene.

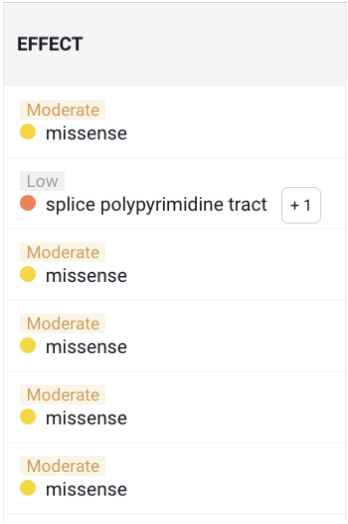
Information relating to OMIM and the databases is only available for germline analyses.

HGVSC. This column provides the HGVS annotation of the variant on the complementary DNA (cDNA) and on the protein.

Effect. This column describes the impact of the variant on the protein and the effect on the transcript.

The impact is calculated according to 4 levels:

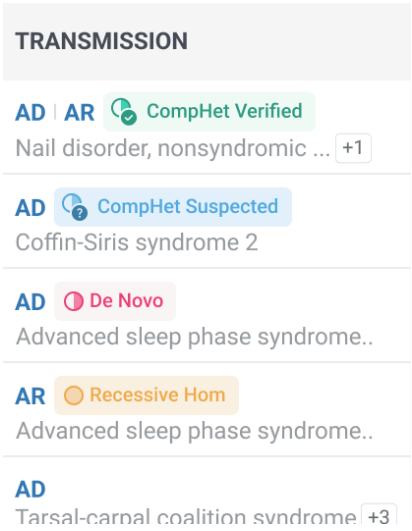
- *High* (high impact, in red): the variant has a high disruptive impact on the protein, probably causing protein truncation, loss of function or triggering its nonsense mediated decay.
- *Moderate* (moderate impact, in yellow): non-disruptive variant likely to modify the protein's efficacy.
- *Low* (low impact, in grey): variant assumed to be harmless or unlikely to modify the protein's behaviour.
- *Modifier* (modifying impact, in blue): generally non-coding variants or variants affecting non-coding regions or genes, for which predictions are difficult or there is no evidence of impact.



For each variant, several effects on the transcript may be available in the Effect column. In this case, only the effect with the highest consequence according to VEP[17] is displayed, followed by a number (e.g. +2) for the other available effects. Details of these are shown in the corresponding tooltip.

[12] https://www.omim.org/
[13] https://www.deciphergenomics.org/genes
[14] https://www.orpha.net/fr/disease/gene
[15] https://www.ncbi.nlm.nih.gov/medgen/
[16] https://search.thegencc.org/
[17] http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

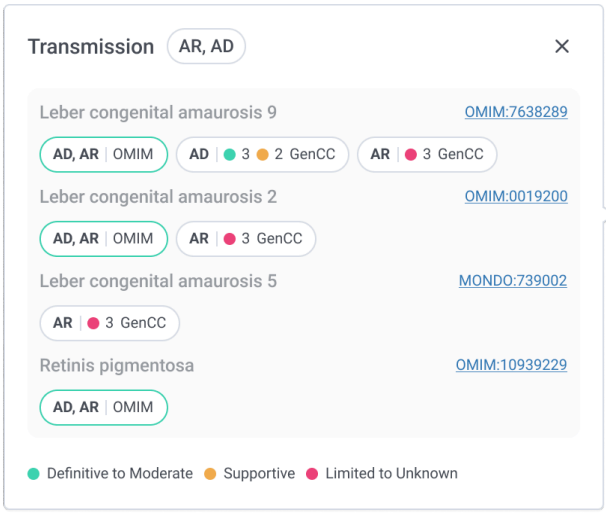
Transmission. This column indicates the mode of transmission of diseases associated with genes reported in the GenCC and OMIM databases. Colored badges are visible to quickly identify transmission patterns based on DiagAI predictions:



- "De Novo"
- "Compound Het - Verified"
- "Recessive Hom
- "Compound Het - Suspected"

De Novo, Compound Het - Verified and Recessive Hom badges are only displayed in family analyses.

A filter can be used to select variants with these badges (see **Filtering variants** - ***Transmission & Phenotypes section***).



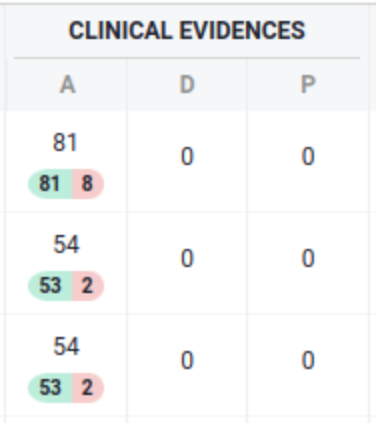
Clicking on the transmission modes links in this column will open a popup window where users can find diseases that are listed with their names and links to databases (OMIM, MONDO -if OMIM is not available- or GenCC).

The diseases are organized in a specific hierarchy: those found in both OMIM and GenCC appear first, followed by OMIM-only entries, then GenCC-only entries.

Each entry includes color-coded tags: OMIM tags have green borders, while GenCC tags use a traffic light system (green/yellow/red) to indicate evidence strength:

- Green: Definitive + Strong + Moderate
- Yellow: Supportive
- Red: Limited, Animal, Unknown

Clinical Evidences. *(SomaVar analysis v1.8 or later, SomaVar LF or SomaCGP only)*. This column is displayed by default in analyses containing CKB annotations. It is divided into three sub-columns, each containing the total number of a given type of evidence: A (Actionability), D (Diagnosis) and P (Prognosis). Column A also displays the number of sensitive (left) and resistant (right) therapeutic assertions as a badge.



The values in these sub-columns are color-coded:

- Resistance, when found (> 0 assertions), is always displayed in red regardless of the indication, and grey otherwise.
- Sensitivity is highlighted in green only when the corresponding indication is selected, and in grey otherwise.

If at least one evidence of this type matches the indication selected in the variants table (Indication panel), a tick appears in the tooltip in front of the indication label. If no match is found, the label will not appear.

Phenotypes. This column, available for Germline projects, allows the association of a gene with HPO terms using the PhenoGenius[18] matrix. If HPO terms are associated with the sample, the highest percentage match between gene and phenotype is displayed, together with the ratio of terms with which the gene is associated. If no genes match the HPOs entered, the column returns the message "no match"; if the sample has no HPO terms, the column displays the message "see HPO". In both cases, known terms for the gene are nevertheless proposed in the associated pop-up window.

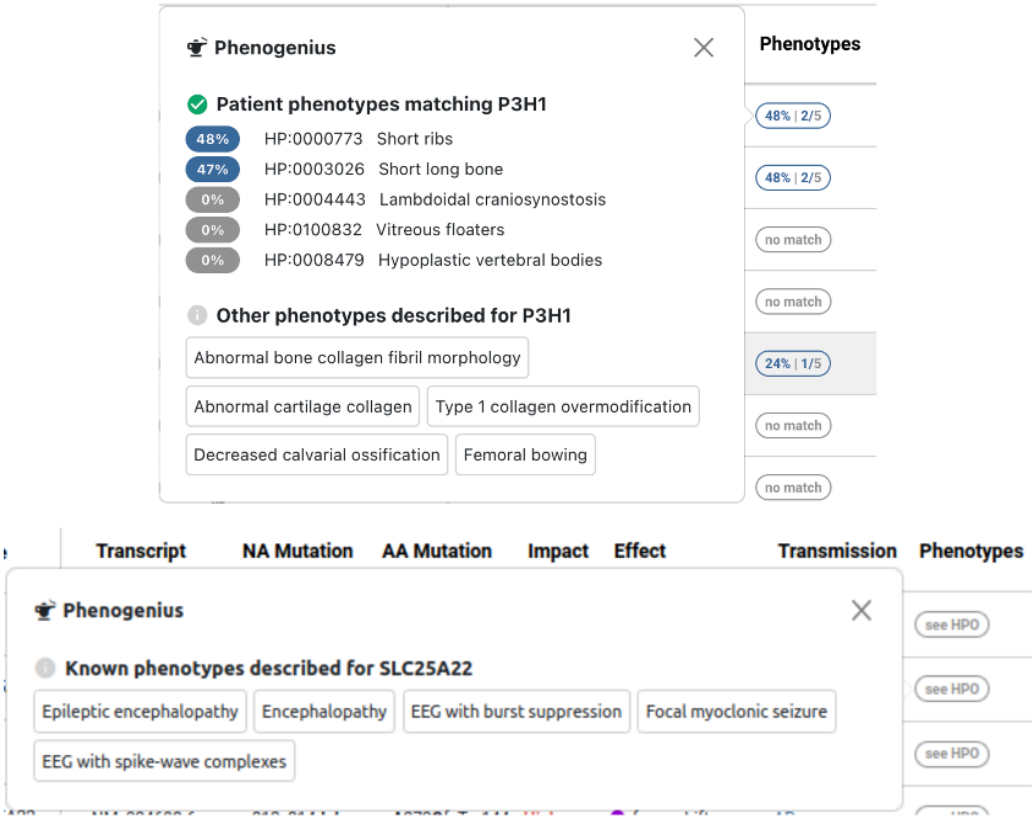
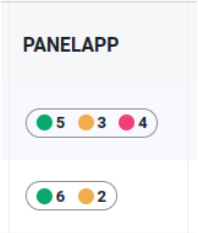


Figure 11 : Viewing information on Phenogenius

[18] Kevin Yauy *et al.*, « Learning phenotypic patterns in genetic diseases by symptom interaction modeling », *medRxiv*, p. 2022.07.29.22278181, janv. 2022, doi: 10.1101/2022.07.29.22278181.

PanelApp. This column integrates annotations from the *PanelApp Australia*[19] and *PanelApp England*[20] panel database. It indicates the number of panels associated with the gene for which a variant has been detected, according to the classification level described below.

If a gene is not found in a panel, the column will display "N/A".



The associated pop-up window displays details of the gene/phenotype association, as well as the mode of transmission associated with the phenotype, the mode of transmission associated with the phenotype and the database from which the panel is derived, PanelApp Australia (**AUS**) or PanelApp England (**ENG**) (**Figure 12**).

Each gene will be assigned one or more panels according to a three-level classification system based on the evidence for the association of gene and phenotype: High evidence in green, Moderate evidence in yellow and Low evidence in red.

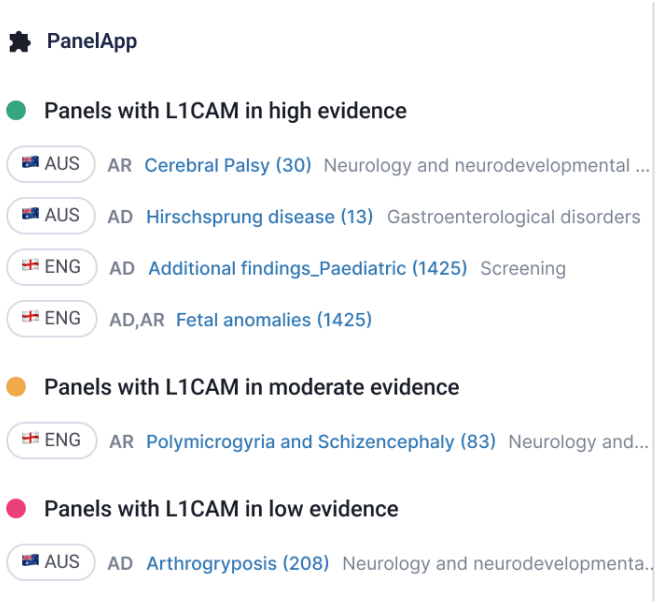
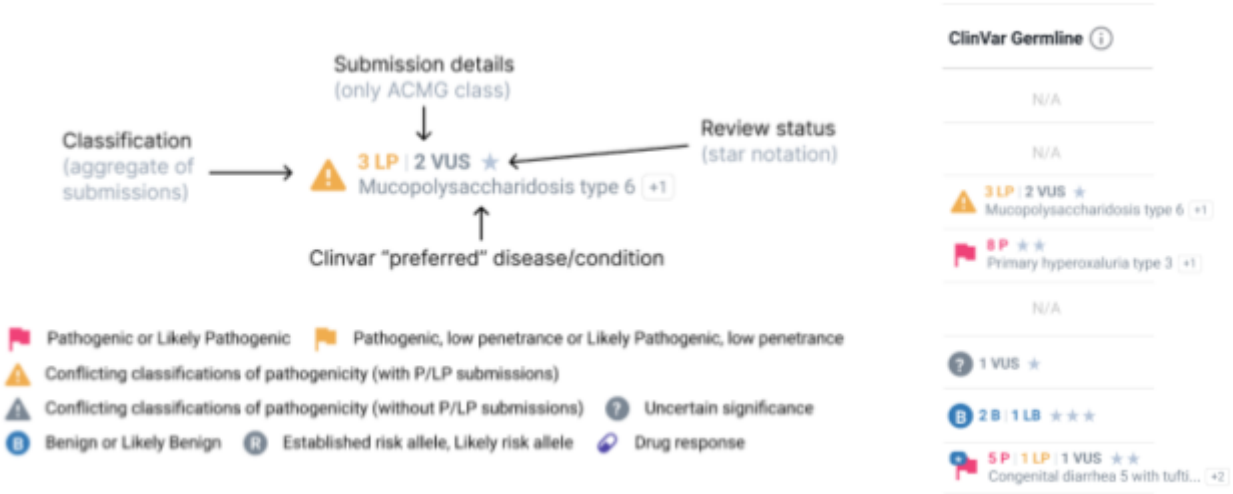


Figure 12: Visualisation of PanelApp information.

[19] https://panelapp-aus.org
[20] https://panelapp.genomicsengland.co.uk/

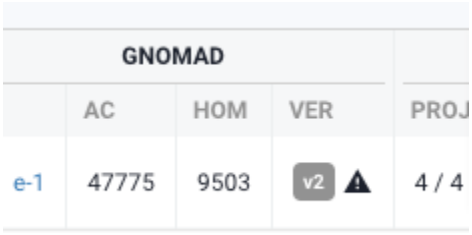
ClinVar Germline. This column displays information extracted from ClinVar during the analysis. It provides a range of information, including pathogenicity class, the details and number of submissions for each ACMG class, revision status and ClinVar preferred condition/disease. By clicking on the cell in the variants table, the user will be redirected to the corresponding ClinVar page.



For more information, see https://www.ncbi.nlm.nih.gov/clinvar/docs/review_status/

GnomAD. This column displays 6 elements from the GnomAD database for a specific variant in the population: the allelic frequency (AF), the number of alleles (AC), the number of homozygotes (HOM), the number of hemizygotes (HEM), the major allele frequency (MAF) and for GRCh37 projects, the version of GnomAD for which this variant is found (VER).

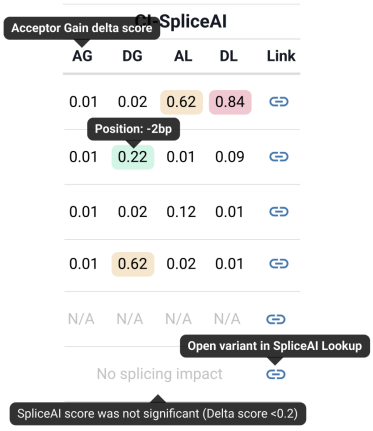
By clicking on the value displayed in the AF column, the user is redirected to the GnomAD database in its v4 version or, failing that, in its v2 version. If a variant is not listed in the database, the value displayed in this column will be N/A.



An alert is displayed in the VER column if this variant is present in GnomAD, but does not pass all the quality filters for at least one of the two datasets (exome or genome). In this case, the values displayed adapt to the values available for this variant, and the filters that triggered the alert are displayed, along with a short message mentioning the offending dataset(s).

Score UP². The UP² (Universal Pathogenicity Prediction) functional prediction score ranging from -1 to 1 classifies the pathogenicity of variants using a Machine Learning approach according to three categories : **P** (Pathogenic : score > 0.1), **VUS** (Variant of Unknown Significance : score ≥ -0.1 et ≤ 0.1) or **B** (Benign : score < -0.1). The UP² score is displayed by default in the variants table, but the user can disable the UP² column from the column management panel by clicking on the  icon.

CI-SpliceAI. This column incorporates prediction scores calculated using a Machine Learning approach for identifying variants with an impact on alternative splicing in the manner of *CI-SpliceAI*[21].



Scores are calculated for a subset of variants meeting the following criteria: splice variants according to VEP (splice_*) with a GnomAD frequency of less than 10^{-3} and meeting the following quality criteria (QA > 20, AO > 10, DP > 20).

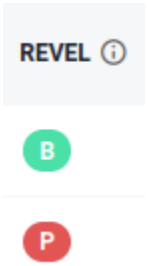
If calculated, the column provides 4 delta scores for the variant concerned, for acceptor gain (AG), acceptor loss (AL), donor gain (DG) and donor loss (DL).

The delta score of a variant, ranging from 0 to 1, corresponds to the probability that the variant modifies splicing. A color code is applied to the delta score thresholds: green (>0.2 ; high recall), yellow (>0.5; recommanded) and red (> 0.8; high precision). If the delta score is below 0.2, we consider the score to be insignificant. The column then specifies 'No splicing impact'.

For each variant, a link is provided to the corresponding *SpliceAI* lookup[22] interface.

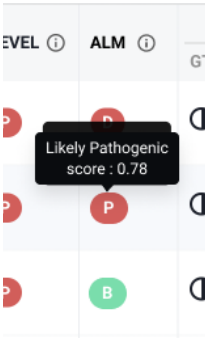
[21] https://github.com/YStrauch/CI-SpliceAI__Annotation
[22] https://spliceailookup.broadinstitute.org/

REVEL. The REVEL prediction score is based on an overall method for predicting the pathogenicity of missense variants using a combination of scores from 13 individual tools: MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP++, SiPhy, phyloP, and phastCons.



The score can vary from 0 to 1 for a missense variant, with higher scores reflecting a greater probability of the variant causing the disease. The thresholds applied are as follows: REVEL score > 0.5 for **P** (Pathogenic) and REVEL score < 0.5 for: **B** (Benign). The score can be displayed in the variants table by the user by ticking the REVEL box in the "Columns" insert, in the "Functional Prediction" section.

ALM. The AlphaMissense[23] prediction score is available in the ALM column from the Columns insert, under Functional Prediction. AlphaMissense is a prediction model that assesses the pathogenicity of missense variants using deep learning. The model is built on the structural predictions of AlphaFold to improve its accuracy.



AlphaMissense integrates evolutionary conservation and structural context to assess the potential impact of missense variants on protein function.

The score ranges from 0 to 1, with higher values suggesting a greater probability that a variant is deleterious. The thresholds applied are: ALM score > 0.564 for **P** (Likely Pathogenic) and ALM score < 0.34 for **B** (Likely Benign). In other cases (0.34 < ALM score < 0.564), it will be labeled **VUS** (Ambiguous).

Sample Freq. This column shows for each variant its frequency in the project (sub-column PROJECT) and in all projects of the entity with similar metadata (sub-column "ALL").

SAMPLE FREQ.	
PROJECT	ALL
2 / 13	7.41%
2 / 13	7.41%

The sample frequency of the project (PROJECT) corresponds to the number of samples in which the variant is detected in relation to the total number of samples analysed in the project.

The sample freq. ALL corresponds to the percentage of analysed samples in which the variant is found on the entity, within similar metadata projects. By clicking on the value displayed in the "ALL" column, the user can access the names of the projects and the number of samples in which the variant is present, as well as

[23] https://alphamissense.hegelab.org/

details of all similar metadata projects (“Show all projects” option).

By default, the ALL sub-column is hidden in the variant table, but it can be reactivated from the column management panel by clicking on the x icon and activating the “Sample Frequency Entity (Entity)” button.

Exon. This column indicates the total number of exons in the gene according to the transcript concerned, as well as the number affected by the mutation, in the form of a fraction. A tooltip indicates the distance between the variant and the nearest splicing site. If the variant is upstream of the splice site, the value will be positive and if it is downstream, it will be negative.

GT (Genotype). This column shows the genotype predicted by the variant detection programs. The various possible pictograms are described in the **Filtering variants** section.

In a trio analysis (GermlineFamily), if the origin of a heterozygous variant can be determined, it will be indicated by the following color code:

- pink → Maternal heterozygous origin
- **blue** → Paternal heterozygous origin
- **red** → *De novo* heterozygous variant
- **orange** → Mendelian violation

No color is applied if transmission is undetermined (e.g. if both parents are heterozygous), or for homozygous variants.

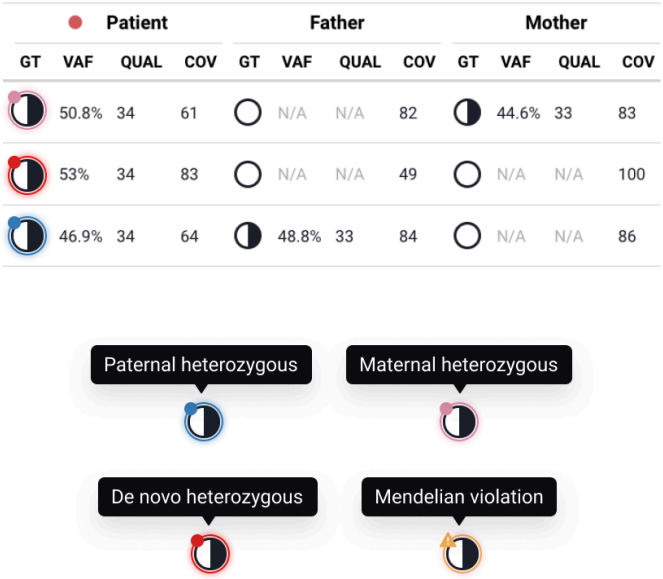
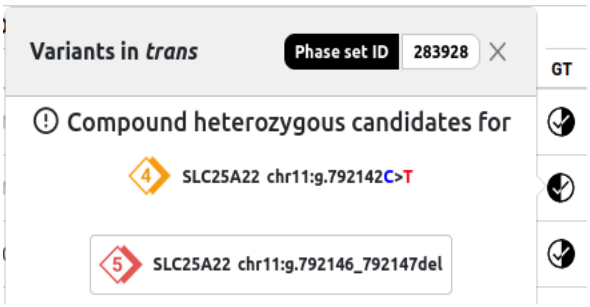


Figure 13: Color code applied to heterozygous genotypes based on GermlineFamily transmission

Variant phasing is available for GermVar Tertiary and GermVar Tertiary Family analyses using VCF files from Oxford Nanopore Technologies sequencers. When variants are phased, the genotype icons appear as follows :  . The check mark indicates that the variant is phased. Phased variants with the black part on the same side and belonging to the same "Phase ID" are in phase, i.e. on the same strand.

You can click on the genotype to obtain a list of cis or trans variants, if present. Only the 5 variants ranked highest by DiagAI are presented, as shown opposite.

The complete list of variants is available by clicking on "Browse all variants".



VAF (Variant Allele Frequency). This column indicates the variant allele frequency, i.e. the proportion of reads supporting the alternative allele (AO) to the total number of reads at that position (COV), expressed as a percentage.

Analyses using worksets SomaVar version ≥ 2.5, SomaCGP version ≥ 0.7 and SomaVar LF version ≥ 2.1 display the following additional information:

- a tooltip displays the VAF, AO and COV values calculated by the 2 *variant callers* Freebayes and Mutect2 detecting the variant.
- a warning icon ⚠ alerts the user to any disagreement between the 2 variant callers for variants detected with a low VAF. Thus, for variants displaying a VAF of less than 5% in the VAF column, the warning is displayed in the event of a VAF discrepancy > 10% between the VAFs calculated by Freebayes and Mutect2.

AO (Alternate Observation). This column indicates the number of reads carrying the variant.

CAL (Caller). This column indicates the name of the *variant caller* (s) having detected the variant.

Cov (Coverage). This column indicates the number of total reads at the position of the variant.

SOR. Strand Odds Ratio (SOR) is a strand-bias metric derived from the Symmetric Odds Ratio test that measures imbalance of variant-supporting reads between forward and reverse strands. Higher SOR suggests a higher likelihood of an artefact, while lower SOR suggests more balanced strand support and is more consistent with a true variant.

Sorting variants

There are several options for sorting variants on the Variants tab. These are available from a drop-down menu in the top right-hand corner of the table. For each sorting option, it is possible to sort in ascending or descending order by clicking on the sort icon . By default, all the options are sorted in descending order, with the exception of the 'Locus' option, which is sorted in ascending order.

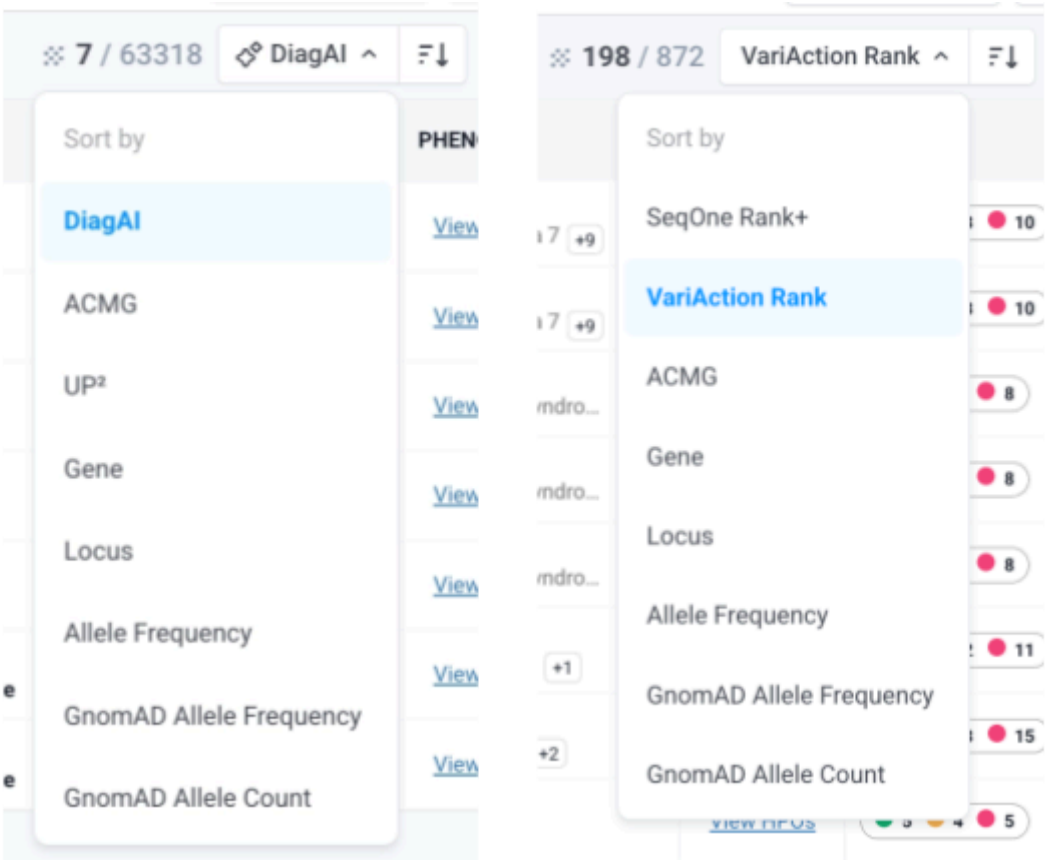


Figure 14: Variant sorting option selection window

DiagAI Score. This sorting option is applied by default to all germline analyses and allows variants to be sorted according to their DiagAI score (replaces SeqOne Rank+ in analyses carried out after October 2, 2024).

SeqOne Rank+. The SeqOne Rank+ prioritization algorithm, based on an older artificial intelligence model than DiagAI Score, remains available and is applied by default to all analyses carried out before 2 October 2024.
It is based on 4 main components:

- A new implementation of the ACMG classification, based on machine learning trained on ClinVar (UP² score)
- The patient's phenotype (if HPO terms are associated with the sample)
- The concordance between the genotype observed and the mode of transmission of the pathology described in the GenCC, OMIM and PanelApp (Australia) databases.
- Variant quality

VariAction Rank. The VariAction Rank algorithm for prioritizing somatic variants is proposed in SomaVar analyses from version v1.8. This system uses the JaxCKB annotation to calculate variant prioritization scores based on their clinical actionability. These prioritization scores take into account the impact of the variant, its pathogenicity using the ComPerMed score, the AMP Tier, the specificity of the variant with the molecular profile, as well as the correspondence with the indication, if the disease was entered before the analysis was launched.

ACMG. Variants can be sorted according to their ACMG classification by selecting this option. By default, variants are sorted in descending order from Pathogenic class 5 to Benign class 1.

Locus. Variants can be sorted according to their genomic position by selecting this option.

Allele frequency. Variants can be sorted by allele frequency by selecting this option. By default, variants will be sorted in descending order. It is possible to sort in ascending order by clicking on the sort icon  .

GnomAD Allele Frequency, GnomAD Allele Count. Variants can be sorted by their GnomAD Allele Frequency (AF) or Allele Count (AC) in ascending or descending order (if the GnomAD data is available).

UP². Variants can be sorted by the pathogenicity UP² score, a component of DiagAI Score.

Gene. Variants can be sorted in ascending or descending alphabetical order according to the gene symbol.
For variants affecting overlapping genes, the gene symbol used for sorting is the one associated with the variant with the highest impact, based on the transcript prioritization rules described above. When expanding the transcript list, the user may see the second overlapping gene, which may appear unsorted. Sorting may also appear not to be applied when multiple filters are applied to the variant table.

Configuring the columns of the variants table

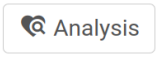
The columns displayed in the variant table can be configured from the *Organise your columns* menu, accessible via the button  at the top right of the variant table. The display of columns can be modified by the user by clicking on the icons  to activate the desired column(s) in the left section. The right section displays the active columns in the analysis and allows users to create and save column profiles from a drop-down menu by clicking on the "*Select profile*" icon.

Active columns can be reordered by clicking and dragging them to the desired location. By default, columns are organized according to "Germline" and "Somatic" profiles.

When users hover over a column profile name in the preset dropdown menu, a tooltip shows which workset created that profile. This warns users if their profile might have columns incompatible with the current workset.

To create a new column profile, simply enter a name in the "*Create a new preset*" field and click on "*Save*".

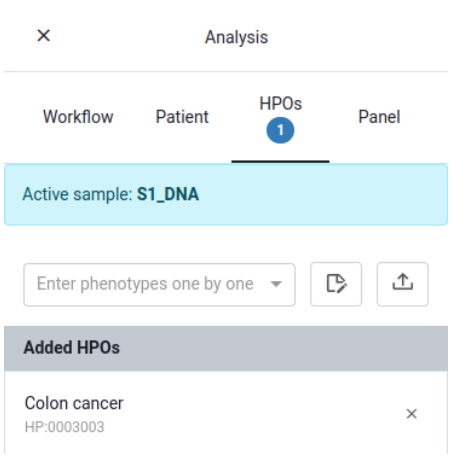
Edit sample metadata from the analysis panel

The user can edit sample metadata directly from the analysis and enter the patient's gender or HPOs (only for germline analyses) from the Analysis drawer, accessible by clicking on the button 

at the top right of the page (see **Side panel for analysis management** section).

From the Patient view, the user accesses gender editing (see the **Patient modal view** section).

In Germline project analyses, from the HPOs view, the user has access to several HPOs term editing functions:



- View terms already entered before launching the analysis
- Manual addition of HPO codes or phenotypes via the "Enter phenotypes one by one" search bar
- Import a text file containing HPO codes
- Generation of HPO terms based on the patient's clinical description, by calling up DiagAI via the icon  . This feature prompts the user to describe the patient's clinical profile in English (description limited to 4000 characters). Then, after clicking on "Extract HPOs with DiagAI", a list of HPOs highlighted in color is generated. Finally, the user can associate all the HPO terms proposed with the patient by clicking on "Add all", or select them individually by hovering the mouse over the terms of choice, then clicking on "Add".

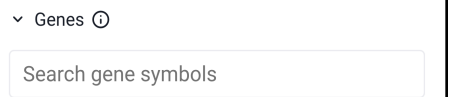
Adding and editing HPOs from this view gives the user the option of reprocessing the DiagAI score and variant classification (Short list or SmartPick) according to the new HPOs entered via the "Re-process analysis" functionality.

Filtering variants

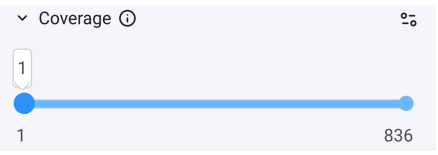
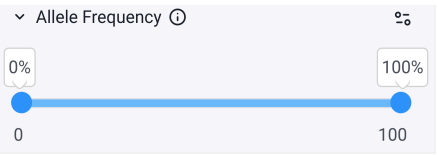
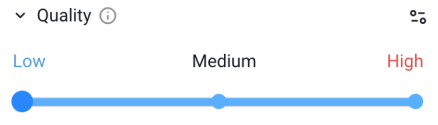
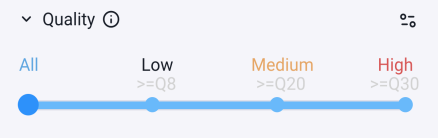
You can refine searches in the *Variants* tab by using the filter categories detailed below. The filter panel can be accessed via the 

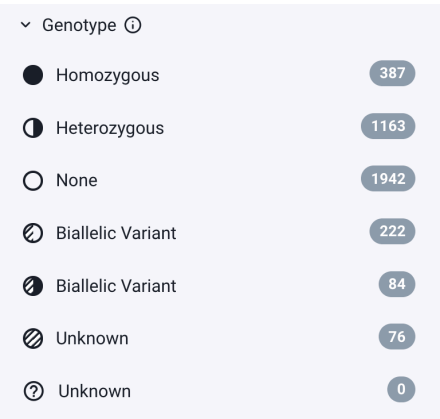
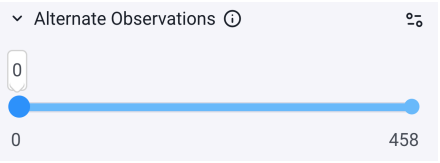
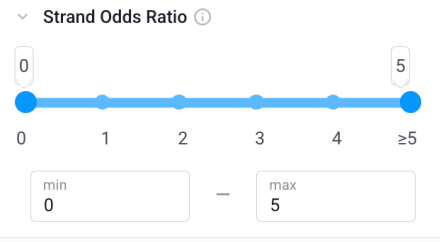
icon located at the top left of the variants table or by pressing the F key on the keyboard. Filters are available under the "Basic" tab. The "Active Filters" section, located at the top of the filters panel, displays the count of filters that are currently in use.

Genes & Variants Filters

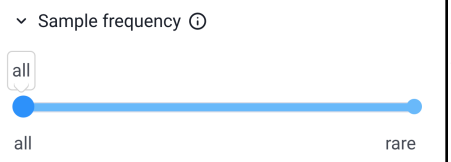
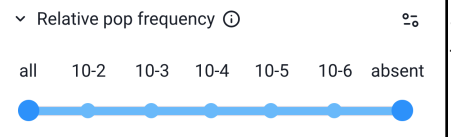
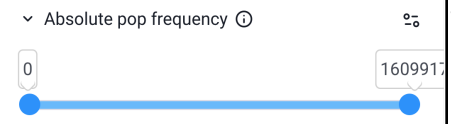
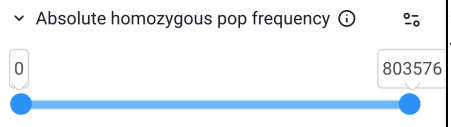
Filter name and overview	Description
	Restricts the list of variants to one or more genes.
	Restricts the list of variants to one or more variants according to the HGVSc or HGVSp nomenclature.

Samples & Quality Section

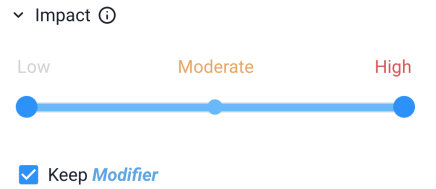
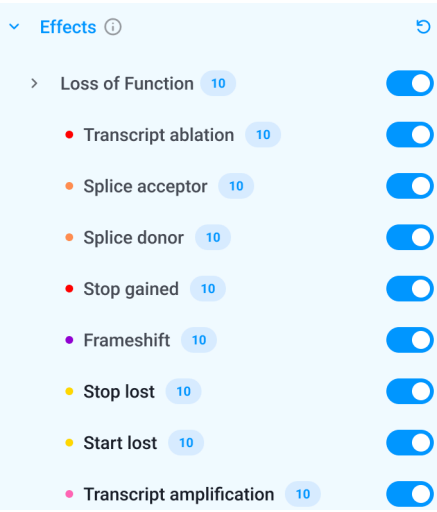
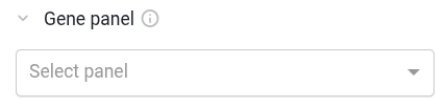
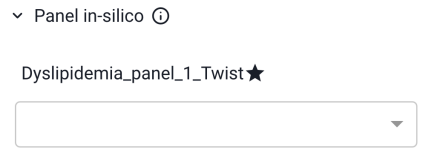
Filter name and overview	Description
	Defines a minimum threshold for the depth of coverage (COV). The maximum value displayed by this filter corresponds to the highest coverage value identified among the variants.
	Filters variants according to the frequency of the mutant allele in the analysed sample.
	Average quality (phred score) of each base carrying the mutated allele, indicating confidence in variant detection: - *Medium* (≥ QUAL=20): 1 chance in 100 of incorrect calling, - *High* (≥ QUAL=30): 1 chance in 1000 of incorrect calling. This filter applies to all samples.
	For GermVar Tertiary and GermVar Tertiary Family analyses on data from Oxford Nanopore Technologies, quality based on the Phred score indicates confidence in variant detection. ≥ QUAL=8 (Low): 1 chance in 8 of incorrect calling, ≥ QUAL=20 (Medium): 1 chance in 100 of incorrect calling, ≥ QUAL=30 (High): 1 chance in 1000 of incorrect calling.

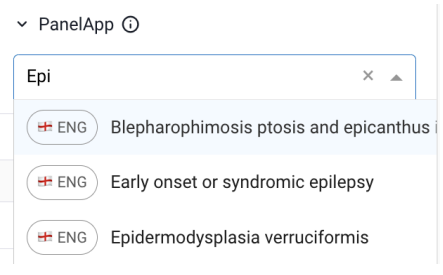
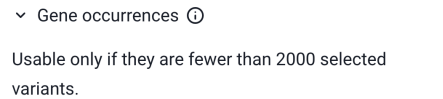
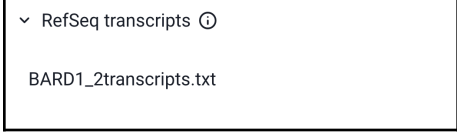
	This filter applies to all samples.
	Filters variants according to the genotype observed.
	Defines a number of sequences carrying the mutation (alternate observation or AO). The maximum value displayed by this filter corresponds to the highest AO value identified among the variants.
	Filters variants according to the strand odds ratio (SOR) score calculated at the variant site.

Frequencies Section

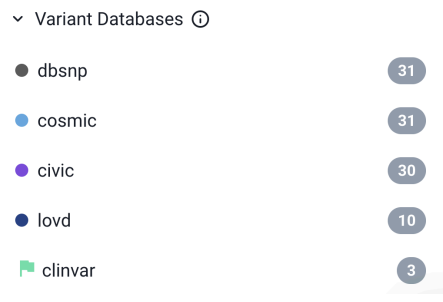
Filter name and overview	Description
	Eliminates variants based on their recurrence in the project under study, from the most frequent (all) to the least frequent (rare).
	Selects variants according to the frequency of their mutant allele in the combined population of the ExAc and gnomAD projects.
	Selects variants based on the number of alleles in the combined ExAc and GnomAD population.
	Selects variants on the basis of their homozygosity frequencies in the combined ExAc and GnomAD population.

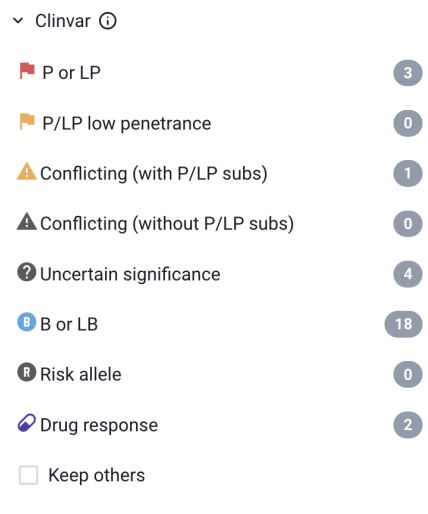
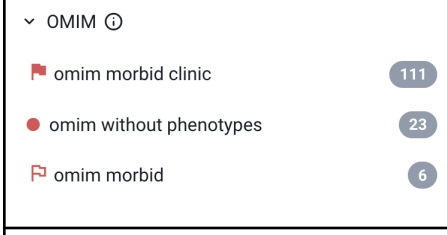
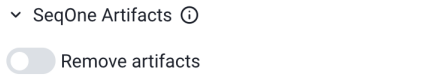
Genetics Section

Filter name and overview	Description
	The impact of the mutation on the protein. *Modifier*: affects the non-coding genes
	SO (Sequence Ontology) terms that describe the most disruptive consequence at gene level. This filter allows you to select the impact of the mutation on the gene, and the options are organized in logical categories to improve navigation: *Loss of Function, Coding, Splicing region, Silent variant, Intronic, UTR, ncRNA, Upstream/Downstream, Regulatory, Others.*
	Select and apply one or more gene panels. This filter displays all the gene panels available for the entity. This filter is not available if a restricted panel has been applied to the analysis from the *Panel* tab of the analysis management pane (see **Panel modal view** section). A message informs the user that the 'Gene panel' filter cannot be used with a restricted panel.
	Selects and applies a panel in-silico. This filter displays all the available BED files for the group. The name of the file used during the analysis is highlighted with a star. Only the variants whose starting position (start) is located within one of the intervals of the selected bed file are retained by this filter.

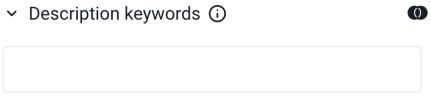
	Allows you to filter variants according to a list of in-silico panels in the PanelApp Australia and PanelApp England catalogs. A badge identifies the database from which the in-silico panel originates: AUS (Australia) or ENG (England). Among both catalogs, superpanels are not available for selection, as well as panels that do not include at least one gene. For example, a panel containing only STRs will not be selectable.
	Defines a threshold for the number of occurrences of a given gene in the variants table.
	Retains variants located on the sex chromosomes and/or the mitochondrial chromosome (M)
	Selects a list of RefSeq transcripts for filtering the variants.

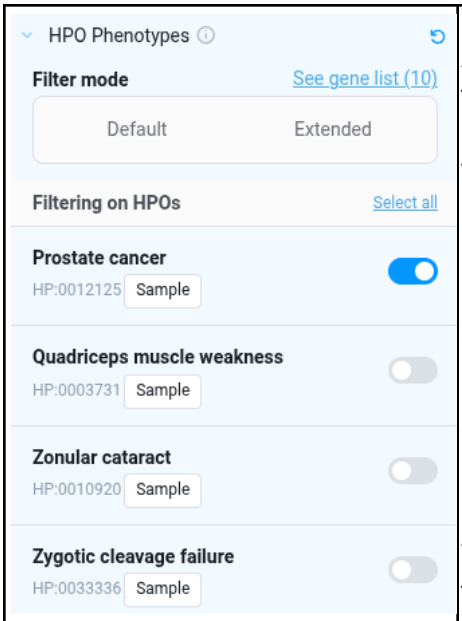
Databases Section

Filter name and overview	Description
	Selects variants annotated by the following databases: dbsnp, cosmic, lovd, clinvar and civic.

	Allows variants to be filtered according to their pathogenicity status on ClinVar. The 'Keep others' box is used to keep the variants without ClinVar annotation.
	Allows you to filter on OMIM morbid clinic, OMIM without phenotypes and OMIM morbid variants. This filter has an in/out button which can be used to select or delete a status.
	Button to remove recurrent variants present in 5% of a cohort of 500 exomes considered as artifacts according to the SeqOne algorithm.

Transmission & Phenotypes Section

Filter name and overview	Description
	Uses keywords to characterize the phenotype(s) associated with the variants by querying the OMIM, Medgen and Litvar databases. This function does not take HPOs into account, as a specific filter is dedicated to them (see below). Two search modes are available: union and intersection.



Allows variants to be filtered on the basis of previously entered HPO terms, either before sample analysis (see **Sample metadata**), or after analysis from the HPO tab of the analysis pane (see **HPO modal view**).

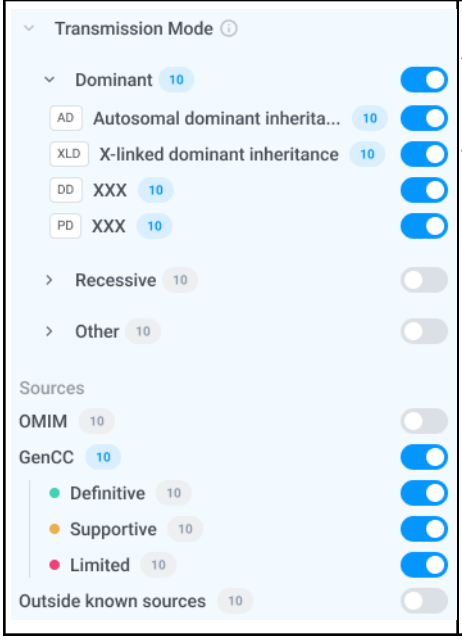
Matches are identified using our PhenoGenius approach, which relies on an HPO-gene matrix to select genes.

2 options are available in Filter mode :

- Default filtering that balanced accuracy and number of genes returned.
- Extended filtering maximizing accuracy with more genes

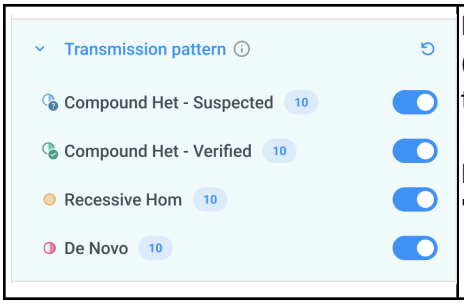
The "See gene list" option provides a list of genes associated with the HPOs selected in this filter. This feature is only available if HPO terms are checked.

It opens a pop-up window showing each gene's ID, symbol, Phenogenius rank, and Phenogenius score. You can download the gene list using the Download icon or copy it using the copy/paste button.

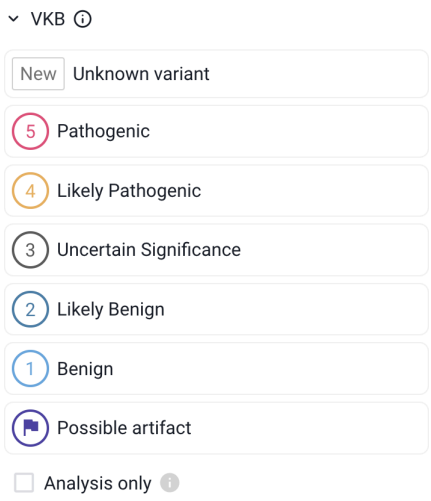
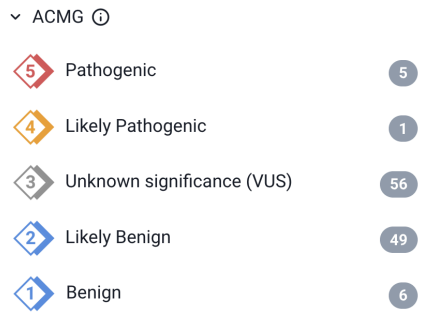
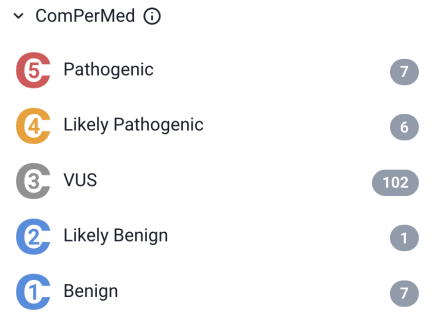
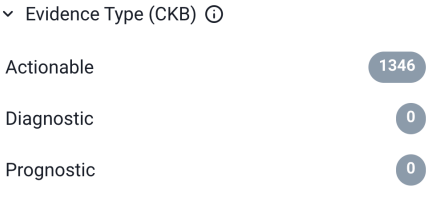


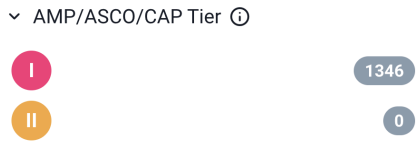
Allows users to filter variants using inheritance links from diseases related to human medical genetics.

Variants can be filtered according to transmission modes (Dominant, Recessive, Other) and transmission sources (OMIM, GenCC, or other databases).

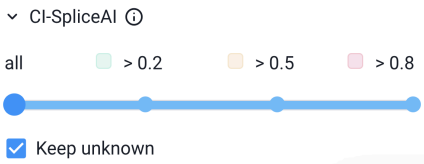
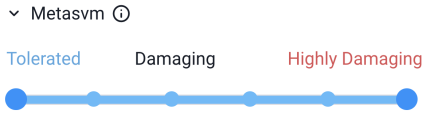
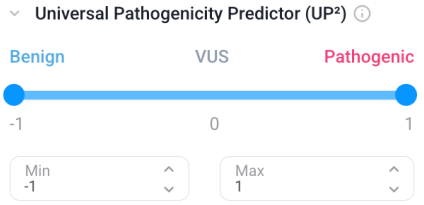
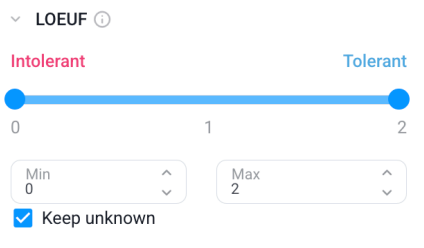
	Enables selection of variants according to heterozygous composite (suspected or verified), homozygous recessive or de novo transmission modes based on DiagAI. Please note that the "Compound Het - Verified", "Recessive Hom" and "De novo" filters are only available in family analyses.

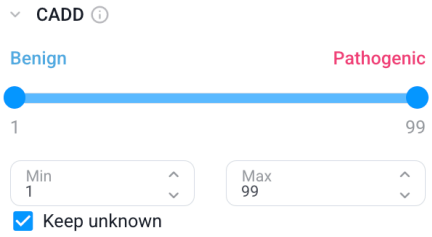
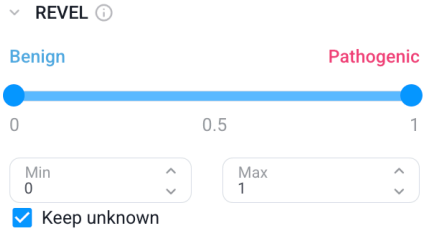
Classification Section

Filter name and overview	Description
	Allows you to select unknown variants not evaluated (New - Unknown variant) or validated in VKB on all the entity's analyses. If the 'Analysis only' box is checked, variants will be searched only on the analysis currently being annotated.
	Pathogenicity classification for germline variants, calculated on a scale of 1 (benign) to 5 (pathogenic).
	Selects somatic variants according to their ComPerMed pathogenicity class on a scale of 1 (benign) to 5 (pathogenic).
	Selects variants according to 3 types of evidence (CKB): Actionable, Diagnostic, Prognostic.

	Select variants according to their assigned AMP level (levels I, II and N/A) in somatic context analysis only.

Scores Section

Filter name and overview	Description
	Select delta scores. The "Keep unknown" box keeps variants with a value of "N/A", but excludes variants with no splicing impact ("No splicing impact" variant).
	Degree of conservation of the site, calculated from multiple conservation scores.
	Takes into account the score of 10 variant prediction applications and the maximum frequency observed in the 1000genomes population.
	Select variants according to the Universal Pathogenicity Predictor (UP²).
	Select variants according to the LOEUF score.

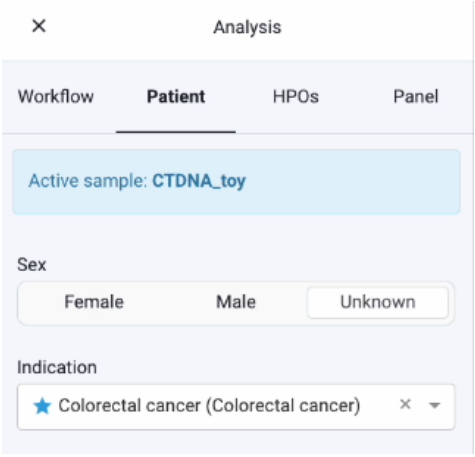
	Select variants according to the CADD Phred score.
	Select variants according to the REVEL score.

Select an indication

In SomaVar, SomaVar LF and SomaCGP, it is possible to select a predefined pathology indication from the JaxCKB database from a drop-down menu available in the *Patient* tab of the analysis management sidebar (see **Patient Modal View** section).

The ontology system is based on *Disease Ontology*[24] and includes terms specific to JaxCKB. Sensitivity information, in sub-column 'A' of the Clinical Evidences column (see section **Variant Page**), is highlighted in green only when the indication for that variant matches the selected indication. If resistance assertions are found for a variant, the corresponding cell in column 'A' will be highlighted in red, regardless of the indication. Otherwise, it will be grayed out.

If the disease has been selected during sample loading and before analysis (see **Sample metadata**), it becomes the default indication defined for the sample and in the *Patient* tab of the analysis management side panel. It is marked with a star icon and appears first in the list. Changing the indication in the analysis management side panel does not update the pre-selected disease for the sample (Disease metadata).



[24] https://disease-ontology.org/

Using a predefined filter profile

Users can select a predefined filter profile from a drop-down menu by clicking on the "*Select profile*" icon. The SeqOne platform offers several predefined filter profiles by default: *GermVar, My Selection, Research, DiagAI*.

Editing and saving filter profiles

Users can save the current position of the cursors for each filter as a filter profile by giving the profile a name and clicking on "*Save as new*". The filter profile will then be available via the same drop-down menu in the *Private* category.

When saving a filter profile, a creation form opens and the user can select the button with the entity's name to make it accessible to other users of the entity. If the user selects the "Private" button, the profile will only be accessible to the user who created it.

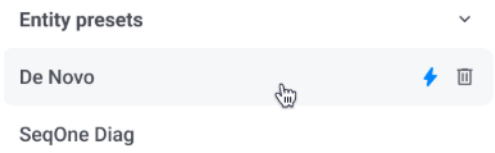
Only the user who initially created the filter profile will be able to edit it by clicking on "*Update*"; if the profile is modified by another user in the entity, that user will be asked to save it under a different name. If the user does not have permission to modify a public filter profile, it will be asked to contact the administrator of its entity. SeqOne filter profiles cannot be modified and saved under another filter profile.

One or more filters can be reset. Simply click on the icon to the right of the filter or in the top-right corner of the "*Filters*" box.

Favorite filter profiles

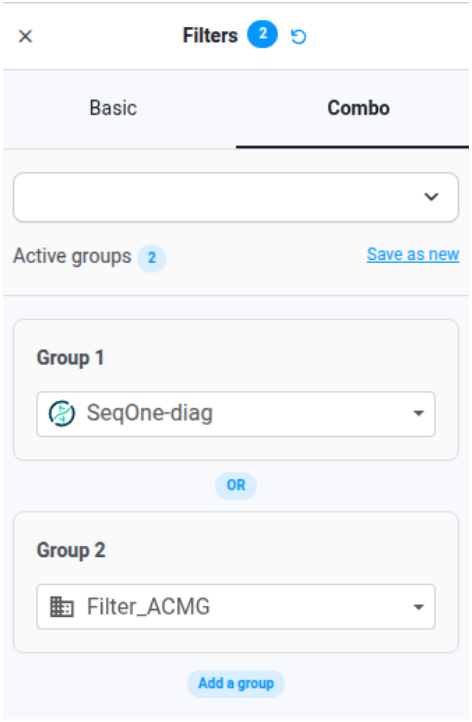
Users can set up filter profiles as favorites, by clicking on the icon to the right of the profile, so that they appear as a shortcut directly in the quick access bar above the variants table.

To remove a filter profile from the favorites, simply click on the same icon . By default, the *My Selection* profile is defined as a favorite.



Combining filter profiles

Filter profiles can be combined under the **OR** condition by clicking on the Combo tab in the filter panel, then on "*Create a combo*". By default, the creation form shows 2 groups of filter profiles, but users can add more by clicking on "*Add a group*".



Once the profiles have been selected, users are invited to save their combo profile by clicking on "*Save as new*". As with filter profiles, the user can make his combo accessible to all users of the entity in the creation form. Only the user who initially created the combo profile will be able to edit it by clicking on "*Update*".

If the profile is modified by another user in the entity, that user will be asked to save it under a different name. If the user does not have permission to modify a public combo profile, it will be asked to contact the administrator of its entity.

SeqOne filter profiles

My Selection. This filter profile enables variants assessed as 3 (*VUS*), 4 (*Likely Pathogenic*) and 5 (*Pathogenic*) to be displayed directly in VKB in the analysis being interpreted.

DiagAI. This filter profile, available in germline analyses (GermlineVar, GermlineFamily, GermVar Tertiary, GermVar Tertiary Family), displays a reduced list of variants with a high probability of being the cause of the patient's disease. This profile exploits the DiagAI prioritization algorithm and the occurrence of the variant in the project (Sample Frequency filter).
It can be applied when HPO terms are or are not associated with the sample, but its sensitivity is increased in the presence of HPO terms. In the case of a family analysis, the filter profile automatically adapts to the size of the family analyzed, as it takes into account that the variant occurrence in the project may be greater. Users can also modify the Sample Frequency filter in the project at any time.
Enter HPOs to display SmartPick variants with the highest probability of causing disease.

Family profiles. In GermlineFamily and GermVar Tertiary Family analyses, *De Novo (new), Recessive Hom (new),* and *Compound Het (new)* filter profiles will only be available for trio analyzes including an affected index case, the unaffected father and the unaffected mother. Otherwise, these profiles will be disabled.

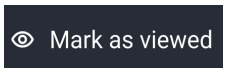
It is not recommended to modify the *De Novo (new), Recessive Hom (new),* and *Compound Het (new)* filter profiles when building and saving a personal profile. It is recommended to start from scratch when creating a variant of these profiles.

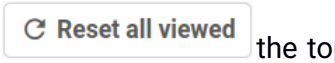
Carrier screening. This filter profile is only available in Germline family for duo analyses (parents without index sample) and identifies recessive genes with a potentially pathogenic heterozygous variant in both unaffected parents for preconception testing.

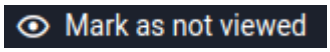
Somatic FFPE. This filter profile displays a reduced list of somatic variants in somatic analyses (SomaVar, SomaCGP) performed on FFPE DNA.

Somatic LF. This filter profile displays a reduced list of somatic variants in SomaVar LF analyses.

Mark as viewed feature

"Mark as viewed" is a feature that allows the user to selectively hide certain variants. To do this, simply select the desired variant(s) and click on the icon 

displayed at the bottom of the variants table. The selection of variants then appears grayed out. To remove the selection, the user can either click on the icon 

the top of the variants table to reset the view, or select the variant(s) and click on 

.

The user can hide the popup by clicking on the cross. This will deselect all variants, but they will remain grayed out.

Selecting variants

Variants can be added to the "Selection" panel alongside any evaluated variants (see section below). To do this, select variants by checking the boxes on the left side of the variant table, then click the "Add to selection" icon at the bottom of the table.

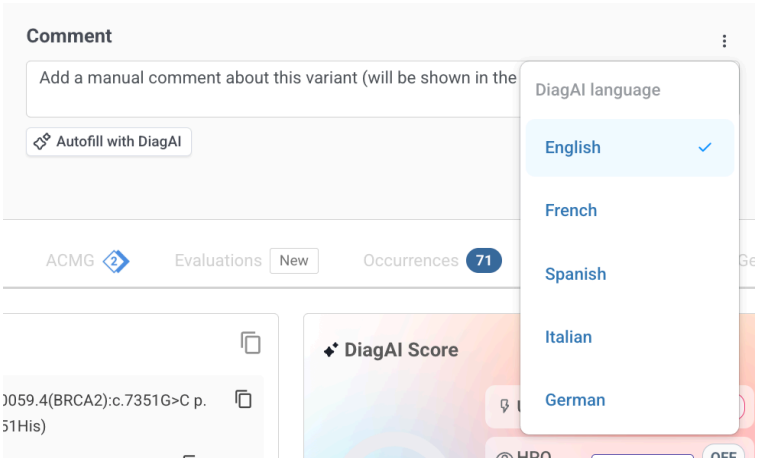
Evaluating variants

Variants can be quickly evaluated from the variant table by clicking in the VKB column and specifying their pathogenicity class and an optional comment. These evaluations and comments are accessible to all platform users within the same entity via the "Selection" icon in

the analysis (see section **Side panel for managing assessments**). A variant can also be evaluated from its dedicated page.

Comments can be entered manually or generated automatically. The interface includes an empty "Classification" field for comments and an "Autofill with DiagAI" button for automatic generation. An indicator appears during comment generation. For existing comments, the "Improve with DiagAI" option enables them to be improved. Each generated comment can be approved, modified or reset. Once validated, comments are saved with their pathogenicity class, date and content. They can be modified or deleted after validation, while retaining their history.

To select the language for DiagAI, click the three dots in the top right corner of the comment box and choose your preferred language from the dropdown menu.

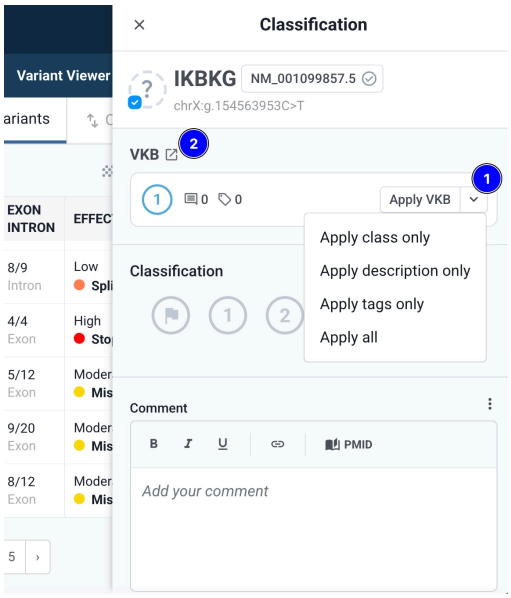


Variants can also be classified by clicking in the classification column, whose header has a flag. To classify a variant, click the + icon. A classification side panel opens, allowing the user to add the pathogenicity class, a comment, and a tag.

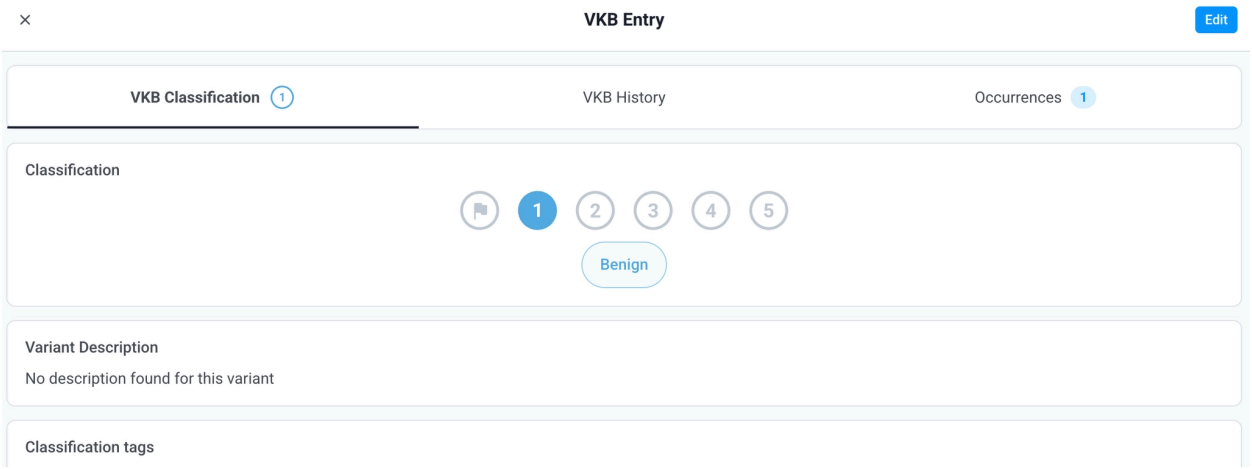
To add this classification to the VKB database, enable the "Add to VKB" toggle. If a classification already exists in the VKB database, click "Apply VKB" to apply it to the current entry. Additional options are available by clicking the arrow next to "Apply VKB" (see step 1 on figure below).

Validate the classification by clicking "Save" at the top right of the side panel.

To edit a classification directly in the VKB database, click on the VKB redirect link (see step 2 on figure below) to open the "VKB Entry" window.



In the "VKB Entry" window, click "Edit" at the top right to modify the evaluation, then click "Save" to apply the changes. The VKB history and variant occurrences are accessible by navigating through the dedicated tabs.



VKB Search. A side panel is available by clicking on the 'VKB' button in the search bar of the SeqOne header. This allows you to search for annotated variants in the VKB database by gene name, transcript, HGVS annotation or SeqOne identifier (Genome_variant_chromosome_location_REF_ALT). The search can be filtered by reference genome (GRCh37 or GRCh38) and/or pathogenicity class.

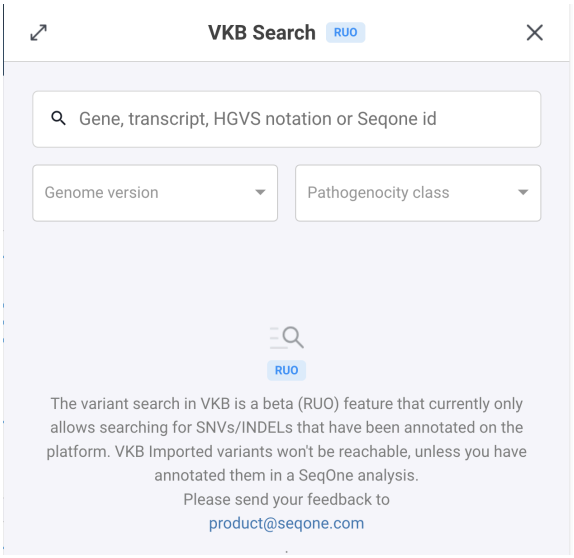


Figure 15: VKB Search window

When you click on an evaluated variant, a more detailed page opens with information about the variant, such as its genomic position, its HGVS annotations on complementary DNA (cDNA) and

on the protein, the pathogenicity classes, the analyses and their associated comments where it was evaluated, and then a link to the analyses.

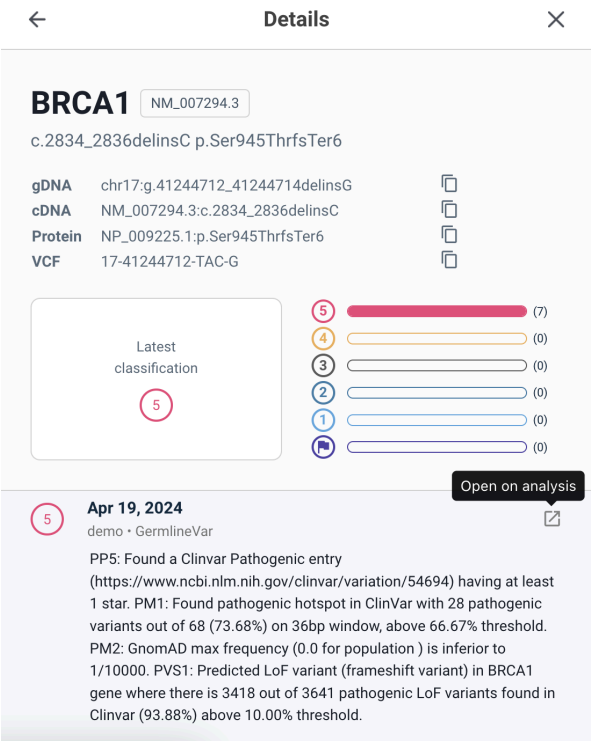


Figure 16 : Variant details window in VKB

Variant page

Clicking on a variant from the variants table opens a side panel containing more information, with several tabs. Users can open the variant page in another tab by clicking on ‘Open in new window’ in the top right-hand corner. The side panel can be closed using the "Esc" key.

Users can evaluate and comment on a variant in this window. Like evaluations and comments made from the main variant table, these evaluations and comments can be seen by all the group members using the platform.

From this page, users can also access variant information, including genotype (GT), allelic frequency (VAF), depth of coverage (COV) and phred score (QA).

The tabs on the variant page are described below.

Variant. The tab is divided into several blocks, described below.

- **Nomenclature**: the different nomenclatures (g., c., p., VCF entry) are listed, with the possibility of copying information from the whole block or from each line to the clipboard using the icon .
- **DiagAI Score:** uses the DiagAI Score calculation summary block available in the DiagAI Score tab on the variant page (see ***DiagAI Score*** paragraph).
- **ClinVar**: information taken from the ClinVar Germline column of the variant table, giving the pathogenicity class assigned to the variant, its clinical status and the number of entries in ClinVar (see **Visualisation and Interpretation** in the **Small Variants** section). The icon opens the ClinVar page in a new tab.
- **Annotations**: transcript impact, exon concerned, distance from the nearest splicing site and type of mutation.
- **Transmission**: links to OMIM and GenCC databases if one or more diseases are associated with the gene.
- **Compound Heterozygous Candidates**, this block is divided into 2 parts:
    - The *Short Variants* section groups all candidate SNVs located on the same gene (based on the gene ID) and ranks them according to their DiagAI score. The variant selected from the Variants page is displayed first and is greyed out in the list.
    - The *CNVs* section indicates whether CNVs have been found, sorted by ACMG class. This section is only available if a CNV analysis has been run, otherwise a warning message is displayed saying '*CNVs not available*'.

  Clicking on one of the results will open a new tab with the variant page (if it's a small variant) or the CNV page (if it's a copy variant). If no CNVs are found, a warning message is displayed with the words "no CNVs found".

- **Tools**: links to other databases (LitVar, LOVD, VarSome, dbSNP, Cosmic, GnomAD, CKB, UCSC, Decipher, Ensembl).
- **GnomAD**: if the variant is known to this database, a table summarising the frequency of the variant in different populations is displayed, and direct access to GnomAD is provided via the icon .

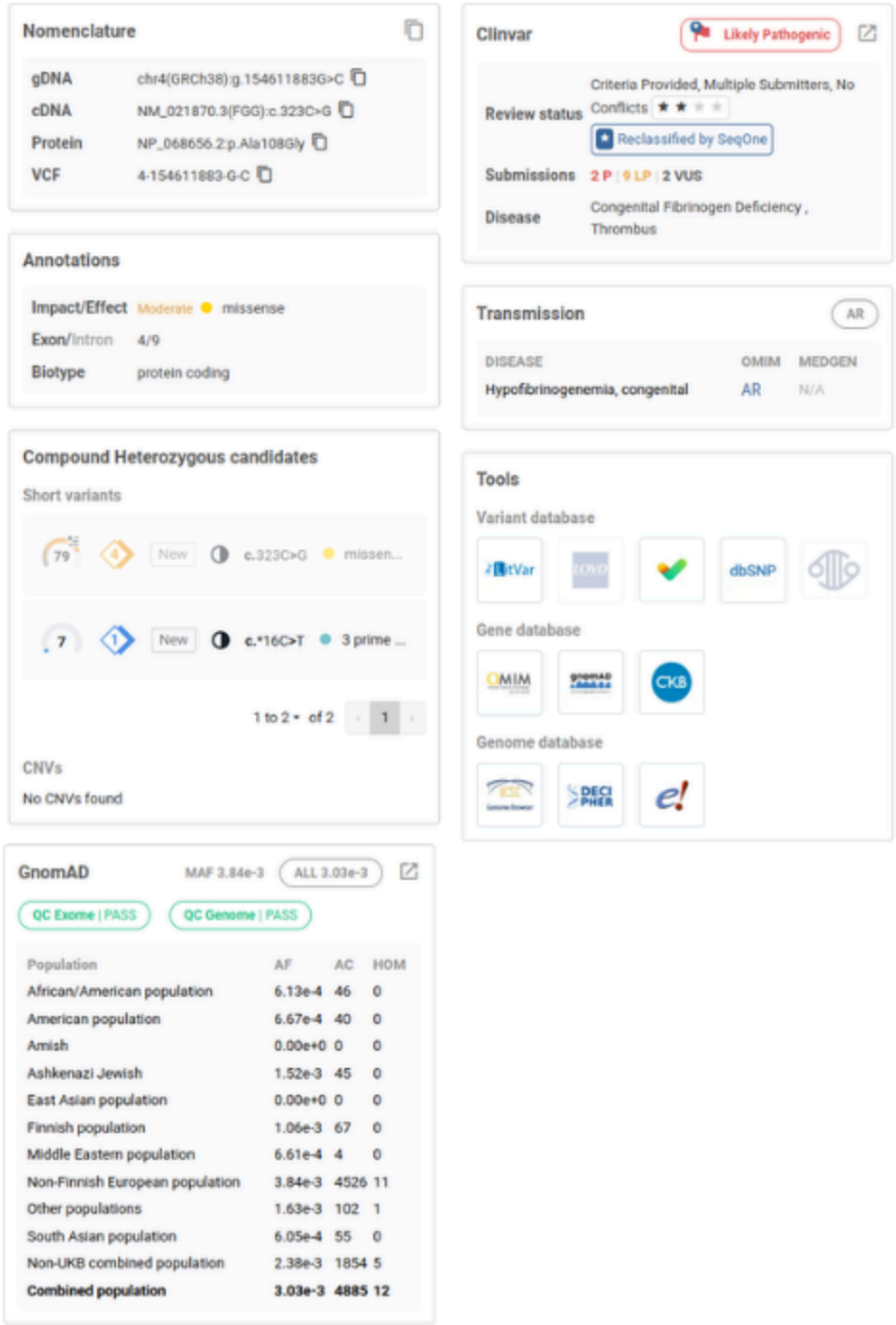


Figure 17: Variant tab on the variant page

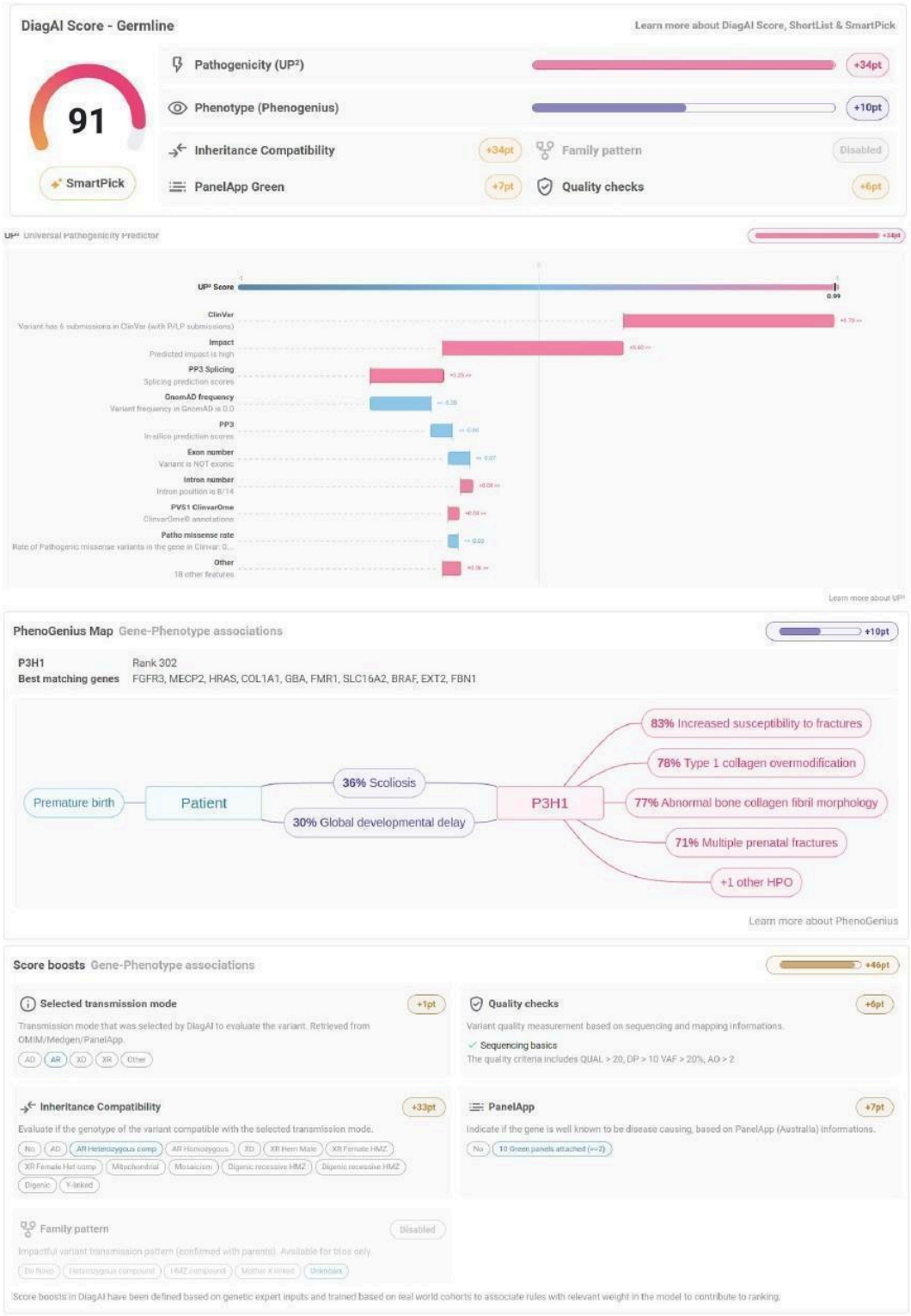


Figure 18: Visualization of the DiagAI tab on the variant page

DiagAI. The tab contains explanations of the DiagAI score assigned to the variant. This tab is only available for GermlineVar, GermlineFamily , GermVar Tertiary and GermVar Tertiary Family (from v1.8) analyses.

The DiagAI tab is divided into 4 main sections:

- A **DiagAI Score - Germline** summary block showing the score, SmartPick status, and listing the main characteristics contributing to the overall score and the associated number of points.
- A waterfall chart explaining the **UP² (Universal Pathogenicity Predictor)** score. A warning message 

is displayed when the user consults the page of a variant whose transcript differs from that used to calculate the UP².
- A **PhenoGenius Map** explainability block, colored if HPOs terms are associated with the sample. The block details the gene rank, a list of the 10 best-associated genes, and the PhenoGenius map, color-coded according to the correspondence between the patient's HPOs and the gene: blue for patient HPOs only, purple for patient HPOs and corresponding to the gene, red for gene HPOs only. The section is grayed out if no HPO terms are entered for the patient.
- A **Score boosts** section integrating 5 sub-features that may or may not contribute (Disabled) to the score: Transmission mode, Inheritance compatibility, Family pattern, PanelApp and Quality check.

ACMG. The tab allows users to view and review the criteria used during the automated ACMG classification. They can modify the selection by checking the different criteria required and then use the classification to evaluate the variant by clicking on "*Use for Evaluation*". It is also possible to modify the level of evidence (*Supporting*, *Moderate*, *Strong* and *Very Strong*) to assign a different number of points to each criterion, which will be taken into account when calculating the final ACMG score. Each level of evidence will award: 1 point for *Supporting*, 2 points for *Moderate*, 4 points for *Strong* and 8 points for *Very Strong*. The points for each criterion will then be added together to assign an ACMG status to the variant:

- *Pathogenic*: 10 points
- *Likely Pathogenic*: 9 to 6 points
- *Uncertain Significance*: 5 to 0 points
- *Likely Benign*: -6 to -1 points
- *Benign*: from -7 points

For criterion BA1, the user cannot select the level of proof, as this is a stand-alone class.

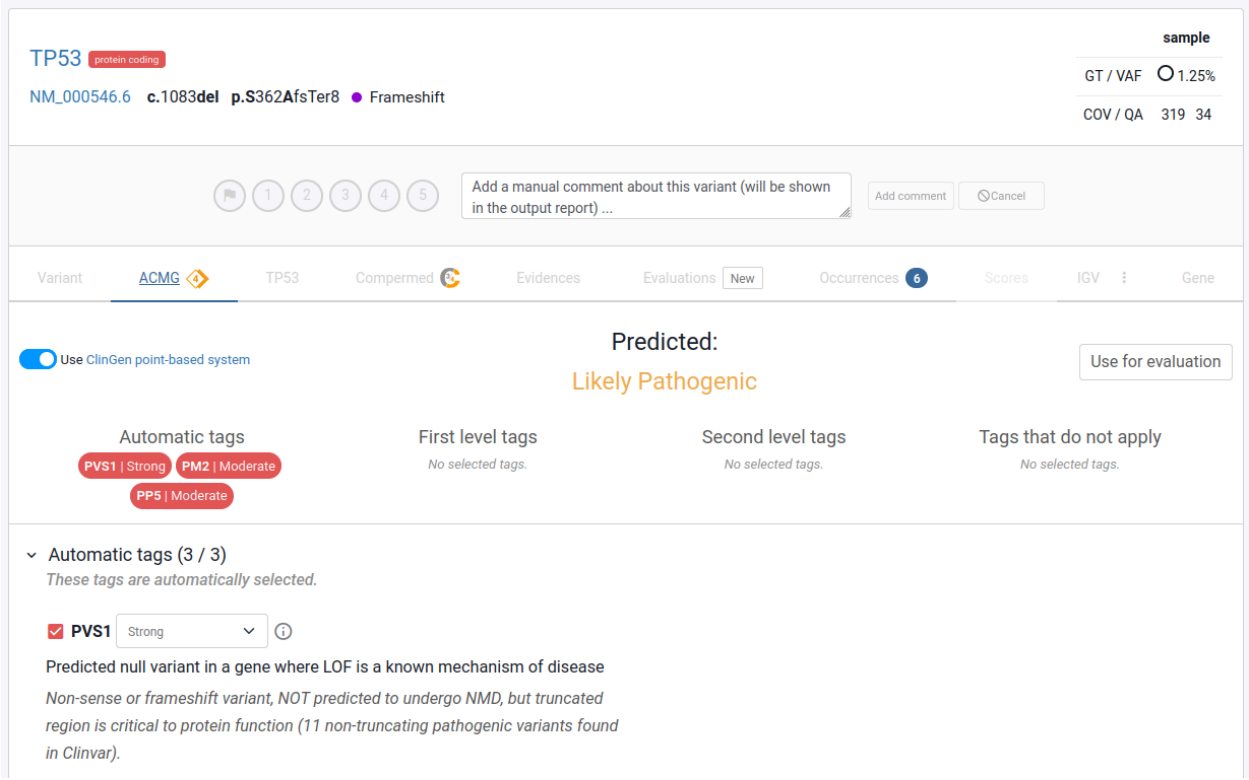


Figure 19: Visualization of the ACMG tab on variant page

TP53. The tab is available only if the variant is annotated on the TP53 gene. This tab provides information on TP53 mutations and their clinical significance according to the IARC[25] database. The TP53 tab is divided into several blocks, described below.

An annotation block (*Annotations*) displays the following information associated with the variant:

- Mutation Rate: substitution rate associated with variant position
- Hotspot (yes/no): presence of the variant in a recurrent mutation site (hotspot)
- Structural Motif : structural motif affected by the mutation
- Domain: functional domain of the protein affected by the mutation

The Functional Data block displays classifications and functional information associated with the variant:

- DNE Class: dominant-negative effect (DNE) on transactivation of wild-type p53 protein
- DNELof Class: functional classification for loss of growth suppression and negative dominant activities
- Transactivation Class: functional classification based on global transcriptional activity (TA)
- Structure Function Class: functional predictions.

[25] https://tp53.cancer.gov/

The ‘Graphics’ and ‘Sample details’ blocks provide details of the samples referenced in the IARC database for the variant. The user can choose to view data associated with either somatic or germline samples, by selecting Somatic or Germline.

The ‘Graphics’ block displays two graphs, with drop-down menus, showing the distribution of samples according to their characteristics.

The Sample details block displays detailed sample information in tabular form.

ComPerMed. (somatic analysis only). The pathogenicity class of each variant is calculated following the recommendations of the Belgian expert committee ComPerMed (Commission for Personalized Medicine). The user has access to detailed information on this classification from a dedicated tab on the variant page.

Evidences. The tab provides a list of therapies associated with the variant selected from the JaxCKB database. The user can scroll down each line to find details of the therapies or indications.

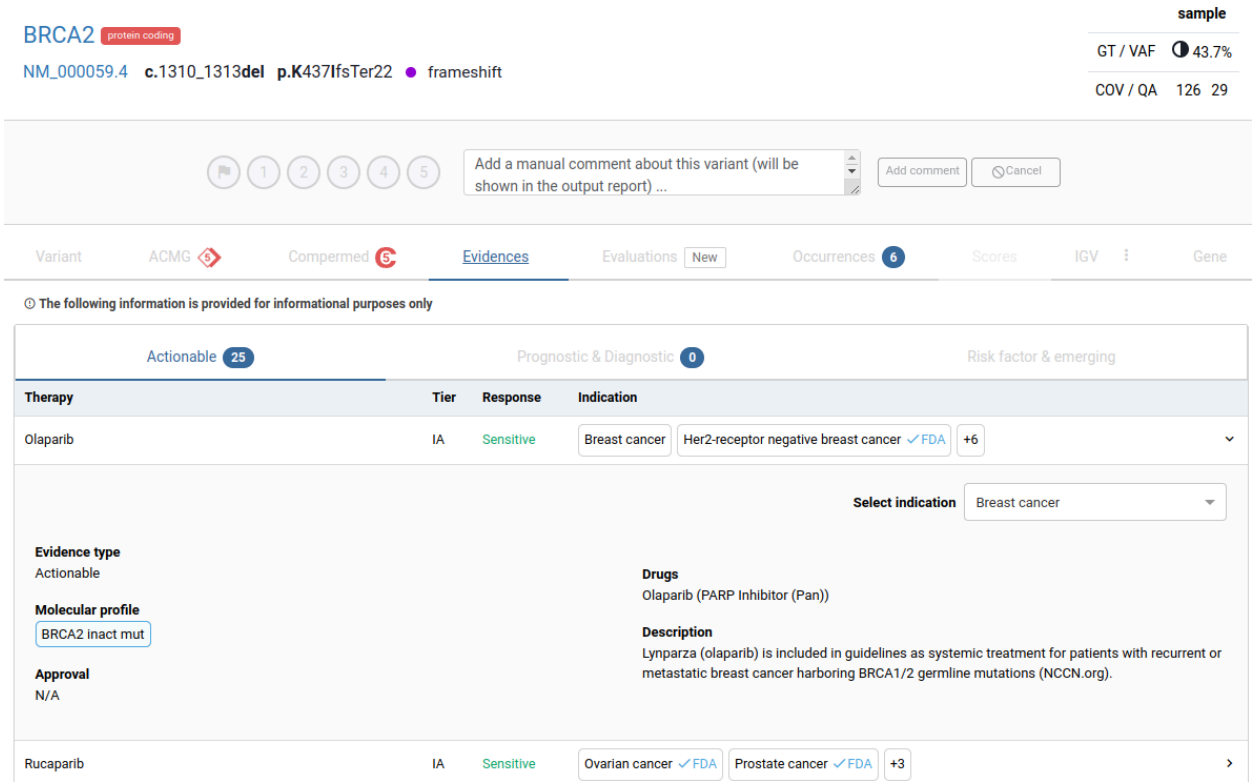


Figure 20: Visualization of the *Evidences* tab in the variant page

Evaluations. This tab provides access to the history of evaluations and interpretations for the variant over time:

- In the current analysis (left section of the tab)
- In the other analyses carried out within the group (right section of the tab)

Occurrences. This tab lists all other samples where the variant is present, without considering any evaluation results. For each instance, the following information is provided: sample name, allelic frequency, a link to the corresponding analysis, and the project name.

Publications. This tab provides access to scientific literature associated with the variant and the gene. Click either the "Variant" or "Gene" sub-tab to view PubMed publications referencing the variant or the gene. A PMID link redirects to the PubMed record, and the full text Link opens the full article when available.

Scores. The tab available for variants with a missense, start lost or stop gained effects. With the exception of these cases, the score will be greyed out. This tab allows you to view the results of the prediction scores of several tools such as *REVEL, DANN, FATHMM, FATHMM-MLK, LRT, MetaLR, MetaSVM, MutationTaster 2, PROVEAN, SIFT and PLi. It also provides access to conservation scores such as GERP++, PhastCons100way Vertebrate, PhastCons20way Mammalian, PhyloP100way Vertebrate, PhyloP20way Mammalian* and *SiPHhy*.

IGV. This tab allows the user to view reads aligned to the position of the variant through the web version of the IGV Genome Browser. In areas of high read coverage, reads are subsampled by default, i.e. not all reads are displayed. The default viewing parameters are applied, i.e. a maximum of 100 reads per 50-nucleotide window. These regions can be identified by the black line below the coverage profile (downsampling). Clicking on it shows the number of reads removed from the display (# downsampled). A menu is available from the icon  , offering three options:

- *Open in new tab*: opens the IGV variant view in a new tab.
- *Track setting*: allows you to add other samples from the same or different projects to the view.
- *Open local IGV*: redirects the IGV Desktop application to the coordinates of the variant currently being interpreted. To view SNVs or other variants, the sample BAM must first be loaded into the IGV Desktop application. This option is only compatible with IGV Desktop versions ≥ 17.

IGV supports split views, for example, by copying two sets of coordinates into the locus search bar, separated by a space. Various settings are available by clicking on the cogwheels to the right of the Genome Browser. For example, to view unaligned reads (soft clips) directly, select the "Show soft clips" option.

Gene. The tab contains additional information about the gene and the associated bibliography.

Generating a clinical report

The user can edit an exportable clinical report in PDF and .docx format by clicking the "Generate report" button from the “Selection” tab while navigating in the analysis.

At the top of the editing page, the user can select a template, set the report status to new or draft, select a sample (multiple samples can be selected for family analyses), and select the editing language. The user can also switch between the “Edition” view and the “Preview” view before saving a draft or proceeding to the next step using the “Save Draft” and “Next Step” buttons.

Report templates can be created in the “Reporting” section, under “Template list” in the entity settings.

In the Edition view, the user can edit the different sections of the report and select multiple options. In the Preview view, the user can review the report before continuing the edition, saving the draft, or validating the report.

The user can save a draft and access it again from the “Selection” tab, then "Generate report", as when creating a new report. When a PDF report is generated, it is automatically displayed in the "Reporting” tab of the analysis. Only after validation can the report be exported as .PDF or .DOCX. The user can delete the report or create additional reports for the same analysis.

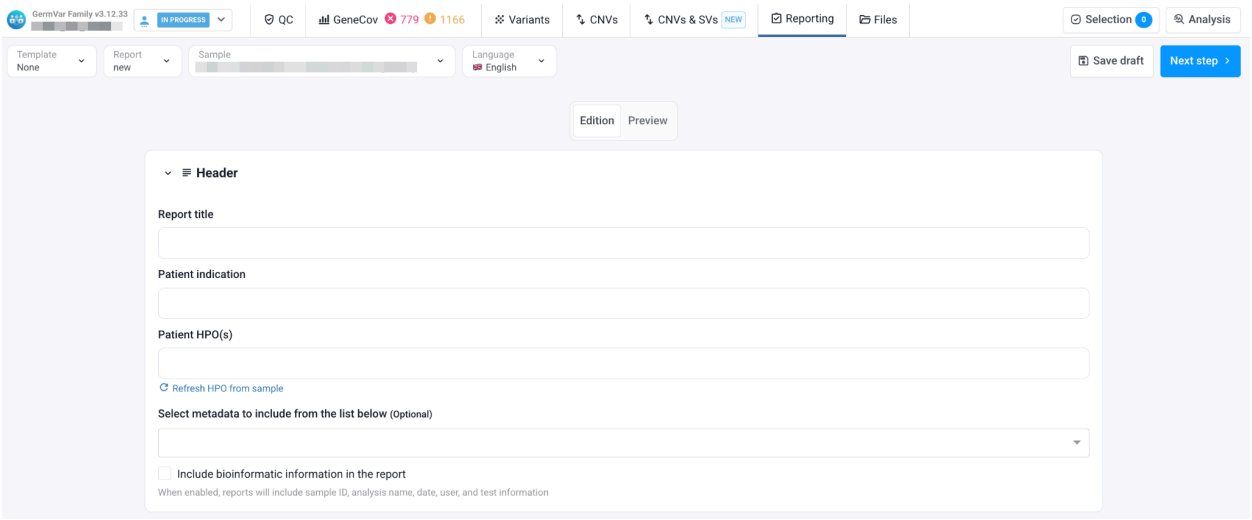


Figure 21 : View tab of the clinical report.

12.1.5. CNV

12.1.5.1. CNV in gene panels

Methodology

Detecting CNVs in the gene panel data, in GermlineVar, GermVar/GermVar Family, SomaVar and SomaHemato, is based on selecting control samples from the co-analysed cohort, according to their correlation.

The correlation is calculated from a coverage table, which in turn is calculated from all the regions in the manifest, but excluding sex chromosomes.

The criteria for a sample to be eligible for the CNV analysis are as follows:

- For a region to be retained in the CNV analysis, the minimum coverage is 50 reads.
- The threshold correlation value for selecting control samples is 0.90 for the germline context and 0.75 for the somatic context.
- The minimum number of controls required for each sample is six.

The gender of the samples (sample metadata) needs to be defined before launching CNVcapture or SomaCNVCapture in order to perform the calling on genes localized on sex chromosomes. Lastly, at least eight samples are needed to perform a CNVCapture or SomaCNVCapture analysis on SeqOne.

The criteria applied to the user manifest for a panel gene to be eligible for CNV analysis are as follows:

- manifest regions are annotated from the canonical transcript determined according to the NCBI MANE Select resource. If no MANE Select transcript is available, annotation is based on the NCBI RefSeq resource.
- only exons for which there is ≥ 80% overlap between the manifest interval and the exon are analyzed. If no gene exon overlaps ≥ 80% with the manifest interval, the gene will not be included in the CNV analysis.
- only introns for which there is ≥ 60% overlap between the manifest interval and the intron are analyzed.

CNVCapture analysis (version ≥ 2.4) includes regions located 20kb upstream and downstream of genes covered by the manifest, if such regions are included in the manifest design. Specifically, each of the 20 segments of 1kb constituting the 20kb region is analyzed if at least 20% of the segment overlaps the initial manifest.

View and interpretation

The *CNVs* tab will display a chart representing the list of genes in which copy variations have been identified. This tab is available in the worksets described in the Methodology section above.

The terminology used to describe the status of called genes is as follows:

- Gain: the gene is affected by partial or total amplification
- Loss: the gene carries a partial or total deletion
- Normal: the gene carries no CNV
- Partially not called: the gene is partially called
- Not called: the gene is excluded from CNV analysis

The partially not called status is assigned to a gene when its copy status is normal but at least one exon has not been called (with "failed" status)

Despite its eligibility for CNV analysis, a gene is excluded (Not called) from the analysis if :

- the gene is on a sex chromosome and the gender metadata has not been filled in,
- all exons fail according to the criteria specified below: number of controls, percentage of variation. The "failed" status of a gene is defined on CDS only. Intronic and upstream/downstream regions are excluded if CDS exons are called.

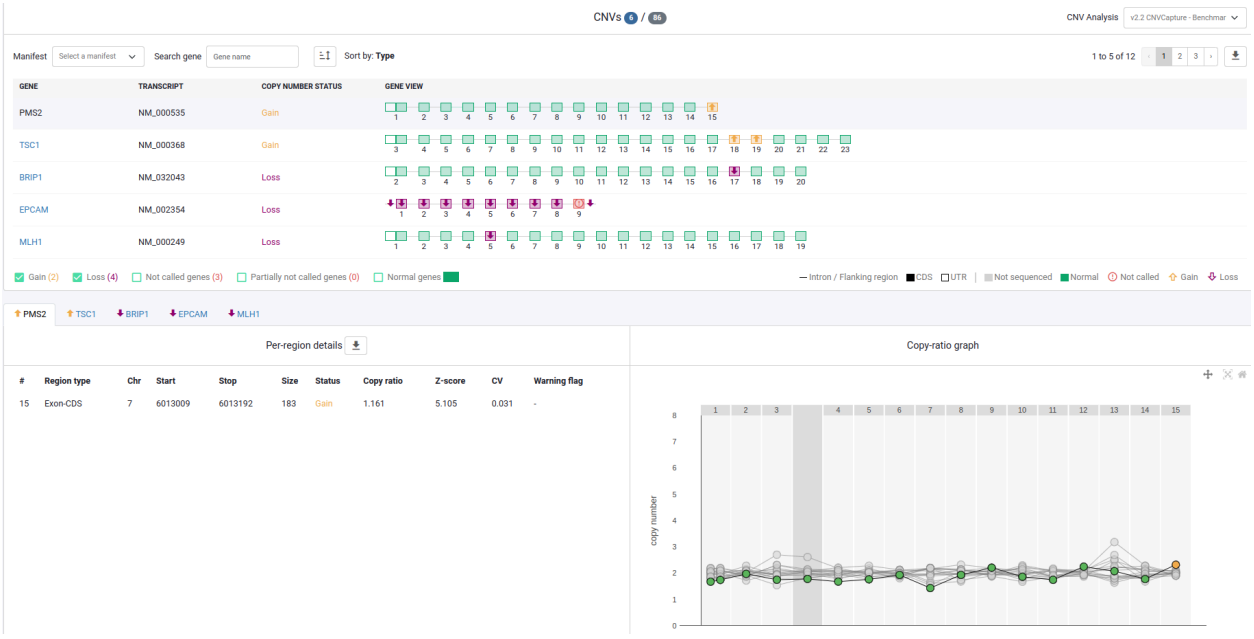


Figure 22: Visualization of the CNVs tab in a GermlineVar analysis.

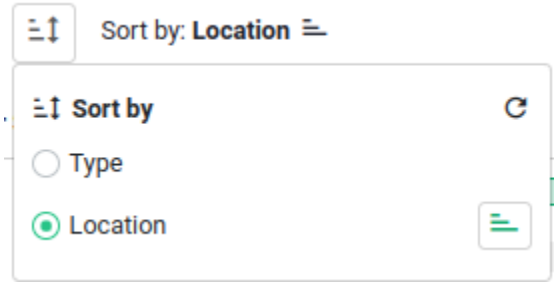
Filters and sorting options

The filter menu can be used to:

- Restrict the display to a subset of regions/genes using a BED file that has previously been uploaded to the interface with the '*Manifest*' feature
- Search for a specific gene with the '*Search gene*' feature

Two CNV sorting options are available:

- **Type** (default): current sorting logic (Gain then Loss)
- **Location**: used to sort CNVs according to chromosomes and their starting position, in ascending or descending order, by clicking on the sort icon



The user can also select genes according to their status: Gain, Loss, Normal genes, Partially not called genes or Not called genes.

Details by region

The Per-region details table shows the details for each region for each gene:

- **#:** region number (exons and introns)
- **Region type**: type of region (upstream, utr-exonic, coding-exonic, intronic, downstream)
- **Chr**: chromosome
- **Start/Stop**: CNV start and end position
- **Size**: size of CNV (bp)
- ***Status*:** region status (gain = duplication, lost = deletion, normal = no CNV, not called = region was not included in CNV analysis)
- **Copy-ratio**: ratio of copy number to normal (n=2).
- **Z-score**: number of standard deviations separating a result from the mean. High positive or negative values indicate strong confidence in the CNV call.
- **CV**: coefficient of variation measuring the relative dispersion within the cohort samples.
- **Warning flag**: An error message may occasionally be displayed according to the following criteria:
  - *Low zscore*: Z-score absolute value <4
  - *High variation*: coefficient of variation (CV) > 0.1
  - *Low coverage*: region coverage for this sample <100
  - *Number of controls:* The number of controls is insufficient to be able to perform the calling on this region

- *Percentage of variation:* The variation is too large across samples for this region to be able to perform the calling
- *Read depth*: insufficient region coverage (<50) and copy ratio >0.1
- *Missing Gender*: the region is located on a sex chromosome and the gender metadata has not been entered.

Warning: a region rich in soft clips, such as a large indel or an Alu insert, may be interpreted as a drop in coverage by the CNVCapture and SomaCNVCapture worksets and detected as a loss of copy, if the size of the insertion is of the same order as that of the called region.

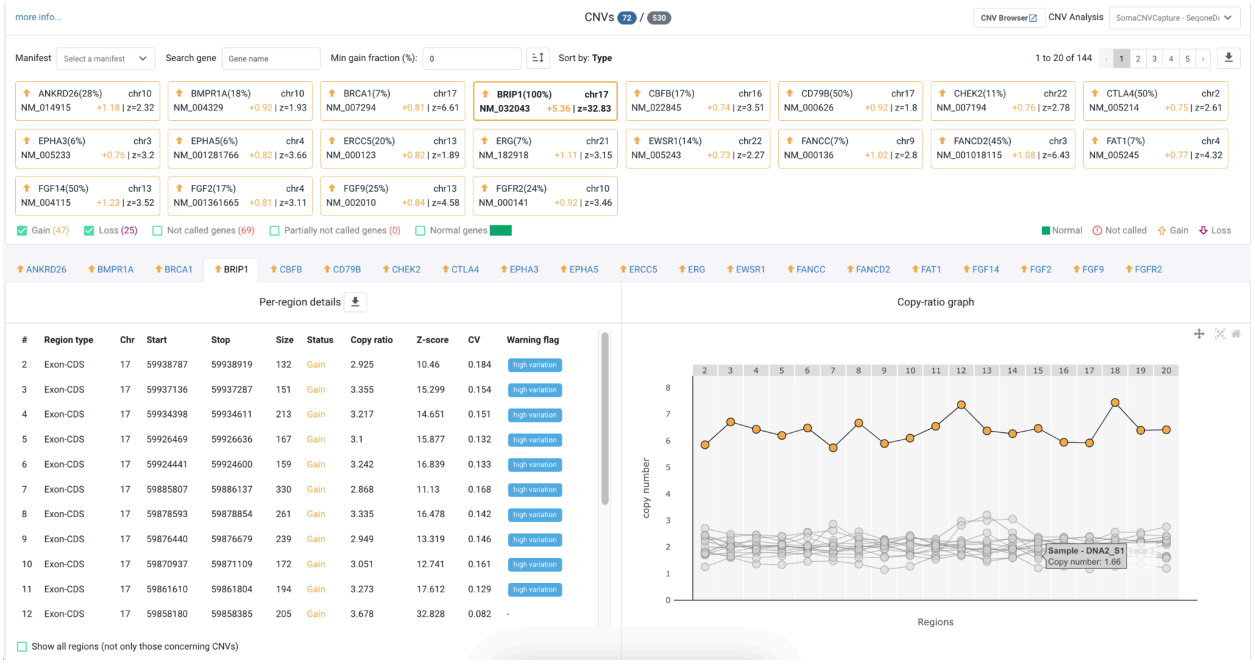
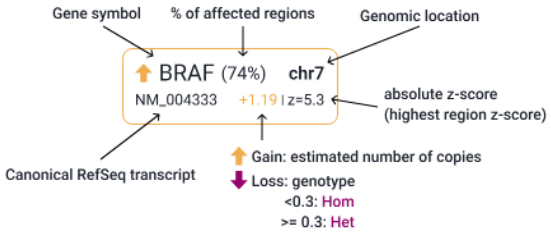


Figure 23: Visualization of the CNVs tab in a SomaVar analysis

In case of **SomaCNVCapture** analyses, although the detection method is similar to that of CNVCapture the scale used for the graphical representation of the variants is the gene, rather than the exon. The representation is shown as a series of boxes. This provides the gene symbol, genomic position, canonical Refseq transcript, absolute z-score and, in the case of gain, copy number or, in the case of loss, genotype.



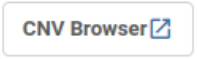
A pop-up window explaining these boxes is available above the filters in the '*more info...*' section.

The percentage displayed in each box reflects the fraction of the gene (*min gain fraction*) affected by the detected CNV. The chunk ratio can be used to filter the results. By default, the filter is set to 0% and only applies to gains of copies.

Deletions detected by this pipeline are not subject to this filter, and should be interpreted with caution. Indeed, the ability of this tool to detect variations of a gene copy depends on the nature and quality of the sample, in particular the tumor fraction.

Warning: If both an amplification and a deletion are detected on the same gene, the gene will be considered as amplified by default, and therefore subject to the min gain fraction filter described above.

CNV browser

An icon 

is available in the CNV display tab of the GermlineVar, GermlineFamily, SomaVar or SomaHemato analyses to open a CNV Browser in a new tab. The CNV Browser graph displays the B allele frequency (BAF) of SNPs (Single Nucleotide Polymorphism) present in the analyzed sample manifest and present in a dataset of SNP arrays from Illumina and SNPs from other backbones such as Agilent's OneSeq, to facilitate visualization of certain types of events such as LoH (Loss of Heterozygosity).

The BAF is the frequency of the non-reference allele (B allele) observed for a germline heterozygous SNP.

By default, the graph represents the BAF of variants (y-axis) over the whole genome (x-axis), broken down by chromosome (1-22, X and Y). A filter allows the user to display the variants on a given chromosome in the top left-hand corner.

In the context of CNV analyses, this BAF information can be used to show an alteration in allele-specific copy number, loss of heterozygosity or allelic imbalance.

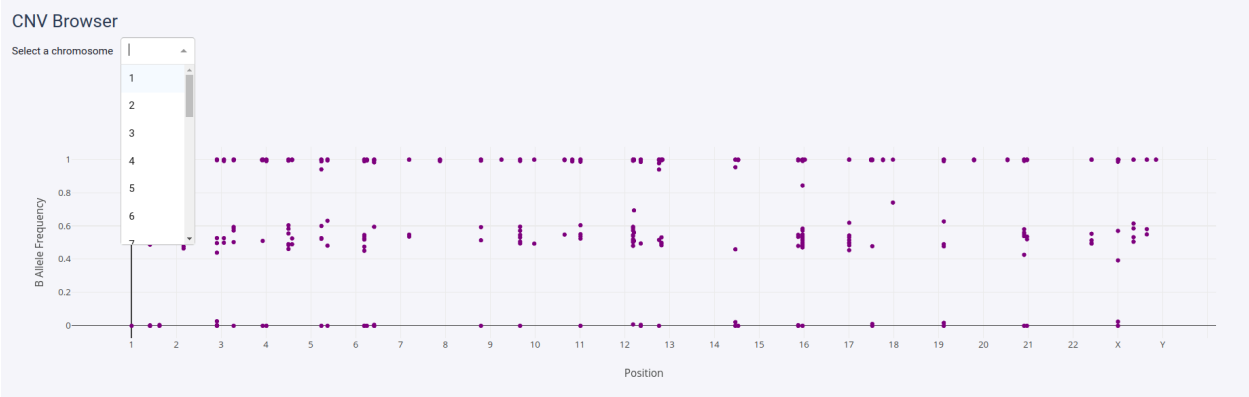


Figure 24: Visualization of the CNV Browser tab in a GermlineVar analysis.

CNV data export

The icon at the top right can be used to export the CNV analysis data in the form of a .tsv file containing the following columns:

- gene (gene_name),
- RefSeq transcript (refseq_id),
- gene status (status),
- chromosome (chr),
- genomic coordinates of the CNV: start and end positions,
- fraction of the gene affected by the CNV (gene_fraction_cnv),
- ratio of copy number (copy_ratio) to normal (n=2),
- number of standard deviations separating a result from the mean (zscore)
- number of exons overlapping the CNV (exons), in the case of a CNVCapture analysis.

For each gene, the icon next to "Per-region details" can be used to specifically export the CNV analysis data for the selected gene in the form of a .tsv file containing the following fields:

- gene (gene_name),
- chromosome (chr),
- genomic coordinates: start and end positions,
- exon or intron number (exon_intron_number),
- status,
- type (exon or intron, type),
- associated error messages (warning_flags, fail_flags),
- adjusted critical probability value (padj),
- critical probability value (pvalue),
- ratio of the number of copies to the normal (ratio),
- measure of the relative dispersion within samples (variability),
- standard deviation of the result from the mean (zscore)

12.1.5.2. CNV in SomaCGP and SomaLBx analysis

Methodology

SomaCGP and SomaLBx use PureCN[26] to detect CNVs and LOHs (loss of heterozygosity). PureCN identifies somatic variants from variant detection data in order to assess the ploidy and tumor purity of the sample. This information is used to determine the copy number of each event.

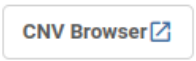
Visualization and interpretation

The CNV visualization tab in SomaVar Tertiary analysis works identically to the one in SomaCGP and SomaLBx as described below.

CNVs are presented in a filterable table. Each CNV displays the following information:

- **Gene**: gene name
- **Location**: cytoband
- **Event coordinates** / **Event size**: event coordinates and size
- **Transcript:** reference transcript
- **Exons**: Exons affected
- **Type**: type of event (LoH, Loss or Gain)
- **Copy number**: absolute number of copies
- **Function**: gene function if applicable, Oncogene or Tumor Suppressor Gene (TSG)

The user can search for CNVs affecting a given gene via a search field and choose whether or not to display genes present in LOH regions by checking the '*Show genes in LOH regions*' box. The results table can be exported using the download icon in the top right-hand corner.

The icon 

opens a *CNV Browser* in a new tab. This tab displays a graph showing the B allele frequency (BAF) of SNPs in the sample analyzed and LOH SNPs from selected publications and the OneSeq Agilent backbone. A second graph represents the logRatio data generated by PureCN. More details are available in the section **CNV: Gene Panels** under **CNV browser.**

[26] https://github.com/lima1/PureCN

12.1.5.3. CNV in exomes and genomes

The CNV visualization tab for exome data is available in GermlineVar, GermlineFamily, GermVar, and GermVar Family analyses when using FASTQ files as primary input and running a CNVExome analysis. This tab also displays CNV results for genome data in GermVar and GermVar Family analyses using FASTQ files, as well as in GermVar Tertiary and GermVar Tertiary Family analyses using VCF files or CGH microarray data.

Methodology for exome

This methodology applies specifically to the CNVExome tool. Just like the CNV analysis pipelines on gene panel data, the process of identifying variants in this case depends on an analysis of the depth of coverage. However, the CNVExome analysis using the GATK GermlineCNVCaller tool is individual, since each sample is compared to a model HHMM (*Hierarchical Hidden Markov Model*) calculated by SeqOne.

To take account of the technical variations involved when preparing samples, this model is based on a large cohort ranging from one to several hundreds of samples sequenced by the laboratory using:

- The same library preparation protocol
- The same capture kit
- The same sequencing platform and read length

These requirements and the variable number of samples required to set up the model can be explained by the need for a cohort offering a homogeneous level of coverage. Several factors must be taken into account, such as the sequencing depth, the degree of variability within the cohort, and the reproducibility of the sample preparation protocols.

The gender metadata (see **Sample metadata**) must be filled in before starting the CNVExome analysis to enable ploidy verification for the sex chromosomes. If it is not completed, a "Ploidy" warning message is displayed on the CNV display interface.

Limitation: The GATK GermlineCNVCaller documentation recommends excluding PAR (PseudoAutosomal Regions) regions, i.e. PAR1 and PAR2 regions, if their contribution to total X coverage is disproportionate. In fact, ploidy is difficult to calculate for the ends of the X chromosome corresponding to the PAR regions for many models calculated by SeqOne. As a result, ploidy is not calculated correctly for these regions, which are detected artifactually at 1 copy, whereas the entire X chromosome is detected at 2 copies.

Users are invited to contact the user support team at support@seqone.com to find out more about the procedure for setting up CNVExome in their group.

Methodology for genome

For versions up to v3.6, GermVar and GermVar Family analyses on genome data integrate the Sentieon CNVScope[27] tool for CNV detection. Starting with version v3.7, GermVar and GermVar Family worksets use the Cue[28] tool for CNV calling on genome data.

Segmental duplication data are detected using Sentieon's SegDup Caller, which is specifically designed to identify copy number events in highly homologous genomic regions on WGS data only using GermVar. This caller targets a curated set of clinically relevant genes known to be challenging for standard variant detection approaches. The resulting events are automatically integrated into the analysis pipeline and visualized in the CNVs & SVs viewer as CNV (both gain and loss) enabling the identification of variants in genomic regions that were previously inaccessible to standard analysis.

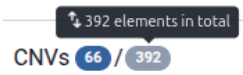
View and interpretation

An interpretation interface is available in the ***CNV*** tab for CNVExome, GermVar and GermVar Family, GermVar Tertiary or GermVar Tertiary Family analyses. The interface is divided into two sections: a CNV results table on the left and a detailed CNV page with multiple tabs on the right.

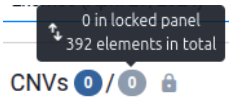
CNV results table

A results table displays the CNVs identified, along with a menu of filters based on their presence in morbid OMIM genes [https://omim.org/], size, frequency of CNV appearance in the cohort, or LogRatio in the case of MicroArray CGH analysis.

The number of filtered CNVs (selected CNVs) and the total number of CNVs detected (elements in total) are displayed at the top of the CNV table.



A padlock icon indicates that the analysis uses a restricted panel, and the CNVs details appear in the associated tooltips.



A warning message 

is displayed when an abnormal ploidy value is detected and/or when the gender metadata (see **Sample metadata**) has not been filled in before the analysis is launched.

By clicking on the "Ploidy" icon, the user can view the ploidy for each chromosome and the patient's gender. Abnormal values are highlighted in yellow.

[27] https://support.sentieon.com/appnotes/cnvscope/
[28] https://github.com/PopicLab/cue

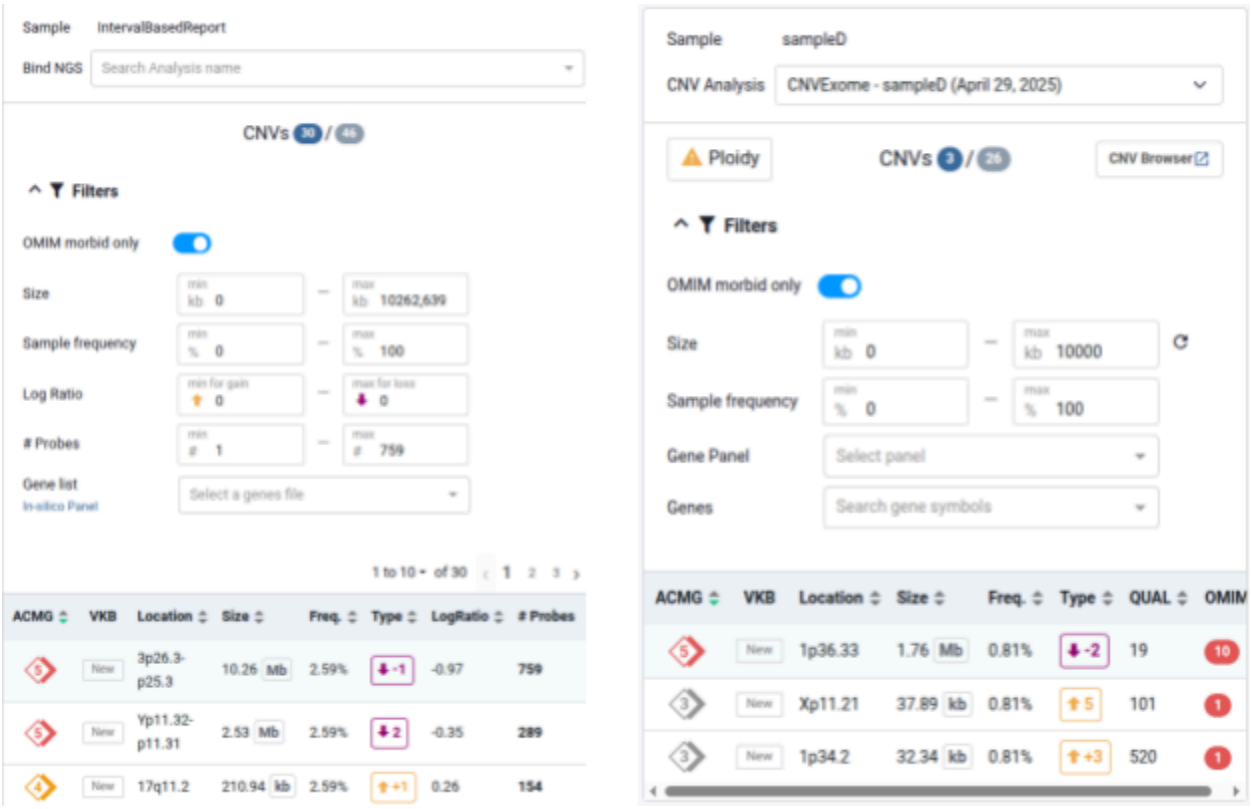


Figure 25: Overview of detected CNVs and filter menu of a CGH analysis (left) and a standard analysis (right).

In the case of a CGH MicroArray analysis combined with an SNP array (CGH+SNP), the loss of heterozygosity (LOH) events detected are reported in the CNV results table. These events correspond to copy-neutral loss of heterozygosity (CN-LOH). In other words, the region retains two copies, but these are identical, characterised by the presence of homozygous SNPs.

In the case of a CGH MicroArray analysis, it is possible to select a non-CGH SNV analysis (GermlineVar, GermlineFamily, GermVar Tertiary and GermVar Tertiary Family) from the 'Bind NGS' drop-down menu in order to display the variants that overlap with the detected CNV.

Filters

OMIM morbid only. This filter is activated by default to easily identify CNVs affecting clinical genes, annotated OMIM Morbid on the basis of their NCBI identifier (NCBI gene_id) or, failing that, on the basis of the gene name and synonyms (gene_symbol and gene_synonyms). Simply lift this filter to display all the CNVs detected in the sample.

Size. The user can filter the CNVs according to their expected size (in kb).

Sample frequency. The user can filter CNVs according to their size and frequency. This frequency corresponds to the occurrence of the CNV among all the samples analysed up to the analysis date and compared with the same model, including the samples used to build the model. The Sample frequency is based on a 95% overlap and considers the same type of CNV (gain/loss/deletion). By hovering the mouse over its value displayed as a percentage, the user will see the corresponding fraction appear. The frequency may take several minutes to update after opening the analysis.

Gene Panel. Like an in silico panel, this filter allows the user to select and apply a list of genes to filter the CNVs detected in the analysis using the drop-down menu (see **Managing user parameters**). This filter is not available if a restricted panel has been applied to the analysis from the Panel tab of the analysis management side pane (see **Panel modal view** section).

Genes. Allows you to filter the CNVs detected containing the genes selected via this filter.

Log Ratio. For analyses based on MicroArray CGH data, the Log Ratio (between the raw signal and the control signal) present in the Agilent input files is displayed in an additional column. The associated filter allows the user to select minimum values for copy gains (min for gain) and maximum values for copy deletions (max for loss). LoH (Loss of Heterozygosity) events are still retained even after this filter has been applied.

#Probes. For analyses based on CGH MicroArray data, the #Probes column indicates the number of probes used in each region (exonic, intergenic or intronic). These values are taken from the IntervalBasedReport.xls input file, and the associated filter allows the user to select the minimum and maximum number of probes used. In a combined CGH+SNP analysis, the number of probes indicates the number of homozygous SNPs in the LOH region.

CNV sorting

The list of CNVs can also be sorted by ACMG pathogenicity class (*ACMG*), genomic position (*Location*), size (*Size*), sample frequency (*Freq.*), copy number (*Type*), quality score (QUAL) in GermVar Tertiary and by LogRatio in CGH MicroArray (*LogRatio*).

ACMG. The pathogenicity class of each variant is calculated semi-automatically following the recommendations of the *American College of Medical Genetics and Genomics*. The pathogenicity class is the default sorting mode in the CNV table.
LoH (Loss of Heterozygosity) events in CGH MicroArray are always assigned to class 3 (VUS).

Type (Copy number). The copy number variation is displayed (e.g.: +2, -1, +3) for CNVs detected on autosomal chromosomes (1 to 22) as the copy number is set at 2 by default.

However the absolute copy number is displayed (ex: 0, 1, 2) for CNVs detected on gonosomes (X and Y sex chromosomes) and not the fold change.

Unlike gene panel analysis, gender-related sample metadata is not required to include sex chromosomes in exome CNV calling.

For combined CGH+SNP analyses, CN-LOH events, which retain two copies, are represented in the Type column by the badge LOH .

QUAL. The Phred quality score is calculated in CNVExome by the GATK caller. When a CNV is made up of several genomic intervals, the highest QUAL score among all the intervals is then assigned to the CNV. For tertiary analyses, the QUAL score displayed is that obtained from the input VCF file (DRAGEN or ONT).

CNV browser

The icon opens a CNV Browser in a new tab displaying the B allele frequency (BAF) of the SNPs (Single Nucleotide Polymorphism) present in the sample analysed. The graph is described in the previous section (CNV: Gene Panels, under CNV browser).

Detailed CNV page

CNV summary. This initial tab features a wealth of information and annotations about the selected CNV, including:

- Details of automatic ACMG classification
- The clinical regions and genes included in the interval affected by the variant. The annotation of clinical regions is based on the ClinGen and OMIM databases. Only OMIM Morbid annotated clinical genes are displayed by default. To display all the clinical genes, the user can activate the "Show all genes" function on the right of the page.
- Overlaps between the detected CNV and the DGV database.
- The small variants (SNVs and indels) detected in genes affected by CNV. By default, the SeqOne-diag filter is activated. To deactivate this filter, click on SeqOne-diag and

deselect SeqOne-diag from the filter profiles. To display all the variants, users can activate the "Show all variants" function on the right of the page.

- A dedicated section for the variants knowledge base (VKB) for identifying variants over 95% identical that have already been validated in the group and their respective evaluations.

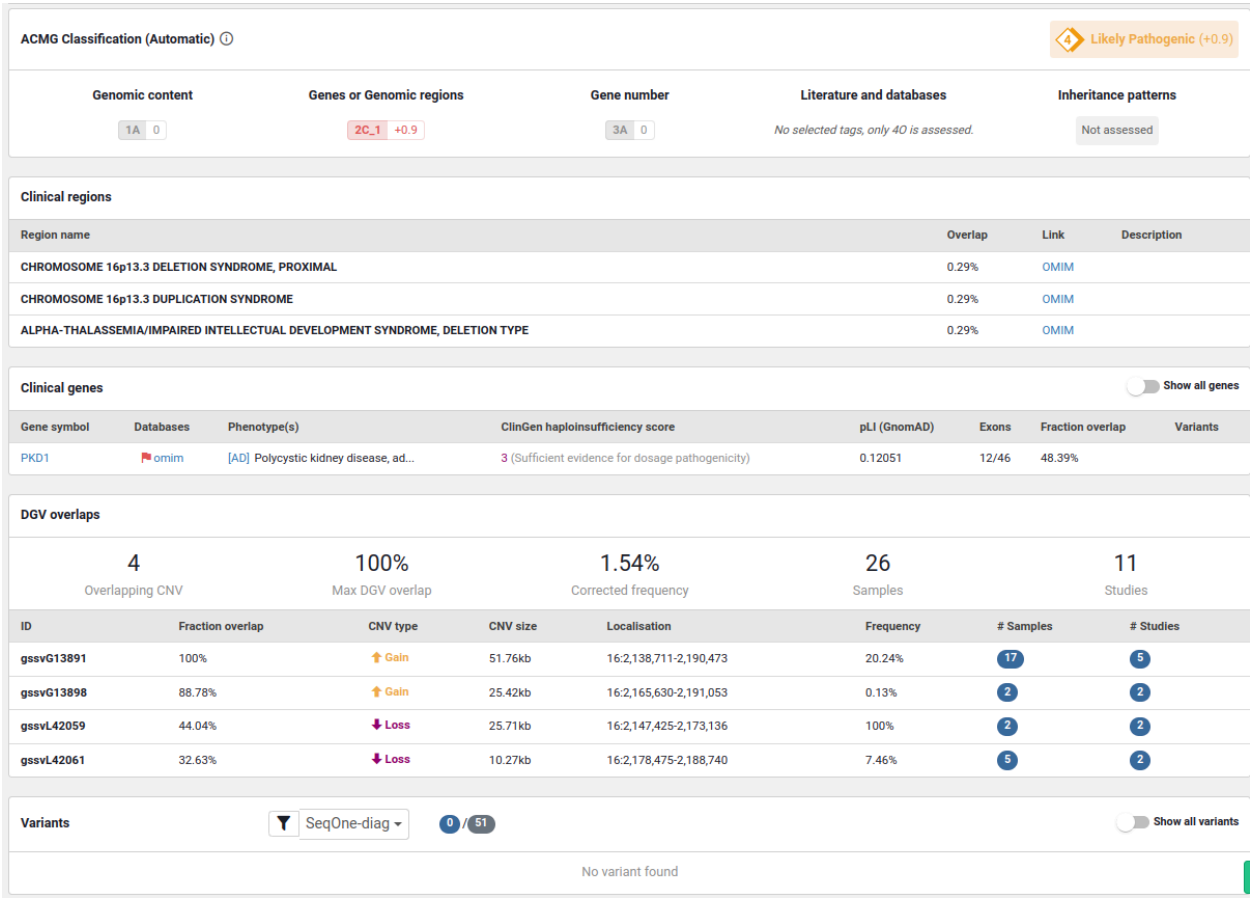


Figure 26: Overview of the CNV summary tab

- The Genome browser IGV where CNVs can be visualised (*not available in GermVar Tertiary and GermVar Tertiary Family analyses*). To check the stability of the signal, up to five other patients can be displayed in addition to the individual analysed from the configuration menu of the chart representing the copy number (DCN).

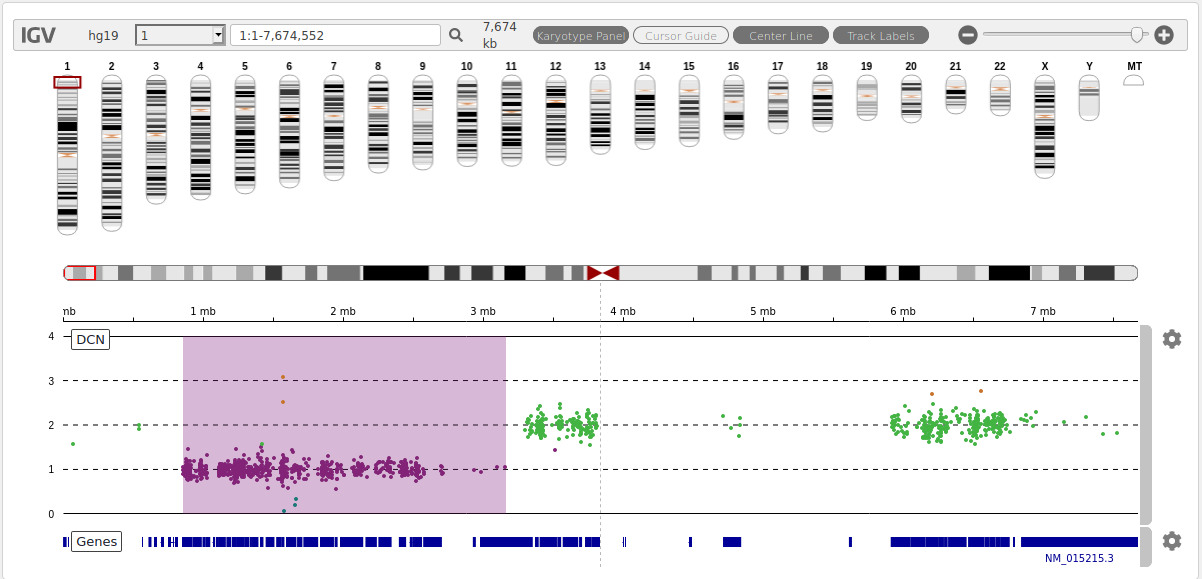


Figure 27: View of the CNVs from an integrated IGV browser

The color code is as follows for the selected patient: blue (0 to 0.5 copies), purple (0.5 to 1.5 copies), green (1.5 to 2.5 copies), orange (2.5 to 3.5 copies) and red (>3.5 copies). Other samples are displayed in gray.

CNV Occurrences. This tab is used to identify other samples containing the CNV, independently of any evaluation. For each occurrence, the following information will be entered:

- the name of the sample (Sample),
- the name of the analysis and a link to it (Analysis),
- the name of the project (Project),
- the genomic location of the CNV (location),
- the size of the CNV (Size),
- the CNV type,
- the fraction overlap.

The hits are sorted in descending order of the overlap percentage. It is possible to filter the CNVs displayed according to the minimum fraction of overlap by filling in the ‘min fraction overlap (%)’ field at the top of the list. By default, this filter is set to 80%.

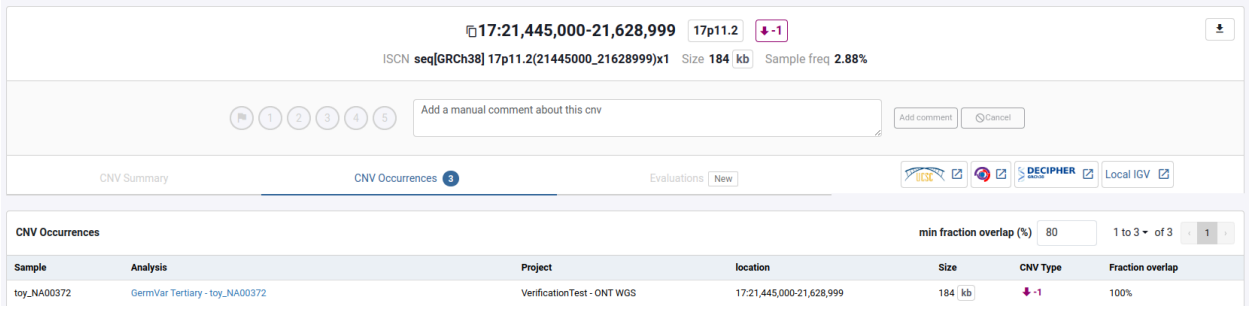


Figure 28: Overview of the CNV occurrences tab in the CNVExome analysis

Evaluations. The pathogenicity classification for each CNV can be defined by the user, accompanied by interpretations in the form of comments. Each new evaluation also helps populate the laboratory's personal knowledge base.

UCSC. An external link to the UCSC Genome Browser allows users to view the genomic region of the CNV. This page opens in an additional tab of the browser and includes the user's personal settings for that browser, thereby offering a personalised view.

CNV-Hub. An external link to the CNV-Hub classifier lets you visualize CNV classification by taking into account multiple classification tools. The link opens an additional browser tab, and integrates the user's personal browser settings, offering a personalized visualization.

DECIPHER Genome Browser. An external link to the DECIPHER Genome Browser enables you to view CNVs on the genome. The link opens an additional browser tab, and integrates the user's personal browser settings, offering a personalized visualization.

Local IGV. This link redirects the *IGV Desktop*[29] application to the coordinates of the CNV currently being interpreted. To view the CNV, it is necessary to have previously loaded the sample BAM on the IGV Desktop application. This option is only compatible with IGV Desktop versions ≥ 17.

In addition to the alignment files, various files are available for downloading from the *Files* tab in the CNVExome analysis:

- The results of the CNV analysis as a tabular file (cnv.tsv)
- The ploidy for each chromosome (contig_ploidy.tsv)
- The predicted number of copies for each exon in the manifest (denoise.tsv)

[29] https://igv.org/doc/desktop/

CNV data export

The icon  at top right allows you to export a TSV file containing the following data for the CNV concerned:

- its VKB rating and commentary,
- ISCN rating,
- ACMG classification,
- chromosome, start and end positions and cytoband,
- CNV size and type,
- copy number
- copy number
- sample freq
- quality score (Phred scale)
- Log2Ratio
- number of probes involved (Probes)
- number of genes affected by CNV.

Information on the associated gene(s) is described in the following columns:

- gene name (Gene_Symbol),
- OMIM phenotype (OMIM_Phenotypes),
- probability of gene intolerance to LoF (Pli_Score)
- ClinGen haploinsufficiency and triplosensitivity scores (HI_Score, TS_Score)
- the number of exons overlapped by the CNV (Nb_overlap_exon),
- the total number of exons in the gene (Nb_total_exon).

If certain descriptive fields concern only certain CNVs (in the case of MicroArray CGH, for example), these fields have the value "NA".

12.1.6. Large Variant Viewer for CNVs & SVs (LVV)

The Large Variant Viewer (LVV) allows the user to visualize and analyze CNVs and Structural Variants (SVs) in a single tab. The interface consists of a central variant table, a filter panel on the left, and a side panel that opens on the right when a variant is selected, similar to the variants tab (section Small Variants).

In compatible analyses, the viewer is located in a tab labeled "CNVs & SVs" or "SVs", depending on the data type:

- CNVs & SVs tab displays CNVs (Loss/Gain), SVs (Deletion/Duplication), and LOH/UPD events.
- Legacy "CNV" Tab: CNVs are also displayed in the legacy viewer to maintain backward compatibility with existing workflows.
- SVs are only accessible in the "CNVs & SVs" or "SVs" tab

Data availability in the Large Variant Viewer depends on the analysis type:

- WGS: Fully compatible (CNVs, SVs, UPDs).
- WES: Fully compatible (CNVs, SVs, UPDs). CNVs appear only if the CNV-calling module is executed as part of the analysis.
- Panels: Partially compatible (SVs only).
- Tertiary analysis (VCF/aCGH):
    - aCGH: All copy-number events are displayed.
    - VCF: CNVs and SVs (symbolic ALT alleles) are displayed. SVs with specific ALT sequences are displayed in the Small Variant Viewer.

CNVs & SVs results table

Similar to the small variant table, the main table displays one row per event, with the following columns providing variant-specific metrics:

DiagAI. Prioritization score from 1 to 100, adapted for large variants, ranking variants by how likely they are to be disease-causing and diagnostically relevant in the context of the case (especially when HPO terms are provided). The DiagAI Score combines multiple components, including: pathogenicity prediction via UP² score, phenotype matching via PhenoGenius (HPO-driven) and expert/inheritance & quality rules.

ACMG classification indicator based on large variant is shown as a 1–5 classification score in the LVV table. In the details drawer an ACMG tab shows the underlying criteria and a live classification summary from Pathogenic to Benign (mainly for CNV Gain/Loss).

Classification (Flag icon) which allows the user to classify the variant in the VKB.

CNV/SV type. CNV/SV type (gain or loss), size, location, and cytoband,

Regions (exonic, intronic, UTR) and genes overlapping the CNV (HGNC gene names), OMIM presence, fraction of coding genes, and details on overlapping exons. Variants are annotated with overlapping genes and genes located within 5 kb of the breakpoints (5 kb padding).

Genes associated with a variant are sorted by clinical relevance (e.g., OMIM morbid status, overlap percentage, constraint scores). The three most relevant genes are displayed in the main table. The full list is available in the "Genes & Regions" tab of the side panel ("Details Drawer").

RMSK %. Percentage of the event overlapping RepeatMasker annotations (e.g., repetitive regions).

ClinVar & Decipher (Clinical databases). Both columns display the number of matches from the corresponding clinical database. The best match is defined as an event where Database Overlap is > 50%. ClinVar and Decipher classifications are mapped to Benign (B), VUS, or Pathogenic (P).

DGV Gold & GnomAD (Population databases). The table displays the maximum population frequency found for the variant. Events are considered matching if Variant Overlap is > 80% and the event type is equivalent. The Frequencies tab in the side panel lists all overlapping events with > 1% overlap.

Phenotypes. By clicking the numbers displayed in the columns corresponding to the results of the Phenogenius algorithm, the user can access:

- the match percentage, which is the highest match score between a single affected gene and the patient's HPO terms.
- the match fraction, which is the number of patient HPO terms that match at least one affected gene.
- For family analysis, matching is performed using the proband's HPO terms.

SampleFreq (Internal sample frequency). The frequency of the variant within the current project and entity is displayed in the corresponding columns.

- Matches are based on > 80% Variant Overlap and identical event type.
- Update Frequency: data is based on a snapshot of analyses performed prior to the current analysis launch. A manual "Refresh" button is available in the Details Drawer to update this data.

Scores columns. The user can display multiple scores:

- UP²: pathogenicity prediction score that is a component of DiagAI.
- pLI, LOEUF: Loss-of-Function intolerance.
- pTriplo, pHAPLO: Triplosensitivity and Haploinsufficiency.
- HI ClinGen, TS ClinGen.

Sample columns. These columns display event metrics specific to the data source:

- For family analyses (excluding aCGH), sample metrics are displayed for all family members.
- Genotype: for CNVs where a genotype is not provided by the caller, the system infers the genotype based on absolute copy number relative to ploidy:
    - Homozygous (hmz) with full dark dot
    - Heterozygous (htz) with half-dark dot
- VAF / #Reads: Variant Allele Frequency and supporting reads (available for SVs).
- #Copy: absolute copy number (available for CNVs).
- LogRatio / Probes: available for aCGH data.
- QUAL: quality score from the variant caller or input file.

Sorting, Filtering, Viewing, and Interpretation

Filtering. The filter panel allows the user to filter variants based on column data, similar to the Small Variants tab filtering system (see the corresponding section).

- Gene Panel: filters for events overlapping genes in a specific gene list.
- Quality: filters based on the QUAL column. By default, events with no quality score (N/A) are hidden unless "Keep unknown" is selected.
- Presets: filter configurations can be saved as presets for future use.
- For family analyses, the Quality filter uses "OR" logic. The variant is displayed if at least one family member meets the quality threshold.

Sorting. By default, the table is sorted by DiagAI Score (descending). The user can sort by UP2, Size, Location, or Sample Freq. Entity, in ascending or descending order.

As in the legacy CNV tab, the user can open the ploidy windows and the CNV browser at the top of the table (see section **CNV: View and interpretation**). A column management icon is also available at the far right of the header.

CNV/SV side panel (Details Drawer)

Similar to the variant page (Small Variants section), selecting a variant in the Large Variant Viewer opens a detailed page on the right side of the screen. This page contains multiple tabs:

Variants. This summary tab includes variant nomenclature, the DIAGAI score, the list of suspected compound heterozygotes (Suspected Het Compound), gene and region information, and links to databases.

ACMG. Detailed classification.

Genes & Regions. The full list of genes and regions associated with the variant, with detailed clinical information such as OMIM morbid status, overlap percentage, constraint scores, and phenotype match.

Classification. Classify the current variant in your VKB at the top of the side panel, and save it. The current classification and occurrences of this variant in other analyses can be found on this page.

Frequencies. The user can access details on variant frequencies in the DGV Gold and GnomAD databases, as well as for the entity and the current project.

Clinical databases. ClinVar & Decipher.

Viewer. The user can choose to view the variant in the SeqOne viewer or IGV. In the IGV sub-tab, the user can view the “Start” or “End” of the event, or “Start to End”. In the SeqOne sub-tab, the user can access the Genome Browser and visualize the event in one of the multiple layers available: CNV, SV, UPD, Gene, CN (copy number), BAF, DGV.

12.1.7. Signatures in SomaCGP analysis

Methodology for calculating MSI score

MSI analysis is an individual analysis performed against a standardized reference. This reference consists of a set of previously selected and pooled negative samples. The tumor sample is compared with this reference using the *MSIsensor* tool. A score is then calculated, represented by the percentage of unstable microsatellite regions covered by the panel.

The panel of normals currently supported by SomaCGP corresponds to the Agilent SureSelect Cancer CGP panel. The use of another kit or a custom design derived from the CGP will require the creation of a dedicated panel of normals.

Methodology for calculating TMB score

TMB is calculated in three steps:

1. Calculation of the size of coding regions included in the panel
2. Identification of somatic variants
3. Calculation of the number of somatic mutations per megabase (mut/Mb)

The identification of somatic variants is based on a set of filters described below:

- Variant type and effect filters
    - SNVs only (MNVs are excluded)
    - Effects retained: stop codon gain, reading frame shift, splice variant (acceptor and donor), transcript ablation, in-phase deletion/insertion, missense variant, synonymous variant.

Synonymous variants are excluded from the calculation of the non-synonymous TMB score.

- Variant quality filters
    - Variants detected by Mutect2 that pass FilterMutectCalls filters
    - Depth of coverage (DP) ≥ 50X
    - Quality score > 30
    - Allelic Support Ratio (ASR) > 0.1
- Somatic variant filters
    - Variant allele frequency (VAF) between 5% and 40%
    - General population frequency (GnomAD) ≤ 0.0001

<u>Risks of Using Synthetic TMB Control Samples</u>

- Unnatural mutation distribution:
    - Engineered mutations might not reflect the genomic context of real tumors (e.g., mutation clustering, strand bias, sequence context).
- Lack of tumor heterogeneity:
    - Synthetic samples don't mimic clonal diversity or tumor purity, which affects variant allele frequency (VAF) distribution.
    - No background noise or FFPE artifacts: real samples have sequencing artifacts, FFPE damage, and copy number variations, which are absent in synthetic materials.
- Different fragment size and chemistry behavior:
    - Synthetic constructs often behave differently during library prep, hybrid capture, and sequencing.

Visualization and interpretation

The "*Signatures"* tab displays the results of the MSI and TMB analyses in 2 sections:

Microsatellite Instability. This view shows information such as :

- the number of microsatellite regions with a coverage greater than or equal to 50X (Sites above 50X coverage),
- the number of somatic sites,
- MSI score in percentage and 95% confidence interval (CI),
- MSI Status

Tumor Mutational Burden. This view shows the following information:

- the size of the coding region expressed in Mb (Coding region size),
- the number of somatic and non-synonymous variants passing filters (Somatic variants passing filters and Somatic non-synonymous variants passing filters),
- the TMB score of all somatic variants passing the filters in mut/Mb (TMB Score),
- the TMB score of non-synonymous variants passing mut/Mb filters.

Each result can be downloaded by clicking on the icon  at the top right of each section as a tsv file.

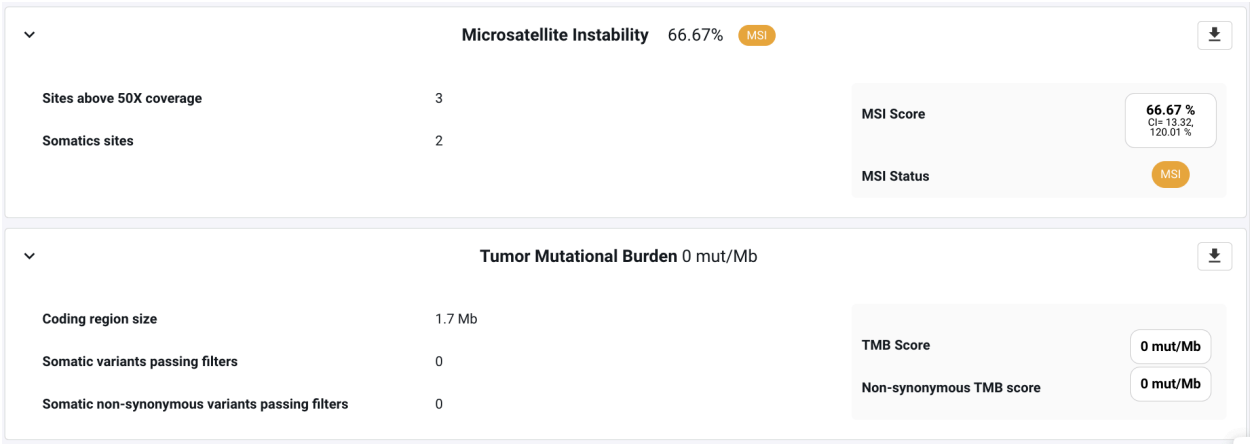


Figure 29: Viewing the “Signatures” tab of a SomaCGP analysis.

12.1.8. Short tandem repetitions (STRs)

Visualization of Short Tandem Repeats (STRs) is available for GermVar Tertiary and GermVar Tertiary Family analysis worksets dedicated to the analysis of Whole Genome Sequencing (WGS) data, when generated from Oxford Nanopore Technologies sequencing.

Short Tandem Repeat (STR) detection is also available in the GermVar and GermVar Family analysis worksets for WGS from FASTQ data generated with Illumina, Element Biosciences, MGI, and Oxford Nanopore Technologies.

Methodology

Short tandem repetitions from FASTQ

Short tandem repeat (STR) data are detected using ExpansionHunter, which processes the BAM file alongside a reference genome and a variant catalog defining the target loci. The resulting VCF is then annotated by Stranger, which classifies each locus as Normal, Pre-mutation, or Pathogenic based on the observed repeat count and known disease thresholds. Both tools are integrated directly into the GermVar and GermVar Family pipelines.

ExpansionHunter: https://github.com/Illumina/ExpansionHunter
Stranger: https://github.com/moonso/stranger

Short tandem repetitions from VCF

Tandem repeat data are extracted from the VCF file (sample.wf_str.vcf.gz) generated from the wf-human-variation v2.3.0 workflow of the EPI2ME application (Oxford Nanopore Technologies) and supplied by the user. As these data have been annotated for each STR-bearing locus, from a catalog of variations maintained by EPI2ME, our pipeline does not modify the annotations. The premutated or pathogenic status of each allele is determined according to the number of sequenced repeats and known premutation and pathogenicity thresholds.

View and interpretation

The STRs tab allows you to view Short Tandem Repeats in the form of a table displaying the following details for each genotyped locus:

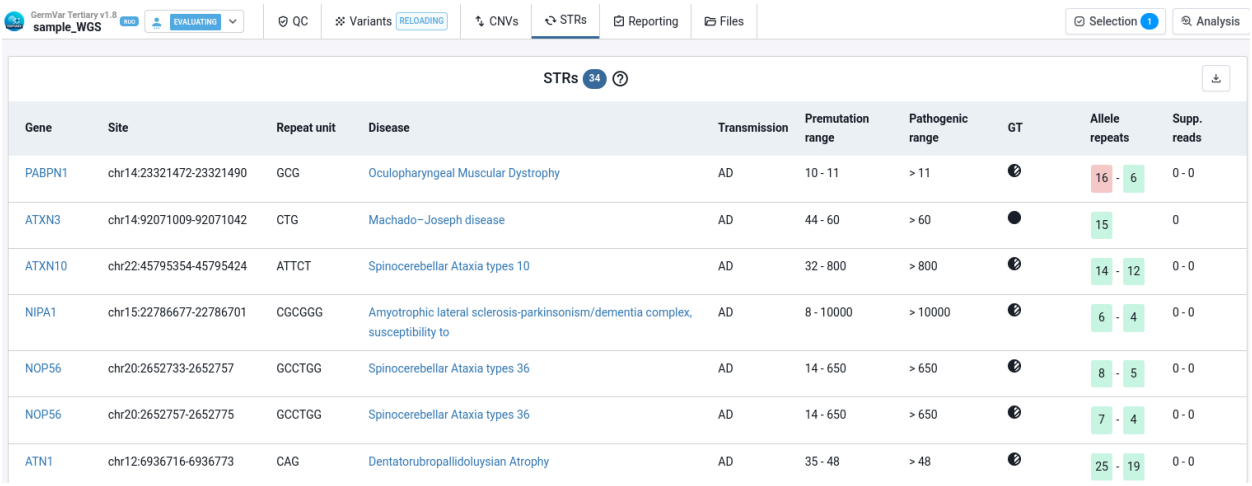
Locus information and annotations

- *Gene*: gene name (HGNC) and corresponding link.
- *Site*: the genomic coordinates of the repeat.
- *Repeat Unit*: the repeated nucleotide motif.

- *Disease*: the associated disease according to the OMIM database and the corresponding link.
- *Transmission:* the known mode of transmission.
- *Premutation range*: the premutation range corresponding to the minimum and maximum number of repeats determining premutated status.
- *Pathogenic range*: the lower threshold of repeats determining pathogenic status.

Details specific to each sample

- *GT*: the homozygous, biallelic or heterozygous genotype of the locus.
- *Allele repeats*: the number of repeats per allele. Homozygous loci carry a single value. For biallelic and heterozygous loci, the number of repeats for both alleles is indicated. The status of each allele is determined by the number of repeats: wild type in green, premutated in orange and pathogenic in red.
- *Supp. reads*: the number of supporting reads for each allele. Homozygous loci carry a single value, while values for 2 alleles are shown for biallelic and heterozygous loci.



Gene	Site	Repeat unit	Disease	Transmission	Premutation range	Pathogenic range	GT	Allele repeats	Supp. reads
PABPN1	chr14:23321472-23321490	GCG	Oculopharyngeal Muscular Dystrophy	AD	10 - 11	> 11		16 - 6	0 - 0
ATXN3	chr14:92071009-92071042	CTG	Machado–Joseph disease	AD	44 - 60	> 60		15	0
ATXN10	chr22:45795354-45795424	ATTCT	Spinocerebellar Ataxia types 10	AD	32 - 800	> 800		14 - 12	0 - 0
NIPA1	chr15:22786677-22786701	CGCGGG	Amyotrophic lateral sclerosis-parkinsonism/dementia complex, susceptibility to	AD	8 - 10000	> 10000		6 - 4	0 - 0
NOP56	chr20:2652733-2652757	GCCTGG	Spinocerebellar Ataxia types 36	AD	14 - 650	> 650		8 - 5	0 - 0
NOP56	chr20:2652757-2652775	GCCTGG	Spinocerebellar Ataxia types 36	AD	14 - 650	> 650		7 - 4	0 - 0
ATN1	chr12:6936716-6936773	CAG	Dentatorubropallidoluysian Atrophy	AD	35 - 48	> 48		25 - 19	0 - 0

Figure 30: Viewing the STRs tab of a GermVar Tertiary analysis.

The table can be downloaded as a repeats.tsv file via the icon at the top right of the table. This file is also available from the Files tab of the analysis.

12.1.9. Fusion genes

12.1.9.1. Fusions detected from DNA

Methodology

GRIDSS (Genome Rearrangement IDentification Software Suite) module, specialized in the detection of genomic rearrangements, and implemented in the SomaVar and SomaVar LF workset, is used to detect gene fusions from DNA sequencing, enriched by capture targeting a panel of genes

View and interpretation

The Fusions tab of a somatic analysis (SomaVar, SomaVar LF, SomaCGP, SomaVar Tertiary) displays the list of fusions passing the various GRIDSS filters in the form of a table detailing information about the fusion and each partner involved:

Fusion information

- fusion with the gene symbols of each partner,
- the type of event (translocation, duplication, inversion or deletion),
- the occurrence of the fusion in COSMIC database,
- the allele frequency (VAF),
- the number of reads supporting the fusion,
- total coverage,
- filters assigned by GRIDSS.

Partner-related information

- the genomic coordinates of the point of fusion (breakend),
- the involved RefSeq transcript,
- the site corresponding to the point of fusion (breakpoint): 5'UTR, 3'UTR, Intron, Exon (non-coding breakpoint), CDS (coding breakpoint) or Splice site,
- the numbering of the exon or intron involved in the fusion, relative to the total number of exons or introns.

A menu of filters allows you to

- search for a specific fusion or a gene involved in a fusion via a search bar,
- display fusions that have not passed GRIDSS filters (LOW_QUAL) by checking the 'Show filtered fusions' box. This box is unchecked by default.

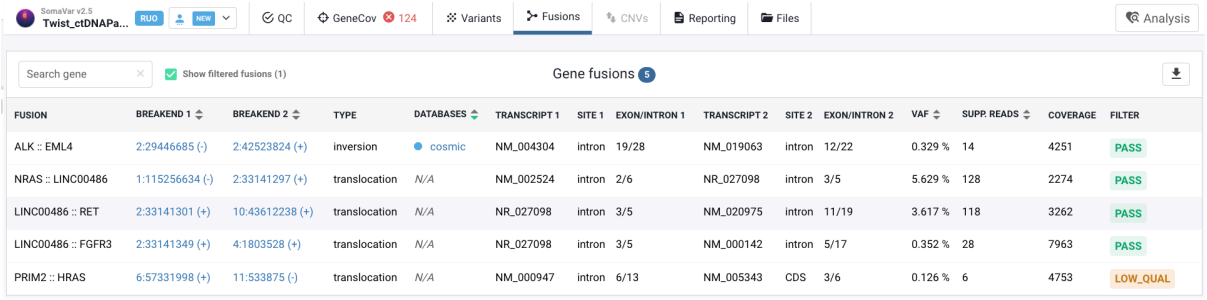


Figure 31: Visualization of the Fusions tab for a SomaVar or SomaVar LF analysis.

The data in the table can be sorted according to genomic coordinates (breakend), the presence of the fusion in COSMIC (Databases), allelic frequency (VAF) and the number of reads supporting the fusion (Supporting reads).

The link associated with the genomic coordinates of each breakpoint opens a split view of the fusion point on IGV. Tooltips detail the number of split reads and mismatched pairs in the Supporting reads column, and the coverage at each breakpoint (Cov. 1 and Cov. 2) in the Coverage column.

The Filter column displays 3 quality filters used by GRIDSS and giving an indication of confidence in the fusion call:

- PASS: the fusion has passed the quality filters. The quality score is greater than or equal to 500 and is considered a medium or high confidence call (if > 1000).
- LOW_QUAL: the fusion has a quality score below 500 for at least one of the two breakends involved in the fusion, and is considered a low-confidence call.
- ASSEMBLY_ONLY: the fusion is called on the basis of an assembly and not on the basis of the read supporting it. Assembly refers to the reconstruction of sequences around potential fusion points, which are then re-aligned to identify the fusion.

If several filters are available for the same fusion, a "+" sign displays all filters in the Filter column.

In addition to the results presented in the interface, a fusions.tsv file detailing the fusions detected is available for download from the Files tab or from the icon  at the top right of the page. It contains the following information:

- fusion name with gene symbols for each partner (fusion)
- genomic coordinates of each fusion point (breakend_1 and breakend_2)
- canonical transcript for each partner (transcript_canonical_1 and 2)

- site corresponding to the fusion point for each partner (site_1 and site_2)
- exon or intron numbering according to the associated RefSeq transcript (exon_1 and 2)
- 60bp of flanking sequences surrounding the fusion point (flanking_region_1 and 2)
- quality filters assigned by GRIDSS (filter)
- event type (type)
- number of split reads (split_reads)
- number of discordant read pairs (discordant_pairs)
- coverage for each partner (coverage_1 and coverage_2)
- fraction of reads supporting the fusion (read_fraction)
- number of reads supporting the fusion (supporting_reads)
- total coverage (coverage)
- COSMIC fusion identifier (cosmic_id)

This file contains:

- Fusions displayed in the interface, having passed the quality filters (filter = PASS),
- Fusions detected but with a quality score below 500 for at least one of the two breakends involved in the fusion (filter = LOW_QUAL).

The quality score for each breakend can be found in the gridss.raw.vcf.gz file (QUAL column), downloadable from the Files tab.

12.1.9.2. Fusions detected from RNA

SomaRNA tool for analysing RNA-seq data from either targeted approaches such as capture and amplicon, meaning that users can identify and view gene fusion events at the transcript level.

Methodology

A combination of two tools is used to detect and view fusions by SomaRNA:

- STAR: an RNA-seq data aligner designed for aligning non-contiguous sequences directly to the reference genome
- Arriba: detects gene fusions from chimeric alignments

Based on the chimeric alignments from STAR, Arriba applies a set of filters to eliminate known artifacts and the transcripts observed in non-pathological contexts. It assigns a confidence score to each potential fusion event that depends on several criteria:

- The number of sequences supporting the fusion
- The balance between split reads and discordant mates
- The distance between breakpoints, their proximity (intragenic or otherwise)

- The type of event

Fusions detected with Medium or High Confidence by Arriba are displayed on SeqOne. Those with Low Confidence are listed in the "arriba_discarded" files (see below). Arriba's confidence reflects the likelihood of an aberrant transcript not observed in healthy tissue, the possibility of an underlying genomic rearrangement, and the exclusion of artifacts.

More details on the confidence level in Arriba's documentation: https://github.com/suhrig/arriba/blob/master/documentation/08-Interpretation-of-results.md

To understand why a fusion is "Low confidence", consult the 'filters' (AA) column in the "discarded" file. It indicates the filter that discarded the event. Details of each filter: https://github.com/suhrig/arriba/blob/master/documentation/11-Internal-algorithm.md

View and interpretation

From the Fusions tab of a SomaRNA analysis, the user obtains a list of fusion predictions, in the form of different tabs.

The top part of the page gives details of the fusion and each of the partners involved:

General information about the fusion

- the number of spanning junction reads, equal to the sum of the respective splits reads of each of the 5' and 3' genes involved in the fusion,
- the number of discordant mates,
- the confidence score assigned by Arriba (High, Medium or Low),
- the type of event leading to the fusion (translocation, duplication, inversion or deletion),
- the impact of the fusion on the reading frame,
- where applicable, the occurrence of the fusion in the COSMIC database.

Information about each fusion partner

- the symbol for each gene,
- the Refseq transcript or, failing that, ENSEMBL for each gene,
- the numbering of the exons involved in the fusion,
- the genomic coordinates of each breakpoint,
- the total coverage at the fusion point (Coverage),
- the read fraction supporting the fusion.

Split reads are reads aligned predominantly in gene 1 or gene 2 and overlapping the fusion. In the case of discordant pairs of reads, one of the reads is on gene 1 and the other on gene 2, without necessarily overlapping the fusion.

The read fraction is the sum of the split reads and discordant read pairs associated with a gene, divided by the total coverage. The read fraction value can be greater than 100% when the split reads associated with a gene are also part of discordant read pairs. These reads are then counted twice instead of once.

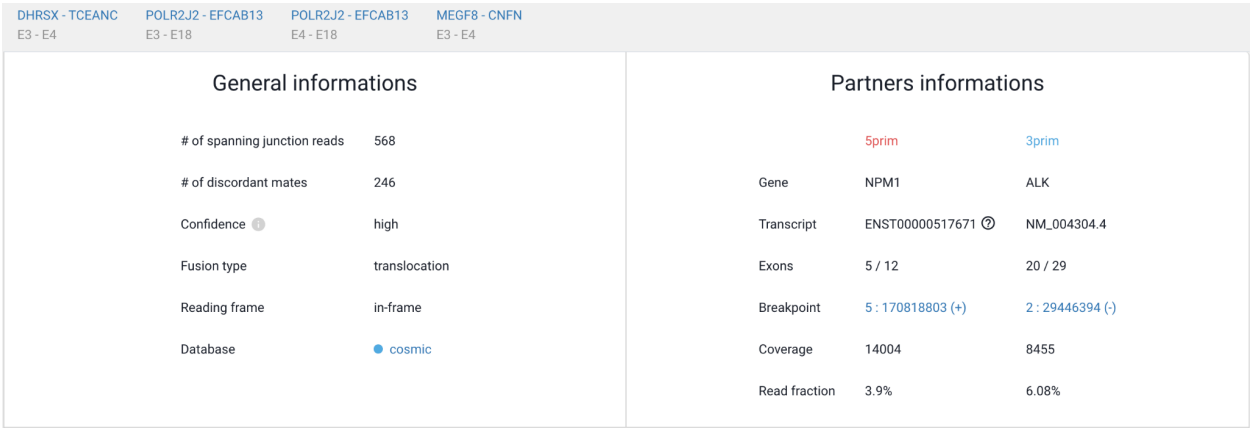


Figure 32: Visualization of fusion information from a SomaRNA analysis.

The lower part of the page shows graphical representations of the fusion, both at gene and exon level, and at protein level. This last section is divided into two tabs:

- The first tab presents the structure of the fusion and the partners involved, where the coverage of the exons covered by the analysis manifest is represented as a diagram on a yellow background:

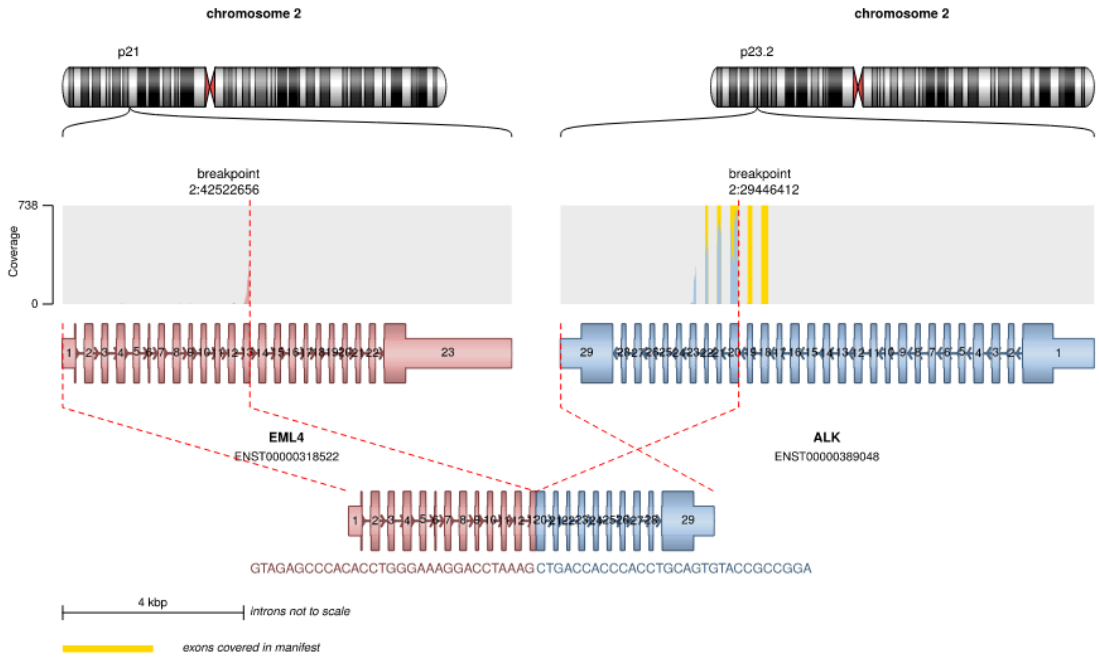


Figure 33: View of a fusion event between two genes

- The retained protein domains following the fusion event are highlighted in a second view:

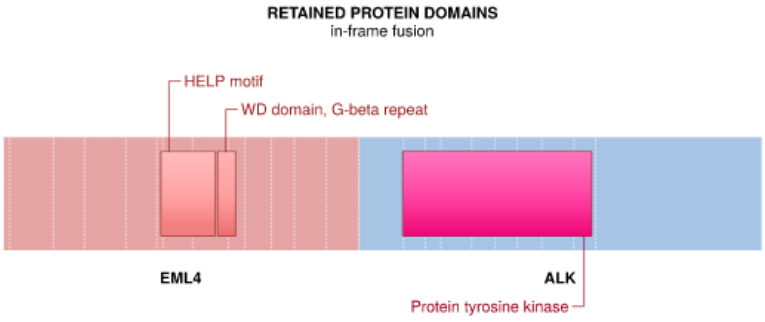


Figure 34: View of the polypeptide resulting from a fusion between two genes

In addition to the results presented in the interface, various files are available for downloading:

- The fusions retained by Arriba in TSV format (arriba.tsv file),
- The fusions eliminated by the Arriba filters in TSV format (file arriba_discarded_standard.tsv),
- the fusions eliminated by the Arriba filters and listed in COSMIC in TSV format (arriba_discarded_cosmic.tsv file),
- the different figures displayed in the interface in SVG format

12.1.10. Clonality

The 'IGHV mutational status report' in the "Clonality" tab of SomaHemato analysis shows mutational status results.

The 'IGHV Results' section displays:

- IGHV Mutational status and percentage of homology:
    - Mutated: if homology is <98% identity to germline
    - Unmutated: if homology is ≥98% identity to germline
    - Borderline: for clones in the 97–98% identity range
    - Polyclonal: no major clone above 20%
    - Non-Conclusive: if any essential QC fails
- The determined subset (with clinical impact) and confidence: *2 (High risk), *8, or None
- A warning notification if warnings appear in the report.

The "Clonality" section lists identified clones with their frequencies and productivity status (Productive, Un-productive).

Each clone has a dedicated section showing its mutational status, frequency, IMGT_quest page link, and additional details (homology, completeness) along with quality control checks.

12.1.11. Splicing variants

Methodology

The SomaRNA workset uses the *Regtools* suite of tools to detect exon jumps and identify alternative splice variants from RNA sequencing, enriched by capture targeting a panel of genes. RegTools extracts exon-exon junctions from the BAM file generated by the STAR RNA-seq data aligner, and annotates them according to a known transcript structure based on a GTF file and Ensembl annotations.

Visualization and interpretation

The Splice variants tab of a SomaRNA analysis is focused on visualizing the splice variants of the following clinically relevant genes and their major actionable variant:

- MET (MET Proto-Oncogene, Receptor Tyrosine Kinase) and the actionable variant METΔ14
- EGFR (Epidermal Growth Factor Receptor) and the actionable splice variant EGFRvIII
- AR (Androgen Receptor) and the actionable splice variant AR-V7

A complete list of splice variants detected in the SomaRNA analysis, in addition to those listed above, can be downloaded directly as a TSV file via the Download icon in the interface or from the Files tab in the analysis.

For each gene mentioned above, the detection status (Detected or Not detected) of the main actionable variant is displayed, along with the number of other alternative events detected for that gene.

Users can view details of the main splice variants detected by opening a dedicated insert, which is divided into three parts:

- On the left, information about the main actionable splice variant is shown: genomic coordinates, gene name, RefSeq transcript, type of variant detected detailing missing exons, read phase, fraction of reads and number of reads supporting the variant. Finally, the average canonical support reads are specified. This corresponds to the average of the 5' and 3' reads supporting the inclusion of the exon.

- on the right, a sashimi plot (Junction Plot) illustrates the junctions detected between exons, as well as the number of reads supporting each junction.

- at the bottom, information about the alternative events detected is listed, i.e. isoforms other than the main variant of interest described above: type of variant detected

(missing exon numbers in brackets), reading phase, fraction and number of reads supporting the variant, average number of reads supporting the canonical transcript (Avg. canonical support), PASS or FAILED status according to the filters described below.

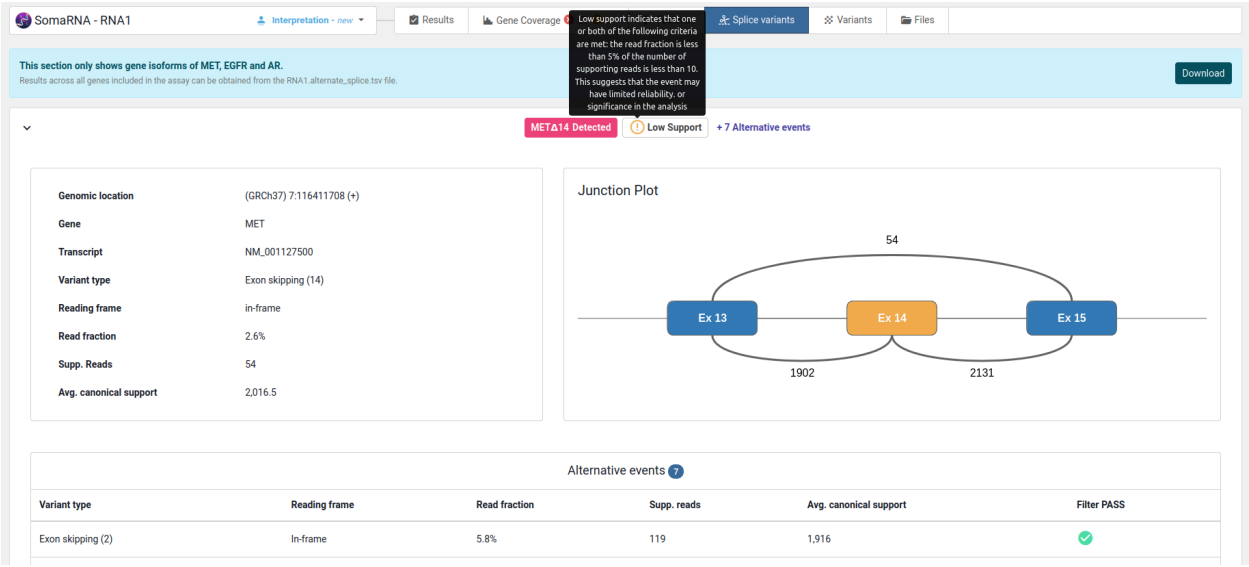


Figure 35: *Visualization of the Splice variants tab of a SomaRNA analysis.*

Detection of splice variants in the genes displayed in the Splice variants tab is not subject to the application of detection filters.

Nevertheless, a "Low support" icon associated with METΔ14 and EGFRvIII detection indicates to the user that the metrics related to supporting reads are below the following thresholds: fraction of reads < 5% or number of supporting reads < 10.

The PASS or FAILED status of other alternative events also uses these thresholds:

- *PASS :* fraction of reads >= 5% and number of reads >= 10.
- *FAILED :* fraction of reads < 5% or number of reads < 10.

The downloadable TSV file details the following information for all splicing events detected: RefSeq transcript, gene, strand direction (+ or -), fraction of reads and number of reads supporting the variant, number of missing exons and corresponding exon numbers, reading phase, genomic coordinates of junctions. The file also specifies under FILTER_PASS (yes/no) whether the variant passes the thresholds applied to supporting reads (fraction of reads >= 5% and number of reads >= 10) and under Actionable (yes/no), the actionability of the variant.

12.1.12. MSI/MSS state

The SomaMSI analysis workset is available when running a new analysis on gene panel data and is used to determine the MSI or MSS state of a tumour sample.

Methodology

This non-individual cohort analysis is similar to a CNVCapture or SomaCNVCapture analysis. At least eight samples are required to run the analysis.

The tool generates a reference sample from the samples in the cohort that are estimated to be negative. Since this state for the samples is not known in advance, this estimate is based on calculating the rate of deletions seen in the microsatellite regions: half the samples in the cohort featuring the lowest rate are therefore selected to generate a reference sample, with each sample thereby contributing a fraction of its sequences.

Each tumour sample is subsequently compared to the reference sample using the MSIsensor tool. A score is calculated, which represents the percentage of microsatellite regions studied that present instability. The regions studied are the microsatellite regions covered by the panel and which are detected by MSIsensor.

Based on this score, the MSI or MSS state of the different samples is dictated by a threshold that is calculated as follows: max(10, median + max (3*median-abs-dev, 0.3*median))
The principle behind this score is as follows:

- In an analysis presenting MSI and MSS samples, the threshold will generally be median + 3*median-abs-dev.
- However, if all the samples have the MSS state, if the scores are very low and if their distribution is extremely close, the threshold will be set to 10 to avoid generating false positives.

View and interpretation

Once the analysis has been completed, the results for all the samples can be viewed from the SomaMSI analysis. The Results tab is divided into two sections:

MSI graph: a graphic representation of the results where each sample is shown according to its score. The threshold determined by the sample cohort creates two zones in the graph representing the samples considered to be MSI and MSS respectively.

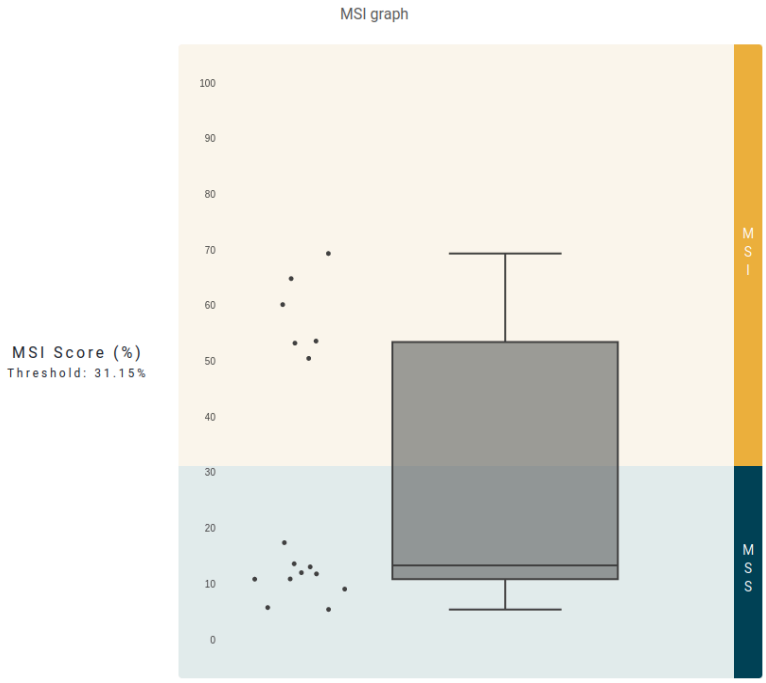


Figure 36: Box plot representing the distribution of the MSI scores for the sample cohort.

Per sample details: the breakdown of the scores and the state attributed to each sample are available in the right panel.

Figure 37: Breakdown of the scores and state of all the samples included in the SomaMSI analysis.

12.1.13. HRD/HRP status

Methodology

The detection of HRD status proposed by SeqOne is based on the analysis of low coverage genome sequencing (sWGS) data, to which can be added the results of a gene panel analysis, if the latter includes the BRCA1 and BRCA2 genes.

The principle is as follow:

- Quantify genomic instability from sWGS data and identify large rearrangements (LGA) as well as losses of parental copies (LPC), which would reflect a loss of heterozygosity.
- Identify copy number variations (CNVs) in key genes described as potentially affecting HRD status.
- Optionally detect pathogenic variants in BRCA1, BRCA2 and TP53 genes, if panel results are available on SeqOne interface for the corresponding patient. In this case, the types of variants detected by the SomaVar workset that are taken into account are SNV, indels and SV (detected by Freebayes, Mutect2, and GRIDSS).

HRD status. HRD positive status is defined by at least one of the following (**SomaHRD v1.2 and v1.3 versions only**): (1) at least one pathogenic BRCA1/2 mutation and/or (2) an HRD probability greater than or equal to 0.5.

From version v1.4 of the SomaHRD workset, the status of the analysis is defined solely by the genomic instability score (GI score). Users can select an HRD status (Positive, Negative, or Non-conclusive) next to the GI status to include it in the report instead of the GI status.

HRD probability. (**SomaHRD v1.2 and v1.3 versions only**) The SeqOne HRD Probability corresponds to the probability that the tumor is HRD positive on the basis of LGA and LPC genomic instability characteristics determined from the sWGS data. Large Genomic Alterations (LGA) are defined as the copy number of breakpoints, where a breakpoint corresponds to a change in copy number between two genomic segments at least 10 Mb long and no more than 3 Mb apart. Losses of Parental Copy (LPC) are defined as the number of haploid segments at least 10 Mb long. The HRD probability is a linear combination of the two genomic instability characteristics, converted using logistic regression to place the values on a probability scale between 0 and 1.

GI probability (GiP). (**from SomaHRD v1.4**) The probability of genomic instability (GI) represents the likelihood of the tumor exhibiting genomic instability, based on shallow whole genome sequencing (sWGS). The calculation of this probability depends on the two types of genomic instability characteristics LGA and LPC. LGA is defined as the number of copy number breakpoints. A breakpoint corresponds to a change in copy number between two sufficiently large genomic segments in close proximity. The LGA score is not a single measure, but rather a set of several variations. These variations are calculated using different parameters for

minimum fragment size and distance between fragments. The LPC score, on the other hand, refers to the number of haploid segments at least 10 Mb long. The SeqOne genomic instability probability is not a direct function of the average LGA and LPC scores. It is calculated using a boosting model that integrates several LGA and LPC features, thus improving prediction accuracy. The final probability is the mean of 100 probabilities (ranging from 0 to 1) calculated by applying different model parameters.

GI status. (**from SomaHRD v1.4**) A positive genomic instability status (GI status) is defined by a probability of genomic instability greater than 50%.

BRCA Status. For the detection of pathogenic variants, determining BRCA status independently of HRD probability or GI probability, the following criteria are applied to BRCA1 and BRCA2 variants from the patient's SomaVar analysis:

- Allele frequency ≥ 15% (**SomaHRD v1.2 and v1.3 versions**),
- Allele frequency ≥ 10% (**from SomaHRD v1.4**)
- Coverage ≥ 35
- Average quality of each base carrying the variant (Phred score) ≥ 20
- Alternate observation (number of sequences carrying the mutation) ≥ 5
- Classification 4 (*likely pathogenic*) or 5 (*pathogenic*) according to our variant pathogenicity classification tool based on ACMG criteria.

TP53 variants. The following criteria are applied to TP53 pathogenic variants from the SomaVar analysis conducted on the patient:

- Coverage ≥ 100
- Average quality of each base carrying the variant (Phred score) ≥ 20
- Alternate observation (number of sequences carrying the mutation) ≥ 5
- Classification 4 (likely pathogenic) or 5 (pathogenic) according to our variant pathogenicity classification tool based on ACMG criteria.

These variants are only displayed without a minimum allele frequency threshold in the SomaHRD analysis report, but are not determinants of HRD status. Only the TP53 pathogenic variant with the highest allele frequency will be reported in the HRD report.

Cellularity. The cellularity metadata (% of tumoral cells in the sample) is required to run the SomaHRD analysis. It can be entered when the sample is loaded or before the analysis is launched, by editing the corresponding sample metadata (see **Sample metadata**). If the value is less than 20%, the result will be non-conclusive. However, the HRD status may be positive if the analysis is accompanied by a panel including BRCA1 and BRCA2, which gives a positive BRCA status.

Tumor Fraction Warning Procedure: SomaHRD analysis includes an alert mechanism to indicate when a sample may have low tumor content. This is based on a statistical measure called the "overlap statistic," which evaluates the similarity between consecutive genomic segments in

terms of copy number ratio. Samples with significant overlap often have lower tumor fractions, making it more challenging to distinguish genomic instability.

To establish a reliable threshold for this warning, we correlated the overlap statistic with the Variant Allele Frequency (VAF) of TP53 mutations using clinical data. An alert is triggered if the overlap statistic exceeds a predefined threshold. This alert highlights potential discrepancies with the manually provided tumor content percentage, which may indicate that the tumor fraction is below 20%. Samples with tumor fractions below this level are at an elevated risk of producing false negatives. However, such samples are not rejected, and the alert serves as guidance for careful interpretation of HRD results.

Minimum coverage. The minimum coverage required to run the test is 0.1X, but we recommend a minimum coverage of 0.5X, ideally 1X. When coverage is reduced to 0.1X, the classification error rate increases by 5%.

Although sWGS data may have been generated from a kit that has UMIs, these are not taken into account during the SomaHRD analysis. A trimming step for both adapters and UMIs is included in the workset.

Visualization and interpretation

Once the analysis is complete, the results are available from the SomaHRD analysis. In addition to QC metrics and the technical report, an HRD report is directly accessible from the HRD Report tab (**SomaHRD v1.2 and v1.3 versions**).

From **version v1.4 of the SomaHRD workset**, the results of the analysis are displayed in a dedicated HRD tab. Each section of the tab reflects a section of the HRD report described below (HRD Summary, Sample quality, Genomic instability features, BRCA1/BRCA2 & TP53 mutations, Copy number variations).

From this tab, the user can export these results as a CSV file or as an HRD report in PDF format. In the case of a PDF export, the user can select the language of the HRD report from a list of available languages: French, Spanish, English, German, Italian. The default selection is English.

The HRD report is divided into the following sections.

A **header** lists patient and analysis metadata, the type of test performed and the HRD status (**SomaHRD v1.2 & v1.3)** or the GI Status (**SomaHRD ≥ v1.4)**.

Sample quality. This section summarizes the main quality metrics of the analysis as well as the minimum thresholds required.
QC controls are displayed as a table. For each parameter, the following information is available:

- ***Name*** : the parameter name
- ***Value*** : the value of the parameter, including :
    - An icon that indicates whether the filter is *Pass*, *Warning*, or *Failed*
    - The parameter value for the sample
- ***Description*** : A short description of the parameter

SAMPLE QUALITY

NAME	VALUE	DESCRIPTION
% of tumoral cells	90%	Low value of tumoral cells is likely to generate false negative results
Coverage	0.43X	Average sequencing coverage of the genome
% correct mapping	96.5%	Proportion of reads correctly mapped
Confidence in genomic instability	1.00	Large values indicates **high confidence** in genomic instability computation
Low tumor fraction	NORMAL	Low tumor fraction detected from the shallow WGS

SAMPLE QUALITY

NAME	VALUE	DESCRIPTION
% of tumoral cells	30%	A percentage of tumoral cells below 20% is likely to generate false negative results.
Coverage	0.7X	The average sequencing coverage of the genome should be above 0.1x.
GI Dispersion	0.93	A large dispersion reflects high variability in the genomic instability computation, indicating low precision in the estimated score.
Low tumor profile	WARNING	Low cellularity detected from the genomic profile.

Figure 38 : Summary of QC metrics related to the sWGS sample in SomaHRD v1.2 (top) and SomaHRD v1.5 (bottom).

Quality checks are applied to the following parameters, according to the thresholds presented at the end of the HRD report (**SAMPLE QUALITY CHECKS** section):

- the percentage of tumor cells, based solely on the value supplied by the user.
- average genome sequencing coverage,
- proportion of correctly aligned reads, calculated after the adapter trimming step (**SomaHRD v1.2 and v1.3 versions only**),
- the metric of confidence in genomic instability (LGA and LPC) based on coverage data (**SomaHRD v1.2 version only**). This quality control is used to measure the confidence of the HRD status returned. Some positive and negative results are marked with a "low confidence" label to identify samples with the lowest confidence in the status.
- a HRD probability confidence interval *(HRD probability confidence interval)* (***SomaHRD v1.3 version only***). This metric will have a *Warning* status when the standard error

(*standard deviation*) is greater than 0.07; it will have a *Failed* status if the HRD probability confidence interval includes the value 0.5.

- a GI probability confidence interval (GI probability confidence interval) (***SomaHRD v1.4 version only***). This metric will have a *Warning* status when the standard error (*standard deviation*) is greater than 0.07; it will have a *Failed* status if the GI probability confidence interval includes the value 0.5.
- a GI dispersion (**SomaHRD v1.5 version only**) that shows the dispersion of the 100 GI probability calculations. This metric will display a "Warning" status with a "Moderate" tag next to the value when dispersion is > 2.6 and a "Failed" status with a "High" tag if > 2.7. Otherwise, a "Low" tag appears beside the score.
- in case of suspicion of low tumor content (*Low tumoral fraction*) based on sWGS, the message *WARNING* is displayed. If this is not the case, the control displays the message *NORMAL*.

If any of the following QC metrics are below the expected failure thresholds, the analysis result will be inconclusive (NON CONCLUSIVE):

- % tumor cells
- coverage
- % reads correctly aligned (**only for SomaHRD v1.2 and v1.3 versions**)
- GI dispersion > 2.7 **if also** 0.37 < GiP < 0.59 (**only for SomaHRD v1.5)**

In this case, the metrics for confidence in genomic instability and low tumor fraction are not calculated and display N/A in their respective sections (for **SomaHRD version < v1.5**).

For SomaHRD v1.3, if the confidence in genomic instability metric is Failed, i.e. below the expected failure threshold, the result of the analysis will be:

- POSITIVE LOW CONFIDENCE if the HRD probability is greater than or equal to 50%.
- NEGATIVE LOW CONFIDENCE if the HRD probability is less than 50%.

In this case, the analysis result is accompanied by a warning.

Nevertheless, if the analysis is accompanied by a panel including BRCA1 and BRCA2, and the panel returns a positive BRCA status, the HRD status will be positive, even if a QC metric relating to the sWGS sample is below the failure threshold (**only for SomaHRD v1.2 and v1.3 versions**).

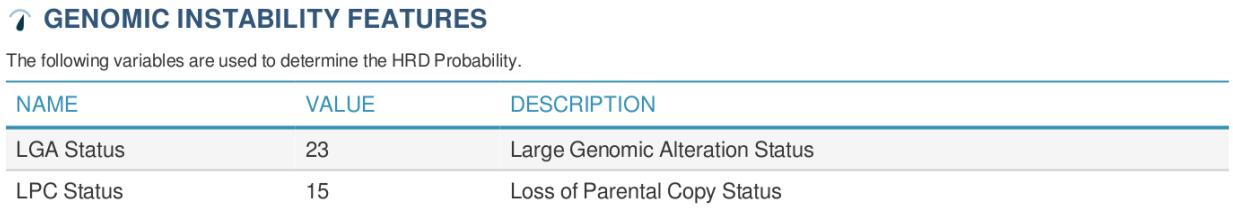
Copy number variations. This section is dedicated to copy number variations of the CCNE1 and RAD51 genes in terms of the number of copies detected. These are not taken into account when calculating the HRD probability and the GI probability.

Genomic instability features. This section provides details of the features that were computed to determine the genomic instability of the tumor. These features are used to determine the SeqOne HRD probability (**SomaHRD v1.2 and v1.3**) and GI probability (**SomaHRD version ≥**

v1.4). This section is intended to provide an understanding of how the probability is calculated. For each of the features, the following information is provided:

- ***Name*** : Name of the feature (eg. LGA Status or LGA Score)
- ***Value*** : The value that has been calculated for the patient
- ***Description*** : Brief description of the features

In case of *NON CONCLUSIVE* results, related to QC metrics below the expected thresholds, the sections *AMPLIFICATIONS* and *GENOMIC INSTABILITY FEATURES* of the report appear grayed out.



GENOMIC INSTABILITY FEATURES

The following variables are used to determine the HRD Probability.

NAME	VALUE	DESCRIPTION
LGA Status	23	Large Genomic Alteration Status
LPC Status	15	Loss of Parental Copy Status

Figure 39: *Features of HRD probability related to genomic instability.*

BRCA1/BRCA2 & TP53 mutations. In the case of a full HRD test including SomaVar analysis, we provide an additional section dedicated to pathogenic variants detected in the BRCA1, BRCA2 and TP53 genes. All BRCA1 and BRCA2 pathogenic variants corresponding to the criteria described in the **Methodology** section are listed. However, only the pathogenic or probably pathogenic TP53 variant with the highest allele frequency will be reported. Other pathogenic TP53 variants can be viewed in the associated SomaVar analysis. This section is hidden for sWGS reports only.

Mutations detected are reported in a table with the following columns:

- ***Gene & Transcript***: HGNC gene name and RefSeq transcript ID
- ***Variant***: HGVSC and HGVSP nomenclatures of the variant
- ***Evaluation***: Variant evaluation according to the ACMG criteria (*Pathogenic* or *Likely Pathogenic*)
- ***VAF***: Allelic fraction of the variant

If no pathogenic BRCA mutations have been identified in the linked panel analysis, this section of the report displays the following message: "No pathogenic mutations have been identified on BRCA1 and BRCA2 genes."

If no pathogenic TP53 mutations have been identified in the linked panel analysis, this section of the report displays the following message: "No pathogenic mutations have been identified on TP53 gene."

BRCA1/BRCA2 & TP53 MUTATIONS

GENE & TRANSCRIPT	VARIANT	EVALUATION	VAF
BRCA1 *NM_007294.4*	c.3116_3117del p.Ala1039GlufsTer6	Pathogenic	60.68%
BRCA1 *NM_007294.4*	c.3113_3117delinsG p.Glu1038GlyfsTer9	Pathogenic	58.18%
BRCA1 *NM_007294.4*	c.3113_3114del p.Glu1038GlyfsTer7	Pathogenic	60.08%
TP53 *NM_000546.6*	c.524G>A p.Arg175His	Pathogenic	39.29%

Figure 40: BRCA1/BRCA2 and TP53 pathogenic variants.

HRD summary. This last section of the report is dedicated to the results. This section aims to provide a detailed view of the results, displaying the following information:

- ***HRD Probability (SomaHRD v1.2 and v1.3)***: The HRD probability value, between 0 and 1. The status is positive when the value is greater than or equal to 0.5.
- ***GI probability*** (**SomaHRD version ≥ v1.4**): The GI probability value, between 0 and 1. The status is positive when this value is greater than or equal to 0.5.
- ***BRCA Status***: status determined by the detection of at least one pathogenic BRCA mutation: MUTATED, WILD-TYPE, or N/A in the absence of panel analysis.
- ***HRD Status or GI Status*** : *POSITIVE, NEGATIVE, NEGATIVE LOW CONFIDENCE, POSITIVE LOW CONFIDENCE* or *NON CONCLUSIVE* as displayed in the report header
- ***Method*** : The method (or mode) that was used to produce the HRD score:
    - Full HRD ⇒ "*WGS SHALLOW + BRCA1/2 GENE PANEL*"
    - Shallow Only ⇒ "*SHALLOW WGS ONLY*"

In case of *NEGATIVE LOW CONFIDENCE, POSITIVE LOW CONFIDENCE or NON CONCLUSIVE* results, those are displayed in light gray.

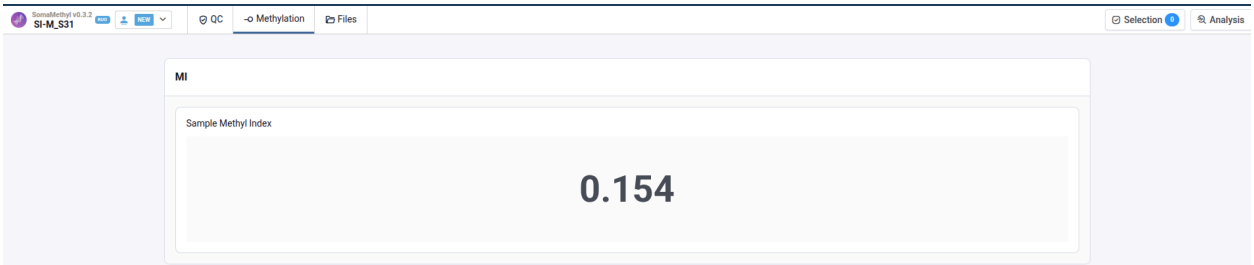
Details about the different values displayed in the report, as well as the overall methodology of the test are present in the "***Footnotes"*** and "***Methodology***" sections, located following the result.

In addition to the HRD report, the genomic instability graphs associated with the HRD analysis are available for download from the Files tab of the SomaHRD analysis. These graphs represent the normalized coverage on all chromosomes from the sWGS data.

In case of *NON CONCLUSIVE* results, related to QC metrics below the expected thresholds, the genomic instability graphs are not generated.

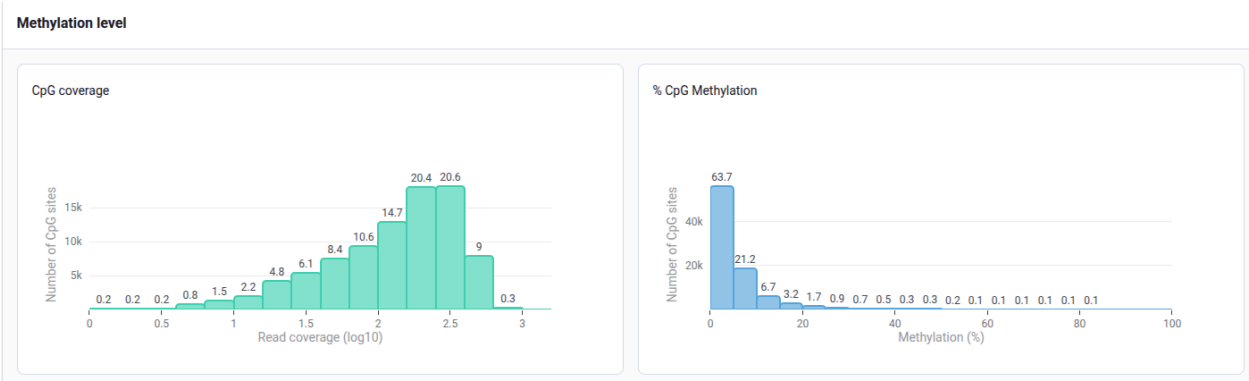
12.1.14. Methylation data for SomaMethyl

Methylation Index



The *Methylation Index* (MI) is a single numeric score that quantifies the overall methylation state of a biological sample and summarizes the tumor-associated methylation signal. It can sensitively detect even low levels of tumor-derived DNA within a background of normal DNA. In SomaMethyl, MI is computed in Standard mode using Agilent's proprietary algorithm (MethylAlgo). When a reference cohort is available, the sample MI can be compared to cohort MI values and interpreted using a user-defined threshold. To evaluate the MI, Agilent provided titration samples with expected methylation levels. The MI values calculated by the SomaMethyl Standard mode module are compared to the expected MI values provided by Agilent, and performance is summarized using a weighted agreement metric defined as 1 − wMAPE (weighted Mean Absolute Percentage Error).

Methylation level



CpG coverage (histogram)

A distribution of read depth across CpG sites within the targeted regions (manifest).

- X-axis (log10 read coverage): coverage per CpG site (number of reads supporting methylation calling at that CpG, i.e., methylated + unmethylated counts).
- Y-axis (# CpG sites): number of CpG sites in the manifest that have that coverage.

You can check whether the targeted CpGs are sufficiently covered and whether coverage is uniform enough to interpret methylation percentages confidently.

% CpG methylation (histogram)

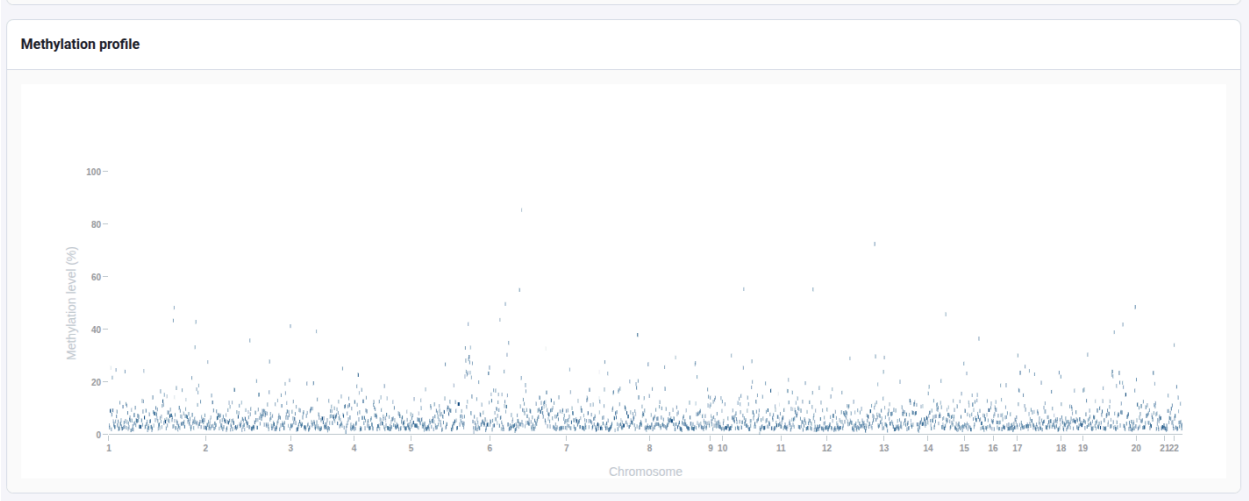
A distribution of methylation levels across CpG sites within the targeted regions (manifest).

- X-axis (methylation %): for each CpG site, the fraction of methylated reads among all reads covering that CpG.
- Y-axis (# CpG sites): number of CpG sites in the manifest showing that methylation level.

It provides a sample-level view of CpG methylation states across the targeted regions (often showing peaks near low and high methylation, depending on the regions captured and sample biology).

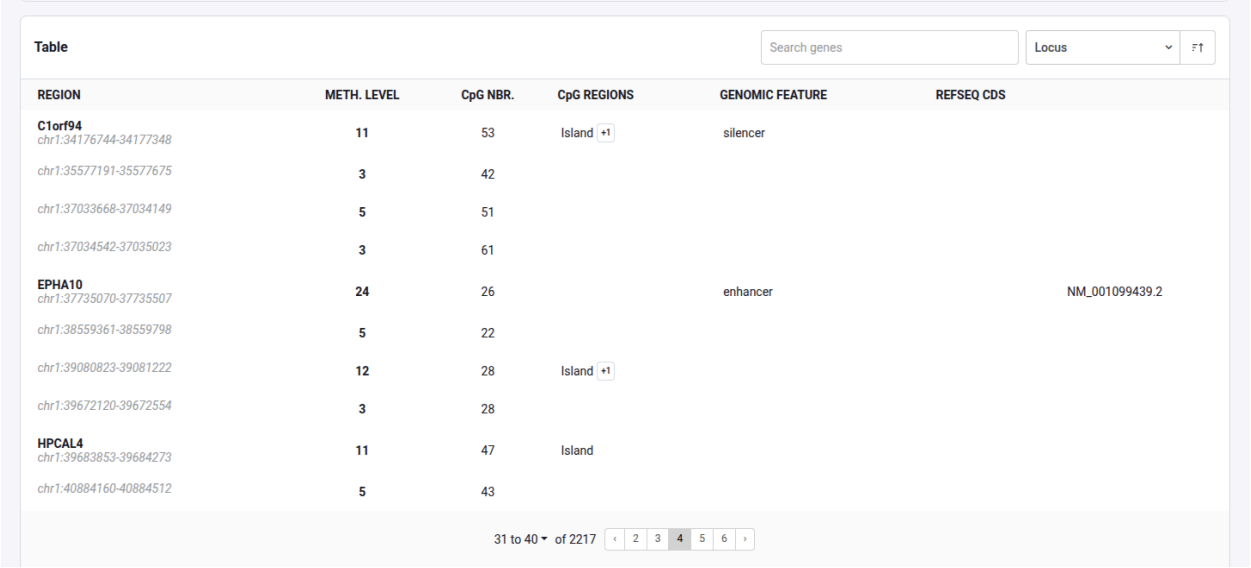
CpG sites with very low coverage produce less reliable methylation percentages because a small number of reads can swing the % up or down. For that reason, the %CpG methylation histogram should always be interpreted together with the CpG coverage histogram (and the other QC metrics).

Methylation profile



The methylation profile plot shows the sample's CpG methylation signal across the genome, displayed by chromosome. Each point corresponds to the average CpG methylation percentage computed over successive genomic windows (as in the `.mcpg.track.bedGraph` output). This view helps visualize broad methylation patterns and identify regions with higher or lower methylation compared with the rest of the genome.

Methylation Table



The Methylation table lists the genomic regions (chromosome:start–end) over which SomaMethyl summarizes CpG methylation (fixed windows or merged segments, depending on the step). For each region METH. LEVEL reports the aggregated methylation level computed from CpG methylation calls in that interval, and CpG NBR. indicates how many CpG sites contributed to the calculation. Each row is then annotated with CpG REGIONS (Island, Shore, Shelf when applicable) and functional annotations, including overlapping GENOMIC FEATURE regulatory elements (promoter, enhancer, silencer) and REFSEQ CDS (coding sequence) information.

13. Report

13.1.1. Output files

The *Files* tab lists the files generated at the different stages in the workset. Each file can be downloaded by clicking on the  icon.

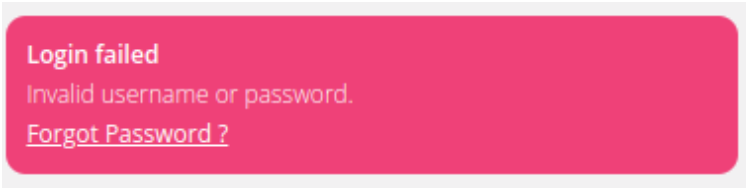
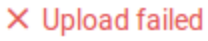
Some files do not include the sample name as a prefix. To download them with this prefix, you need to activate the 'Prefix with sample name' option at the top right of the page.

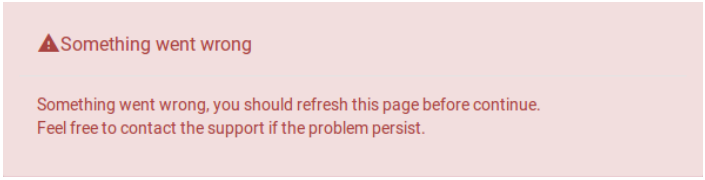
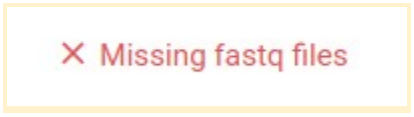
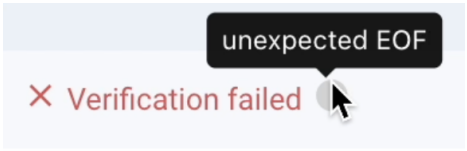
Files are available for downloading from the different sections in the interface:

- FastQ files are available in the project's *Samples* tab by selecting the relevant sample and then the  icon.
- Manifests (BED files) are available in the user's *Settings* via the icon in the top-right corner of the dashboard, and then in the *Manifests* tab.
- The analysis report(s) in PDF format is/are available in the *Reporting* tab after selecting an analysis.
- The files generated during an analysis (BAM, VCF, etc.) and the different HTML reports (quality, technical, etc.) are available in the *Files* tab for each analysis.

14. Troubleshooting and error messages

Several messages are used to notify users of an error

Error message	Description
	The password and/or login is/are incorrect. Users are invited to enter their login and password again, or click on "Forgot password?" to reset their password.
	The upload for the sample is pending. Users are advised to wait, not close the browser window and not cut their internet connection.
	The sample failed to upload. *This error may occur if the browser window was closed or if the internet connection suffered an outage during uploading.* Users are invited to try uploading the file(s) again.
	Indicates the progress of the current upload Users are advised to wait, not close the browser window and not cut their internet connection.
	Page loading. Users are invited to wait.

	The analysis is pending. Users are invited to wait.
	The analysis failed. Users are prompted to check the settings and data used to run the analysis and then restart the analysis.
	An error has occurred. The user is prompted to click on the "Esc" key. If the problem persists, contact support at support@seqone.com.
	This sample is missing the associated FastQ files, meaning that no new analyses can be launched for that sample.
	The "Unexpected EOF" (End Of File) message corresponds to a truncated file. The user is invited to check the data and reload the files concerned.

Users can contact the support team by clicking on the logo  at the top right of the user dashboard, then on the icon  .

15. Warranty

Please refer to the warranty terms and conditions in your maintenance agreement.

16. Legal notice

16.1. Terms of use

Information about the user license can be accessed by clicking on the user account logo followed by *Terms of Use*. Users can view and download the information in PDF format from the pop-up window by clicking on "*OPEN PDF*".

16.2. Limitation of liability

SeqOne disclaims any and all liability for any misuse or non-professional use of SeqOne platform.

SeqOne Platform is a product for research use only in your country and Seqone is not responsible for the diagnoses derived from the results obtained if the device is used in diagnostic procedures.

Consequently, users must exercise due caution when using the software and its results.

16.3. Copyrights and trademarks

All company and product names mentioned in this guide may be trademarks of their respective owners.