CGAR Documentation

In-Hee Lee

Oct 27, 2022

Contents:

CHAPTER 1

Introduction

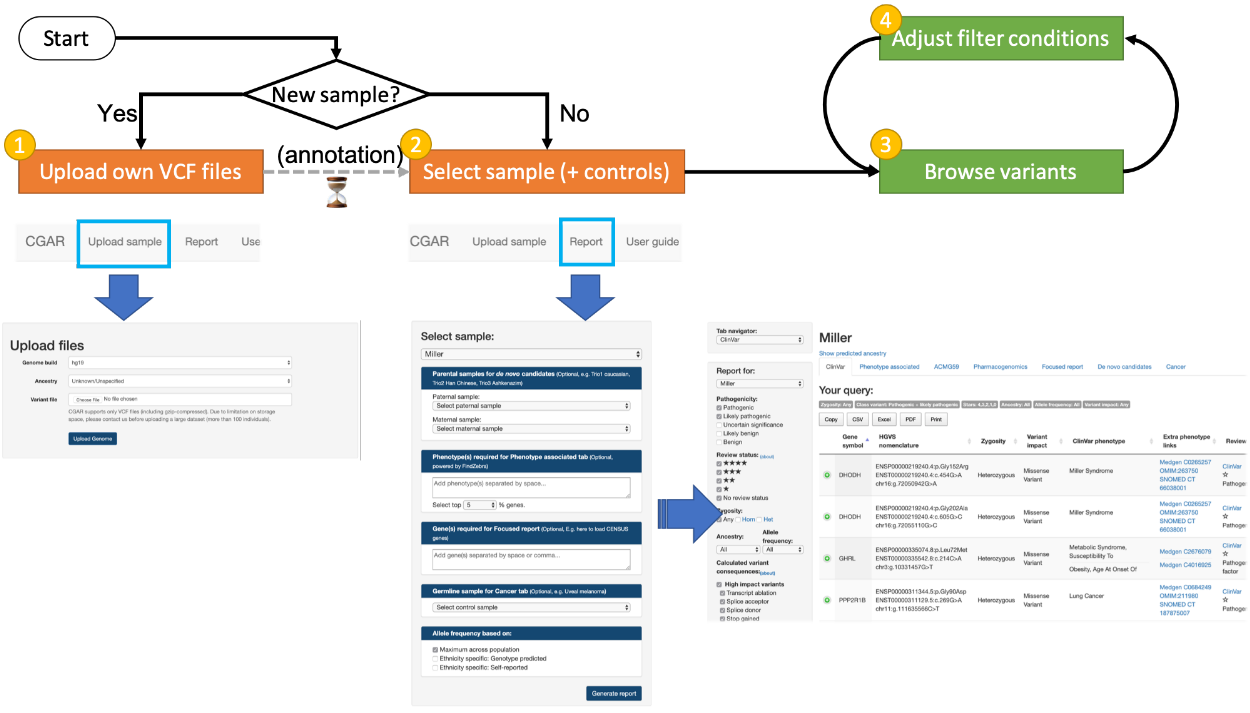
Clinical Genome & Ancestry Report (CGAR) is an interactive web application to dynamically filter and organize clinically implicated variants from variant call files in VCF. Variants are annotated with latest information from diverse sources and can be searched by simple yet comprehensive filtering options, with various links to external resources.

CGAR organizes variants according to the following categories:

- Reported disease-associated variants from HGMD® (license required) and ClinVar,
- Variants in putative disease-associated genes,
- Secondary findings recommended for reporting,
- Variants with pharmacogenomic consequences,
- Variants on user-defined genes,
- *De novo* variant candidates for trio (if parental samples are specified),
- Germline or somatic (if matched control sample is specified) variants implicated in cancer risk, diagnosis, treatment and prognosis.

Note: Some sections of CGAR report require specific license. Only registered users with proper license are allowed to access them.

For each category, variants selected by a set of configurable criteria are presented in intuitive and interactive web interface where users can browse and identify clinically implicated variants from thousands (or millions) of variants from whole exome (or genome) sequencing data.



1.1 CGAR public server

CGAR is freely available at https://tom.tch.harvard.edu/apps/cgar/.

Any interested users can use the full functionality of CGAR without registration. However, the workspace of unregistered users are shared with each other, implying that they can access the same set of samples. In other words, samples uploaded by an user *A* can be viewed by another user *B*.

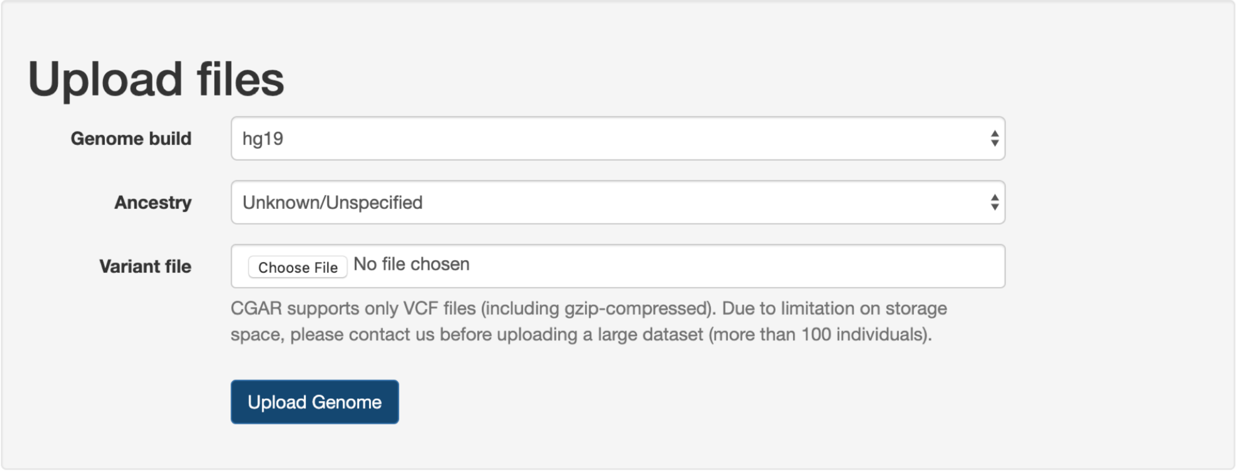
Users who need private workspace are encouraged to register or to set up CGAR locally. Please contact us to inquiry about registration.

Note: Check *Running CGAR locally* for guidelines to set up CGAR locally.

CHAPTER 2

Upload Own Variant Files

Clicking `Upload sample` at the top menu opens a page for users to upload their own variant call file. The files need to follow the Variant Call Format (VCF), and only one file can be uploaded at a time. However, VCF files containing more than one individual or sample are accepted.



Upload files

Genome build: hg19

Ancestry: Unknown/Unspecified

Variant file: Choose File No file chosen

CGAR supports only VCF files (including gzip-compressed). Due to limitation on storage space, please contact us before uploading a large dataset (more than 100 individuals).

First, users need to select correct version of human reference genome assembly (`Genome build`): **hg19** or **hg38**. If uncertain, it is recommended to check the header lines within the file (lines starting with `#`). To do so, you can run the following command in a terminal:

```
# if the file is gzip-compressed
$ zcat /path/to/variant/file.vcf.gz | head -n <arbitrary number of lines to view>
# if the file is not compressed
$ head -n <arbitrary number of lines to view> /path/to/variant/file.vcf

##fileformat=VCFv(version number)
...
```

Then, look for a line starting with `##reference=`. This line usually contains file name for referece genome fasta file, which could provide a hint of reference genome assembly version. For example, if the filename contained **b37**,

hg19, or **human_g1k_v37**, it's most likely that you can choose **hg19**. Or, find lines starting with `##contig=`, which list the set of chromosomes with their names and lengths. From Human Assembly Data in Genome Reference Consortium, you can find chromosome lengths in different versions of assembly, and match with values found in your VCF file.

Next, the global ancestry (`Ancestry`) for the individual or sample to be analyzed, if known, is recommended to be set. The same five continental groups as in the 1000 Genomes Project are used: **African**, **East Asian**, **European**, **South Asian**, and **American**. If unknown or uncertain, it can be left as **Unknown/Unspecified**, and CGAR will estimate the global ancestry from variant file.

In the `Variant file`, local variant file can be selected for uploading to CGAR. Only files ending with the extension of `.vcf` or `.vcf.gz` are accepted.

Note: Due to limitation on storage space, please contact us before uploading a file containing large number of individuals (more than 100 individuals).

Finally, when user clicks on the `Upload Genome` button (at the bottom of dialog), the variant call file gets transferred to the server and placed on a queue for annotation and analysis. The required amount of time to finish the process varies, depending on the number of variants in the file. Under normal circumstances, a file of whole genome containing 3 to 4 million variants gets processed in one hour. Users will receive a notification email each time a file is done processed and available for analysis.

2.1 Browse and track previous files

The bottom of the page displays a table listing all variant files uploaded so far (including the one you have just uploaded).

In this table, users can see:

- The name of sample(s) in the variant file - `Genome label`.
- The name of the variant file - `Genome file name`.
- The version of human reference genome assembly - `Genome build`.
- The global ancestry specified at the time of uploading - `Genome ancestry`.
- The version of annotation - `Annotation`.
- The time and date when the file was first uploaded or done processing - `Uploaded`.

When the file was just uploaded, but not yet finished processed and ready for analysis, it would first appear in the table with **none** as `Annotation`. Also, the `Genome label` would simply show the number of samples within the file, e.g., **(3 sample(s))** or **(1 sample(s))**.

Later, the table will be updated with as many samples as was in the file, with each sample as separate row carrying their own identifier (specified in original VCF file) as `Genome label`. At this point, all samples will have the current version of annotation in `Annotation`, and `Uploaded` will reflect the date and time when the processing was finished.

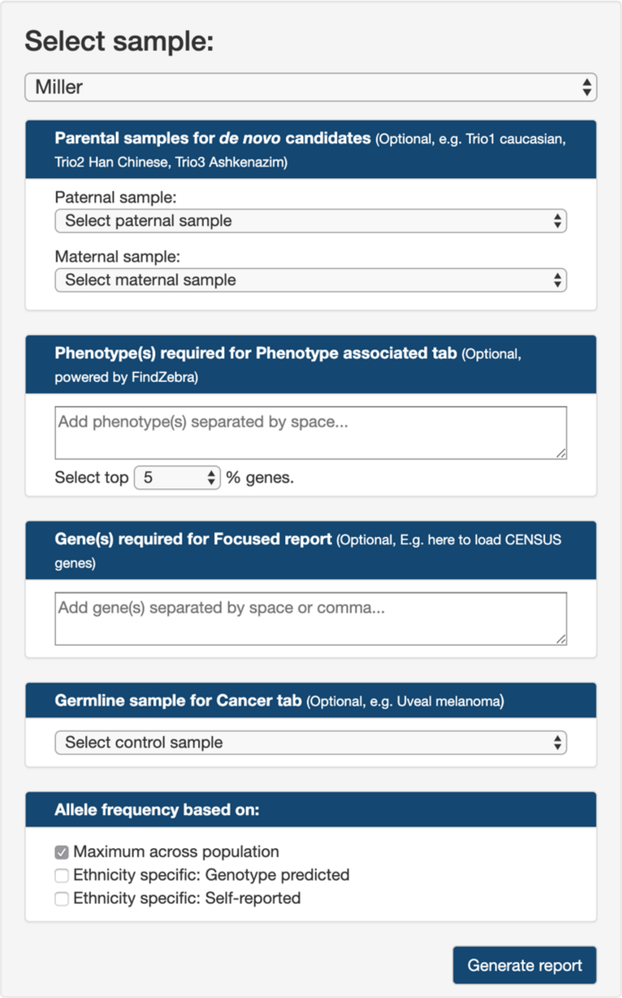
CHAPTER 3

Start Analysis on CGAR

To start analyzing any of finished samples, simply click `Report` from the top menu to open the dialog as follows.

See also:

Refer to *Upload Own Variant Files* for how to add new samples or check if the samples are ready.



Select sample:

Miller

Parental samples for *de novo* candidates (Optional, e.g. Trio1 caucasian, Trio2 Han Chinese, Trio3 Ashkenazim)

Paternal sample: Select paternal sample

Maternal sample: Select maternal sample

Phenotype(s) required for Phenotype associated tab (Optional, powered by FindZebra)

Add phenotype(s) separated by space...

Select top 5 % genes.

Gene(s) required for Focused report (Optional, E.g. here to load CENSUS genes)

Add gene(s) separated by space or comma...

Germline sample for Cancer tab (Optional, e.g. Uveal melanoma)

Select control sample

Allele frequency based on:

☑ Maximum across population
☐ Ethnicity specific: Genotype predicted
☐ Ethnicity specific: Self-reported

The dropdown list at the top contains analysis-ready samples for users to start work on. The CGAR public server provides a list of publicly available samples for all users who want to try out CGAR. See *Publicly available samples*.

If available, users can specify parental samples for the selected case in `Parental samples for de novo candidates`. The dropdown lists `Paternal sample` and `Maternal sample`, both also populated with analysis-ready samples, are used to specify paternal and maternal cases, respectively. For the *de novo* candidate analysis to work, both `Paternal sample` and `Maternal sample` needs to be specified.

See also:

Refer to *De novo variant candidates* for more details on trio analysis using CGAR.

For analysis using *Publicly available samples*, three short-cut links are provided for each of three trios (`Trio1 caucasian`, `Trio2 Han Chinese`, and `Trio3 Ashkenazim`).

In cases where users are interested in variants on genes associated with specific phenotypes, yet not certain about which genes to look for, they can provide a list of keywords describing the phenotype of interest in `Phenotype(s) required for Phenotype associated tab`. Users can type in as many keywords as they want, each separated by space. CGAR will find any variants on genes associated with any of keywords.

See also:

Refer to *Putative phenotype-associated genes* for more details on how the genes associated with the given keywords are used in CGAR.

For more focused analysis on variants in pre-defined set of genes, users can provide their genes of interest through

`Gene(s) required for Focused report`. CGAR will expect a list of official gene symbols (if uncertain, check the approved gene symbols from HUGO Gene Nomenclature Committee, each separated by either comma or space. For the convenience, the list of cancer-associated genes as curated by Cancer Gene Census is provided as short-cut link.

When a matched control for cancer sample is available, specifying it as `Germline sample for Cancer tab` would make CGAR display only somatic variants on `Cancer` subsection. Again, a short-cut for public example of uveal melanoma case (tumor and its matched blood) is provided for convenience (`Uveal melanoma`).

Finally, users can specify how to utilize population-specific allele frequencies in CGAR.

1. **Maximum across population**: Default choice. Limits an allele frequency of a variant to the highest value (most common) across all populations. Results in the most strict selection of rare variants.
2. **Ethnicity specific: Genotype predicted**: Limits an allele frequency of a variant to the value from a population of the same global ancestry as estimated from variants' genotypes.
3. **Ethnicity specific: Self-reported**: The same as the above, but uses the ancestry specified upon uploading the variants file. Not available if **Unknown/Unspecified** is chosen.

See also:

Check *Genotype-based prediction of ancestral composition* for details on genotype-based ancestry prediction in CGAR.

After setting all the above options, click the button `Generate report` for CGAR to start generating report tables.

3.1 Publicly available samples

The following samples are readily available to all users in CGAR public server.

- Maternal trio of CHPH/Utah pedigree 1463
  - Mother (NA12878) as **trio1_NA12878_daughter**.
  - Maternal grandfather (NA12891) as **trio1_NA12891_father**.
  - Maternal grandmother (NA12892) as **trio1_NA12892_mother**.
- The Han Chinese trio from Genome in a Bottle Consortium.
  - The son (NA24631) as **trio2_GM24631_son**.
  - The father (NA24694) as **trio2_GM24694_father**.
  - The mother (NA24695) as **trio2_GM24695_mother**.
- The Ashkenazim trio from Genome in a Bottle Consortium.
  - The son (NA24385) as **trio3_NA24385_son**.
  - The father (NA24149) as **trio3_NA24149_father**.
  - The mother (NA24143) as **trio3_NA24143_mother**.
- A pair of tumor sample with its matched blood control from uveal melanoma (MIM: 155720) case ref.
  - Tumor sample as **patient1_cancer_melanoma**.
  - Control sample as **patient1_blood_melanoma**.
- Artifical examples of specific phenotypes.
  - Artificial sample for Miller syndrome (MIM: 263750) as **Miller**.

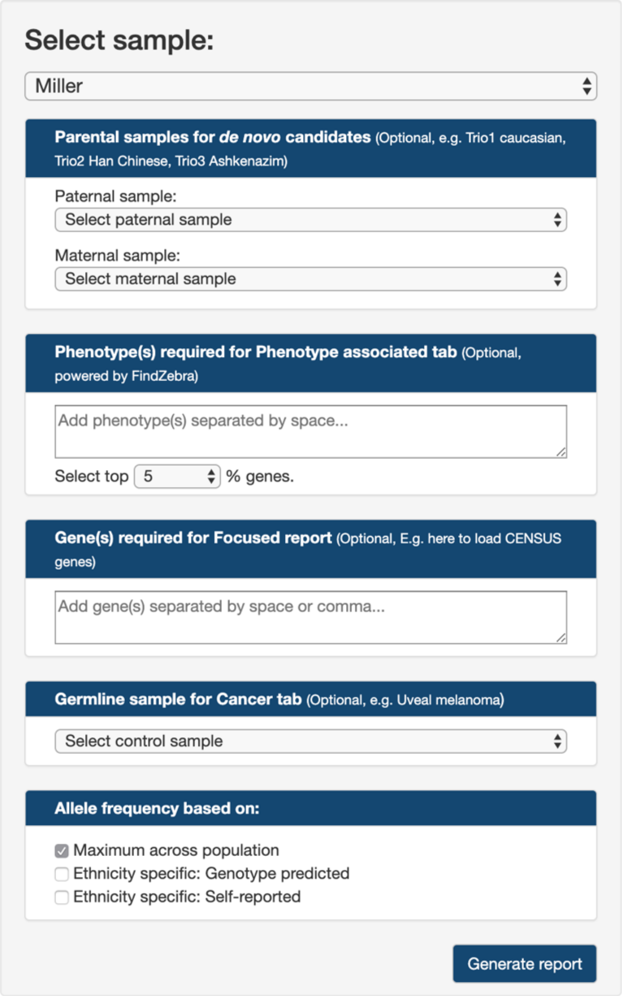
  - Artificial sample for Schinzel-Giedion syndrome (MIM: 269150) as **schinzel_giedion**.

3.2 An Example Use Case

3.2.1 Method - filter variants by allele frequency

Let's use an artificial sample for Miller syndrome to illustrate steps to identify variants potentially associated with phenotype.

First, select the sample **Miller** from the list of samples, then click `Generate report` to start analysis.



Select sample:

Miller

Parental samples for *de novo* candidates (Optional, e.g. Trio1 caucasian, Trio2 Han Chinese, Trio3 Ashkenazim)

Paternal sample: Select paternal sample

Maternal sample: Select maternal sample

Phenotype(s) required for Phenotype associated tab (Optional, powered by FindZebra)

Add phenotype(s) separated by space...

Select top 5 % genes.

Gene(s) required for Focused report (Optional, E.g. here to load CENSUS genes)

Add gene(s) separated by space or comma...

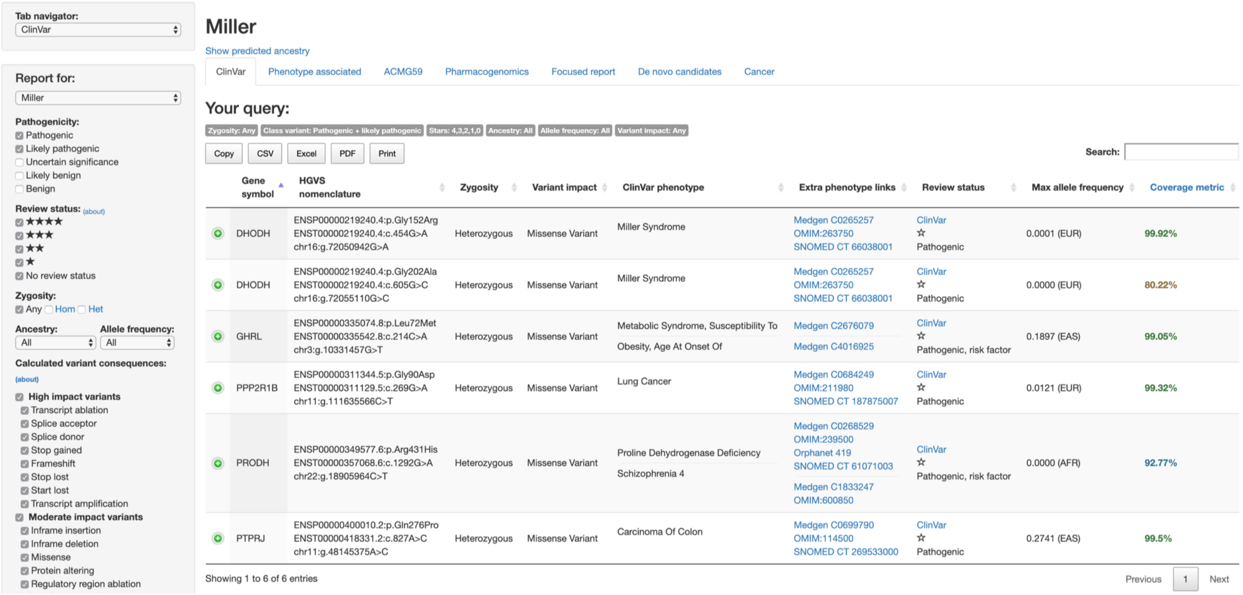
Germline sample for Cancer tab (Optional, e.g. Uveal melanoma)

Select control sample

Allele frequency based on:

☑ Maximum across population
□ Ethnicity specific: Genotype predicted
□ Ethnicity specific: Self-reported

By default, the `ClinVar` tab will show 6 variants as shown below.



Tab navigator:
ClinVar

Report for:
Miller

Pathogenicity:
☑ Pathogenic
☑ Likely pathogenic
□ Uncertain significance
□ Likely benign
□ Benign

Review status: (about)
☑ ★★★★
☑ ★★★
☑ ★★
☑ ★
☑ No review status

Zygosity:
☑ Any □ Hom □ Het

Ancestry: All
Allele frequency: All

Calculated variant consequences:
(about)
☑ High impact variants
☑ Transcript ablation
☑ Splice acceptor
☑ Splice donor
☑ Stop gained
☑ Frameshift
☑ Stop lost
☑ Start lost
☑ Transcript amplification
☑ Moderate impact variants
☑ Inframe insertion
☑ Inframe deletion
☑ Missense
☑ Protein altering
☑ Regulatory region ablation

Miller

Show predicted ancestry

Your query:

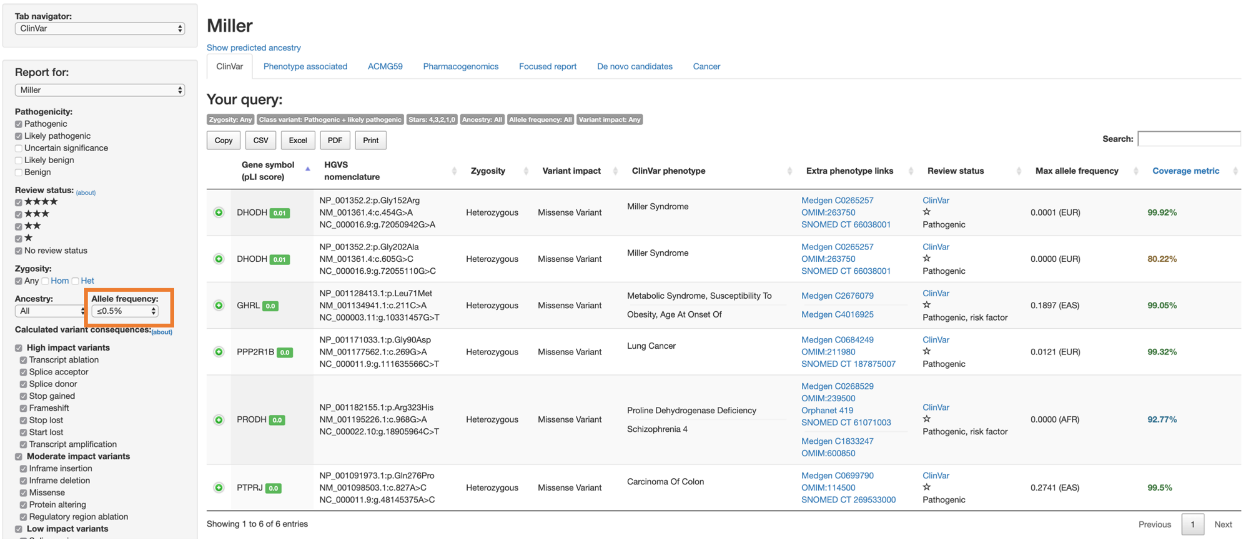
Zygosity: Any | Class variant: Pathogenic + likely pathogenic | Stars: 4,3,2,1,0 | Ancestry: All | Allele frequency: All | Variant impact: Any

Search:

Gene symbol	HGVS nomenclature	Zygosity	Variant impact	ClinVar phenotype	Extra phenotype links	Review status	Max allele frequency	Coverage metric
DHODH	ENSP00000219240.4:p.Gly152Arg ENST00000219240.4:c.454G>A chr16:g.72050942G>A	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0001 (EUR)	99.92%
DHODH	ENSP00000219240.4:p.Gly202Ala ENST00000219240.4:c.605G>C chr16:g.72055110G>C	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0000 (EUR)	80.22%
GHRL	ENSP00000335074.8:p.Leu72Met ENST00000335542.8:c.214C>A chr3:g.10331457G>T	Heterozygous	Missense Variant	Metabolic Syndrome, Susceptibility To Obesity, Age At Onset Of	Medgen C2676079 Medgen C4016925	ClinVar ☆ Pathogenic, risk factor	0.1897 (EAS)	99.05%
PPP2R1B	ENSP00000311344.5:p.Gly90Asp ENST00000311129.5:c.269G>A chr11:g.111635566C>T	Heterozygous	Missense Variant	Lung Cancer	Medgen C0684249 OMIM:211980 SNOMED CT 187875007	ClinVar ☆ Pathogenic	0.0121 (EUR)	99.32%
PRODH	ENSP00000349577.6:p.Arg431His ENST00000357068.6:c.1292G>A chr22:g.18905964C>T	Heterozygous	Missense Variant	Proline Dehydrogenase Deficiency Schizophrenia 4	Medgen C0268529 OMIM:239500 Orphanet 419 SNOMED CT 61071003 Medgen C1833247 OMIM:600850	ClinVar ☆ Pathogenic, risk factor	0.0000 (AFR)	92.77%
PTPRJ	ENSP00000400010.2:p.Gln276Pro ENST00000418331.2:c.827A>C chr11:g.48145375A>C	Heterozygous	Missense Variant	Carcinoma Of Colon	Medgen C0699790 OMIM:114500 SNOMED CT 269533000	ClinVar ☆ Pathogenic	0.2741 (EAS)	99.5%

Showing 1 to 6 of 6 entries

However, given the fact that the incidence of Miller syndrome is one in a million newborns (ref), it could be useful to restrict our search to vary rare variants (allele frequency less than 0.5%), then click `Generate report`. (based on ACMG AMP guideline for the interpretation of sequence variants, Table 4: Criteria for benign variants, BS1 - allele frequency is greater than expected for disorder.)



Tab navigator:
ClinVar

Report for:
Miller

Pathogenicity:
☑ Pathogenic
☑ Likely pathogenic
□ Uncertain significance
□ Likely benign
□ Benign

Review status: (about)
☑ ★★★★
☑ ★★★
☑ ★★
☑ ★
☑ No review status

Zygosity:
☑ Any □ Hom □ Het

Ancestry: All
Allele frequency: ≤0.5%

Calculated variant consequences: (about)
☑ High impact variants
☑ Transcript ablation
☑ Splice acceptor
☑ Splice donor
☑ Stop gained
☑ Frameshift
☑ Stop lost
☑ Start lost
☑ Transcript amplification
☑ Moderate impact variants
☑ Inframe insertion
☑ Inframe deletion
☑ Missense
☑ Protein altering
☑ Regulatory region ablation
☑ Low impact variants

Miller

Show predicted ancestry

Your query:

Zygosity: Any | Class variant: Pathogenic + likely pathogenic | Stars: 4,3,2,1,0 | Ancestry: All | Allele frequency: All | Variant impact: Any

Search:

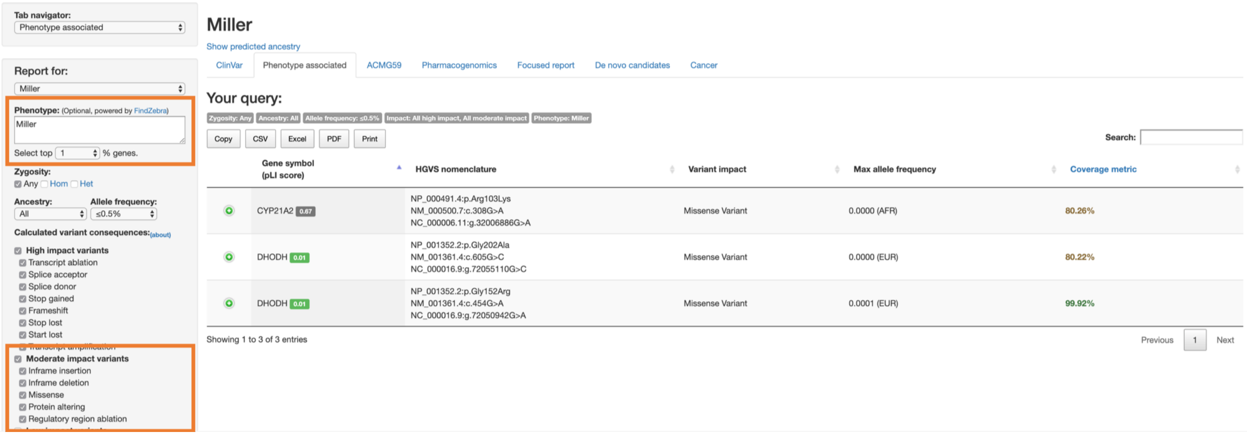
Gene symbol (pLI score)	HGVS nomenclature	Zygosity	Variant impact	ClinVar phenotype	Extra phenotype links	Review status	Max allele frequency	Coverage metric
DHODH 0.01	NP_001352.2:p.Gly152Arg NM_001361.4:c.454G>A NC_000016.9:g.72050942G>A	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0001 (EUR)	99.92%
DHODH 0.01	NP_001352.2:p.Gly202Ala NM_001361.4:c.605G>C NC_000016.9:g.72055110G>C	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0000 (EUR)	80.22%
GHRL 0.0	NP_001128413.1:p.Leu71Met NM_001134941.1:c.211C>A NC_000003.11:g.10331457G>T	Heterozygous	Missense Variant	Metabolic Syndrome, Susceptibility To Obesity, Age At Onset Of	Medgen C2676079 Medgen C4016925	ClinVar ☆ Pathogenic, risk factor	0.1897 (EAS)	99.05%
PPP2R1B 0.0	NP_001171033.1:p.Gly90Asp NM_001177562.1:c.269G>A NC_000011.9:g.111635566C>T	Heterozygous	Missense Variant	Lung Cancer	Medgen C0684249 OMIM:211980 SNOMED CT 187875007	ClinVar ☆ Pathogenic	0.0121 (EUR)	99.32%
PRODH 0.0	NP_001182155.1:p.Arg323His NM_001195226.1:c.968G>A NC_000022.10:g.18905964C>T	Heterozygous	Missense Variant	Proline Dehydrogenase Deficiency Schizophrenia 4	Medgen C0268529 OMIM:239500 Orphanet 419 SNOMED CT 61071003 Medgen C1833247 OMIM:600850	ClinVar ☆ Pathogenic, risk factor	0.0000 (AFR)	92.77%
PTPRJ 0.0	NP_001091973.1:p.Gln276Pro NM_001098503.1:c.827A>C NC_000011.9:g.48145375A>C	Heterozygous	Missense Variant	Carcinoma Of Colon	Medgen C0699790 OMIM:114500 SNOMED CT 269533000	ClinVar ☆ Pathogenic	0.2741 (EAS)	99.5%

Showing 1 to 6 of 6 entries

This will leave us with only 3 variants: 2 missense variants on *DHODH* gene and one missense variant on *PRODH* gene. Of these, the 2 variants on *DHODH* gene were the disease-assoiciated variants that were added to create artificial samples.

3.2.2 Alternative method - search variants by genes associated with phenotype

Alternatively, we could search for genes associated with the phenotype and identify variants on those genes, within `Phenotype associated` tab. Since we know that the variants are from sample with Miller syndrome, type in **Miller** into the `Phenotype` text box. Also, to limit our search for the most relevant genes, we select the top 1% of the genes (select **1** from the dropdown value beneath the `Phenotype` text box). We also choose to see any variants that are of high or moderate impact: select both **High impact variants** and **Moderate impact variants** under the `Calculated variant consequences`, then click `Generate report`.



Tab navigator: Phenotype associated

Report for: Miller

Phenotype: (Optional, powered by FindZebra) Miller

Select top 1 % genes.

Zygosity: ☑ Any □ Hom □ Het

Ancestry: All

Allele frequency: ≤0.5%

Calculated variant consequences: (about)

- ☑ High impact variants
  - ☑ Transcript ablation
  - ☑ Splice acceptor
  - ☑ Splice donor
  - ☑ Stop gained
  - ☑ Frameshift
  - ☑ Stop lost
  - ☑ Start lost
- ☑ Moderate impact variants
  - ☑ Inframe insertion
  - ☑ Inframe deletion
  - ☑ Missense
  - ☑ Protein altering
  - ☑ Regulatory region ablation

Miller

Your query:

Zygosity: Any | Ancestry: All | Allele frequency: ≤0.5% | Impact: All high impact, All moderate impact | Phenotype: Miller

	Gene symbol (pLI score)	HGVS nomenclature	Variant impact	Max allele frequency	Coverage metric
⊕	CYP21A2 0.67	NP_000491.4:p.Arg103Lys NM_000500.7:c.308G>A NC_000006.11:g.32006886G>A	Missense Variant	0.0000 (AFR)	80.26%
⊕	DHODH 0.01	NP_001352.2:p.Gly202Ala NM_001361.4:c.605G>C NC_000016.9:g.72055110G>C	Missense Variant	0.0000 (EUR)	80.22%
⊕	DHODH 0.01	NP_001352.2:p.Gly152Arg NM_001361.4:c.454G>A NC_000016.9:g.72050942G>A	Missense Variant	0.0001 (EUR)	99.92%

Showing 1 to 3 of 3 entries

Again, we could see the 2 missense variants on *DHODH* gene (the same variants we found in the previous method), and another missense variant on *CYP21A2* gene.

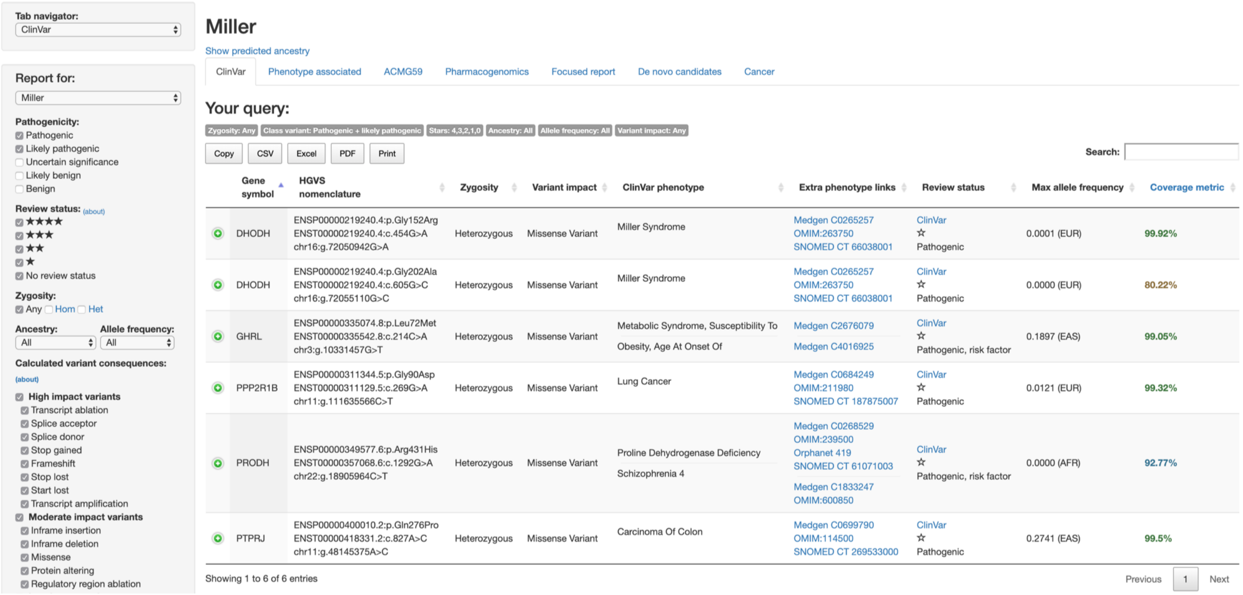
This example illustrates how CGAR can identify known disease-associated variants or rare variants in the genes associated with phenotype.

CHAPTER 4

Inside CGAR Report

For each sample, the generated report shows the name (identifier) of the chosen sample with 9 sections of variants organized by their analytical implications.

- (Restricted) Disease-associated variants reported in HGMD - `HGMD`
- Disease-associated variants reported in ClinVar - `ClinVar`
- (Restricted) Variants on genes associated with rare-diseases in OrphaData - `Orphanet`
- Variants on genes with putative association to user-defined phenotype - `Phenotype associated`
- Secondary findings - `ACMG59`
- Variants with potential pharmacogenomic implications - `Pharmacogenomics`
- Variants on user-defined set of genes - `Focused report`
- De novo variant candidates from trio - `De novo candidates`
- (Somatic) Variants potentially associated with cancer - `Cancer`



Tab navigator: ClinVar

Report for: Miller

Pathogenicity:
- ☑ Pathogenic
- ☑ Likely pathogenic
- □ Uncertain significance
- □ Likely benign
- □ Benign

Review status: (about)
- ☑ ★★★★
- ☑ ★★★
- ☑ ★★
- ☑ ★
- ☑ No review status

Zygosity: ☑ Any □ Hom □ Het

Ancestry: All Allele frequency: All

Calculated variant consequences: (about)
- ☑ High impact variants
  - ☑ Transcript ablation
  - ☑ Splice acceptor
  - ☑ Splice donor
  - ☑ Stop gained
  - ☑ Frameshift
  - ☑ Stop lost
  - ☑ Start lost
  - ☑ Transcript amplification
- ☑ Moderate impact variants
  - ☑ Inframe insertion
  - ☑ Inframe deletion
  - ☑ Missense
  - ☑ Protein altering
  - ☑ Regulatory region ablation

Miller

Show predicted ancestry

ClinVar | Phenotype associated | ACMG59 | Pharmacogenomics | Focused report | De novo candidates | Cancer

Your query:

Zygosity: Any | Class variant: Pathogenic + likely pathogenic | Stars: 4,3,2,1,0 | Ancestry: All | Allele frequency: All | Variant impact: Any

Copy | CSV | Excel | PDF | Print   Search:

	Gene symbol	HGVS nomenclature	Zygosity	Variant impact	ClinVar phenotype	Extra phenotype links	Review status	Max allele frequency	Coverage metric
⊕	DHODH	ENSP00000219240.4:p.Gly152Arg ENST00000219240.4:c.454G>A chr16:g.72050942G>A	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0001 (EUR)	99.92%
⊕	DHODH	ENSP00000219240.4:p.Gly202Ala ENST00000219240.4:c.605G>C chr16:g.72055110G>C	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0000 (EUR)	80.22%
⊕	GHRL	ENSP00000335074.8:p.Leu72Met ENST00000335542.8:c.214C>A chr3:g.10331457G>T	Heterozygous	Missense Variant	Metabolic Syndrome, Susceptibility To Obesity, Age At Onset Of	Medgen C2676079 Medgen C4016925	ClinVar ☆ Pathogenic, risk factor	0.1897 (EAS)	99.05%
⊕	PPP2R1B	ENSP00000311344.5:p.Gly90Asp ENST00000311129.5:c.269G>A chr11:g.111635566C>T	Heterozygous	Missense Variant	Lung Cancer	Medgen C0684249 OMIM:211980 SNOMED CT 187875007	ClinVar ☆ Pathogenic	0.0121 (EUR)	99.32%
⊕	PRODH	ENSP00000349577.6:p.Arg431His ENST00000357068.6:c.1292G>A chr22:g.18905964C>T	Heterozygous	Missense Variant	Proline Dehydrogenase Deficiency Schizophrenia 4	Medgen C0268529 OMIM:239500 Orphanet 419 SNOMED CT 61071003 Medgen C1833247 OMIM:600850	ClinVar ☆ Pathogenic, risk factor	0.0000 (AFR)	92.77%
⊕	PTPRJ	ENSP00000400010.2:p.Gln276Pro ENST00000418331.2:c.827A>C chr11:g.48145375A>C	Heterozygous	Missense Variant	Carcinoma Of Colon	Medgen C0699790 OMIM:114500 SNOMED CT 269533000	ClinVar ☆ Pathogenic	0.2741 (EAS)	99.5%

Showing 1 to 6 of 6 entries   Previous 1 Next

The side manu on the left reflects settings for current section and provides means to change settings and to reanalyze. The main table on the right presents the essential information on variants as well as links to further details or external resources.

In the following sections, each component and section in CGAR report is described in detail.

4.1 Genotype-based prediction of ancestral composition

Since variant allele frequencies vary across populations, the allele frequences needs to be compared with those from population of matching ancestry. CGAR implements an ancestry proportion analysis for WGS, [EIGMIX], to show the predicted ancestral composition. The global ancestry (ancestry group corresponds to the majority in the ancestral composition) is derived from the predicted ancestral composition, and used to decide which population should be used to compare allele frequencies.

4.1.1 Methods to predict ancestral composition

EIGMIX derives principal components from surrogate populations with reported ancestry and projects an individual of interest to the principal components to determine its ancestry. It first assumes center coordinates of principal components for each surrogate population as unit vectors in ancestral composition space. For example, in 3 ancestral populations A, B, and C, centroids of principal components in A, B, and C are mapped to unit vectors in 3-dimensional space of ancestral composition, $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. Then EIGMIX builds linear transformation from the principal component space to ancestral proportions, which is used to calculate the proportions of ancestral populations for an individual. EIGMIX was selected for its accuracy and ability to handle millions of SNPs and large number of query individuals efficiently.

As surrogate ancestral populations, CGAR uses the five continental-level population groups in the 1000 Genomes Project Phase 3 dataset: African, American, East Asian, European, and South Asian. Each continental-level population groups can be further divided into 4 to 7 populations. For example, the continental-level population **European** in 1000 Genomes Project consists of **CEU** (Utah residents (CEPH) with Northern and Western European ancestry), **TSI** (Toscani in Italia), **FIN** (Finnish in Finland), **GBR** (British in England and Scotland), and **IBS** (Iberian population in Spain). Within each continental-level population, CGAR uses the population showing relatively lower admixed ancestral structure than other population as surrogate population as follows:

Continental-level population	Population chosen as surrogate	The size of chosen population
American	PEL (Peruvians from Lima, Peru)	85
African	YRI (Yoruba in Ibadan, Nigeria)	108
East Asian	CHB (Han Chinese in Beijing, China)	103
European	CEU	99
South Asian	ITU (Indian Telugu from the UK)	102

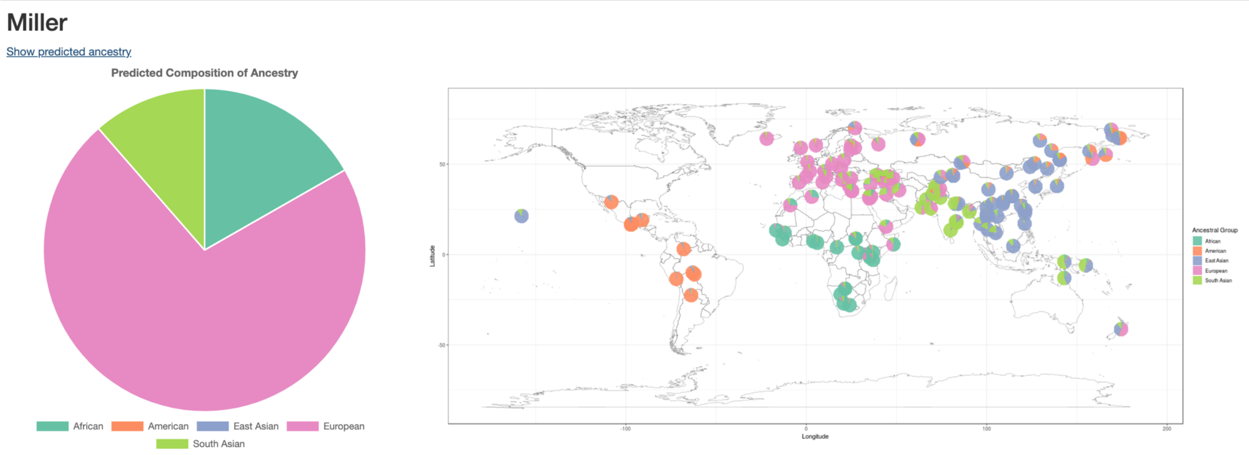
To balance between groups, 85 individuals are randomly selected from each population, making a total of 425 individuals as reference panel.

Next, only common bi-allelic SNPs (minor allele frequency of 5% or higher in 1000 Genomes Project) are collected from the 425 individuals. The selected SNPs were further pruned to make 1058271 SNPs that were at least 2000 bases apart from each other.

For each new individual, only the common variants between the given individual and the 1058271 SNPs in the reference panel are used for EIGMIX analysis. With common variants, EIGMIX builds new principal component space and derives new linear transformation to map between principal component space and ancestral proportions space. Then the coordinate of new individual in principal component space is calculated and converted to predictions on ancestral proportions of the new individual.

4.1.2 Use of ancestral composition in CGAR

In the main screen of generated reports in CGAR, there is a link `Show predicted ancestry` which opens graphs as in the following image:



Miller

Show predicted ancestry

The pie chart on the left shows the predicted ancestral composition of the selected sample (**Miller**). In the current example, European takes the majority of ancestral composition. When multiple samples are selected (as in the case of trio or cancer-control pairs), the ancestral compositions of all selected samples are drawn as nested doughnuts.

CGAR shows the ancestral compositions of 278 individuals from the Simons Genome Diversity Project (chart on the right). These individuals were recruited from 127 populations from disparate locations around the world, and each pie graph was plotted on the matching geographic location. Users can interpret the ancestral origin of the sample by comparing it with these individuals.

In this example, the majority of ancestral composition is European, which is also confirmed by pie charts from individuals on European nations. Therefore, for this sample, when selecting rare variants, it would be accurate to select variants based on allele frequencies in European populations.

4.2 Side menu and main table

4.2.1 Components in side menu

The side menu on the left of main report body provides an interface for users to control and adjust parameters in CGAR, thus facilitating interactive analysis.

The `Tab navigator` on the top reflects current tab in the main report, and provides an alternative to navigate betweet tabs. Selecting the name of different tab from the dropdown list will change the current tab in the main report.

Following `Tab navigator`, the dropdown list under headings `Report for:` shows current sample by default. It shows the same list of analysis-ready samples as in `Upload sample`. Users can also select different sample in this dropdown list, and (optionally) adjust parameters below to generate a new report for different sample, moving between samples continuously.

Some of contents in the side menu changes depending on the current tab in the main screen. For example, if `ClinVar` tab is focused in the main screen, the side menu will show controls specific to the tab such as the review status of variants in ClinVar. However, if `Pharmacogenomics` tab becomes the current tab, the side menu will change to show controls for the new current tab. The controls in side menu specific to each tab will be described in subsections for the tab.

Controls in the side menu that are common to all tabs are as follows:

- `Zygosity`: The genotype of variants. Can be either **Any** (default, includes variants of all genotype), **Hom** (homozygous variants), or **Het** (heterozygous variants).
- `Ancestry`: The population group to take allele frequency values.
- `Allele frequency`: The upper threshold of allele frequencies for variants to be included in the report. Can be either **All** (no restriction on variant allele frequencies), **<0.5%**, **<1%**, **<3%**, or **<5%** (the maximum allowed variant allele frequencies).

The combination of `Ancestry` and `Allele frequency` determines how variants are filtered by variant allele frequencies as follows (the default combination of values are varied in each tab):

1. **Any** for `Ancestry`

1. **All** for `Allele frequency`: No filtering on variant allele frequency.

2. Other values for `Allele frequency`: Use only variants whose maximum value from allele frequency in any population is less than or equal to the specified threshold value.

2. Other values for `Ancestry`

1. **All** for `Allele frequency`: No filtering on variant allele frequency.

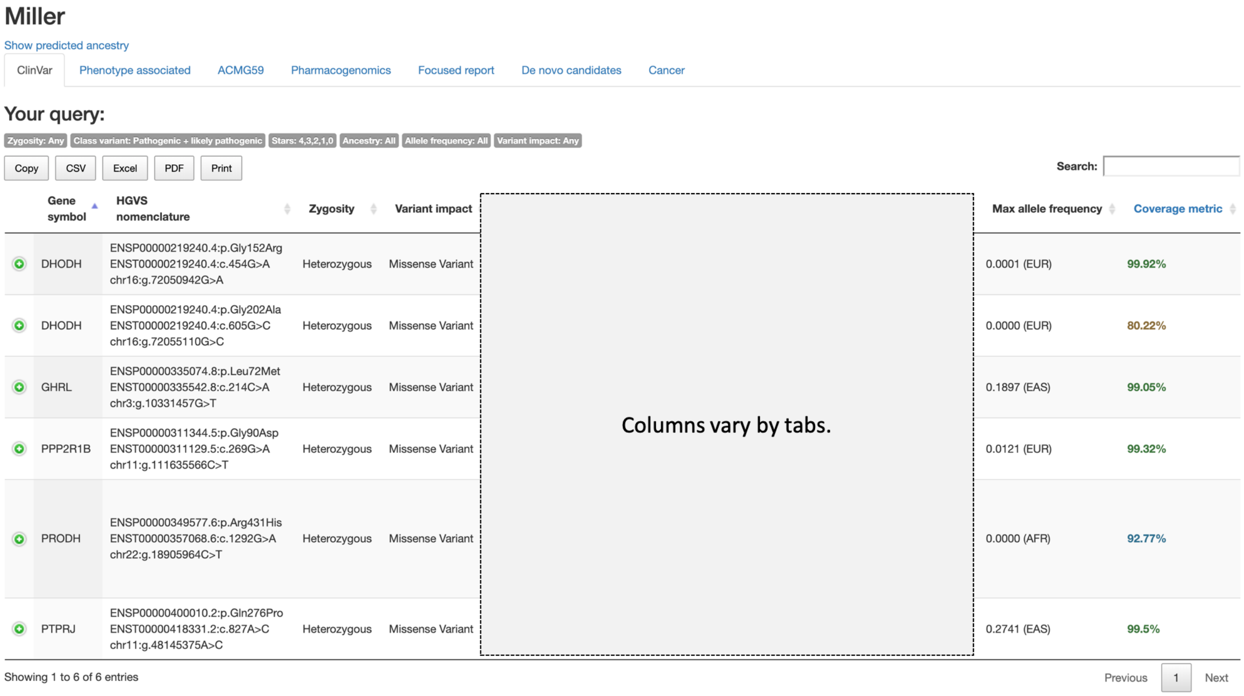
2. Other values for `Allele frequency`: Use only variants whose allele frequency in the specified population is less than or equal to the specified threshold value.

- `Calculated variant consequences`: The consequence of variants on genes or transcripts calculated using the Variant Effect Predictor. Can be either **Any** (no restriction for variant consequences), or specific consequence(s). The full list of possible consequences from VEP is available in this link. CGAR allows variant filtering on individual consequences with high, moderate, or low predicted impacts or any combination of them. The default values also varies by tabs.
- `Allele frequency based on`: Specifies how to use ancestry to determine allele frequency of variants. Can be either **Max** (default, the same as **All** in `Ancestry`), **Predicted by genotype** (sets values for `Ancestry` with majority group in the predicted ancestry composition), or **Self-identified** (only available if `Ancestry` is specified upon uploading, sets values for `Ancestry` with the specified value from uploading time).

Note: This option will be merged into `Ancestry` in future release.

4.2.2 Main report body

The main tables in the report organize variants in multiple tabs that correspond to specific analytic purpose.



Miller

Show predicted ancestry

Your query:

Zygosity: Any | Class variant: Pathogenic + likely pathogenic | Stars: 4,3,2,1,0 | Ancestry: All | Allele frequency: All | Variant impact: Any

	Gene symbol	HGVS nomenclature	Zygosity	Variant impact	Max allele frequency	Coverage metric
⊕	DHODH	ENSP00000219240.4:p.Gly152Arg ENST00000219240.4:c.454G>A chr16:g.72050942G>A	Heterozygous	Missense Variant	0.0001 (EUR)	99.92%
⊕	DHODH	ENSP00000219240.4:p.Gly202Ala ENST00000219240.4:c.605G>C chr16:g.72055110G>C	Heterozygous	Missense Variant	0.0000 (EUR)	80.22%
⊕	GHRL	ENSP00000335074.8:p.Leu72Met ENST00000335542.8:c.214C>A chr3:g.10331457G>T	Heterozygous	Missense Variant	0.1897 (EAS)	99.05%
⊕	PPP2R1B	ENSP00000311344.5:p.Gly90Asp ENST00000311129.5:c.269G>A chr11:g.111635566C>T	Heterozygous	Missense Variant	0.0121 (EUR)	99.32%
⊕	PRODH	ENSP00000349577.6:p.Arg431His ENST00000357068.6:c.1292G>A chr22:g.18905964C>T	Heterozygous	Missense Variant	0.0000 (AFR)	92.77%
⊕	PTPRJ	ENSP00000400010.2:p.Gln276Pro ENST00000418331.2:c.827A>C chr11:g.48145375A>C	Heterozygous	Missense Variant	0.2741 (EAS)	99.5%

Columns vary by tabs.

Showing 1 to 6 of 6 entries

Under the name (identifier) of the current sample (**Miller**) and links to open ancestry prediction, tabs corresponding to report sections accessible to the current user are listed.

The small gray labels under the tag `Your query:` shows the values used to generate current report (tab), to remind users of the current setting and help to change settings for subsequent analysis. In the above example, the `Zygosity` is set to **Any**, `Pathogenicity` (specific to `ClinVar` tab) is set to both **Pathogenic** and **likely pathogenic**, **All** is used for `Ancestry`, **All** is used for `Allele frequency`, and **Any** for `Calculated variant consequences` (*Variant impact* in the label).

The buttons `Copy`, `CSV`, `Excel`, `PDF`, and `Print` provide various options to save or export variants in current tab.

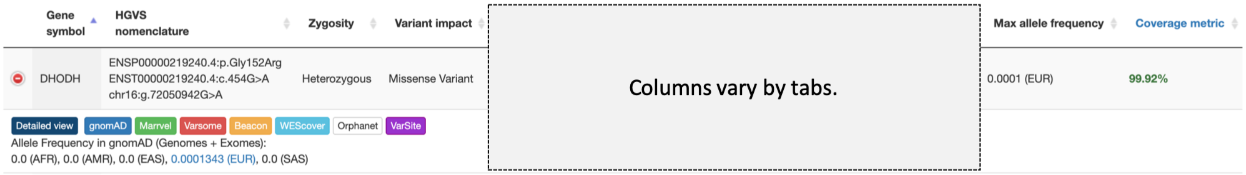
Using the `Search` box on the top right, users can quickly search for variants. Any text or value in the table can be searched here.

Columns common to tables in all tabs are as follows:

- `Gene symbol`: The official symbol for the gene by HUGO Gene Nomenclature Committee (HGNC).
- `HGVS nomenclature`: The variant representation as recommended by the Human Genome Variation Society (HGVS). The latest recommendation can be found in this link.
- `Zygosity`: Zygosity of the variant.
- `Variant impact`: The Sequence Ontology (SO) terms describing the calculated consequence of the variant by VEP.

- `Max allele frequency`: The variant allele frequence of the variant allele from Genome Aggregation Database (gnomAD), release 2.0.2. The value in this column is allele frequency from exomes in gnomAD, and always shows the maximum value from 5 population groups in gnomAD (AFR: African, EAS: East Asian, SAS: South Asian, AMR: Latino, EUR: non-Finnish European (NFE in gnomAD)). The population group of the maximum value is also shown in parentheses. If `Ancestry` is specified, the allele frequencies of the specified population will be shown (under the column heading `Allele frequency`).
- `Coverage metric`: The percentage of gnomAD exomes with minimum of 20x read depth on the variant's locus. The coverage values are graded by 4 different colors: green (99% or more exomes with 20x), blue (90% or more), brown (50% or more), and red (less than 50%).

The green plus sign left to the `Gene symbol` for each variant opens a hidden row containing links to more details on variants and to external sources.



	Gene symbol	HGVS nomenclature	Zygosity	Variant impact		Max allele frequency	Coverage metric
⊖	DHODH	ENSP00000219240.4:p.Gly152Arg ENST00000219240.4:c.454G>A chr16:g.72050942G>A	Heterozygous	Missense Variant	Columns vary by tabs.	0.0001 (EUR)	99.92%

Detailed view gnomAD Marrvel Varsome Beacon WEScover Orphanet VarSite
Allele Frequency in gnomAD (Genomes + Exomes):
0.0 (AFR), 0.0 (AMR), 0.0 (EAS), 0.0001343 (EUR), 0.0 (SAS)

The links in the hidden row are:

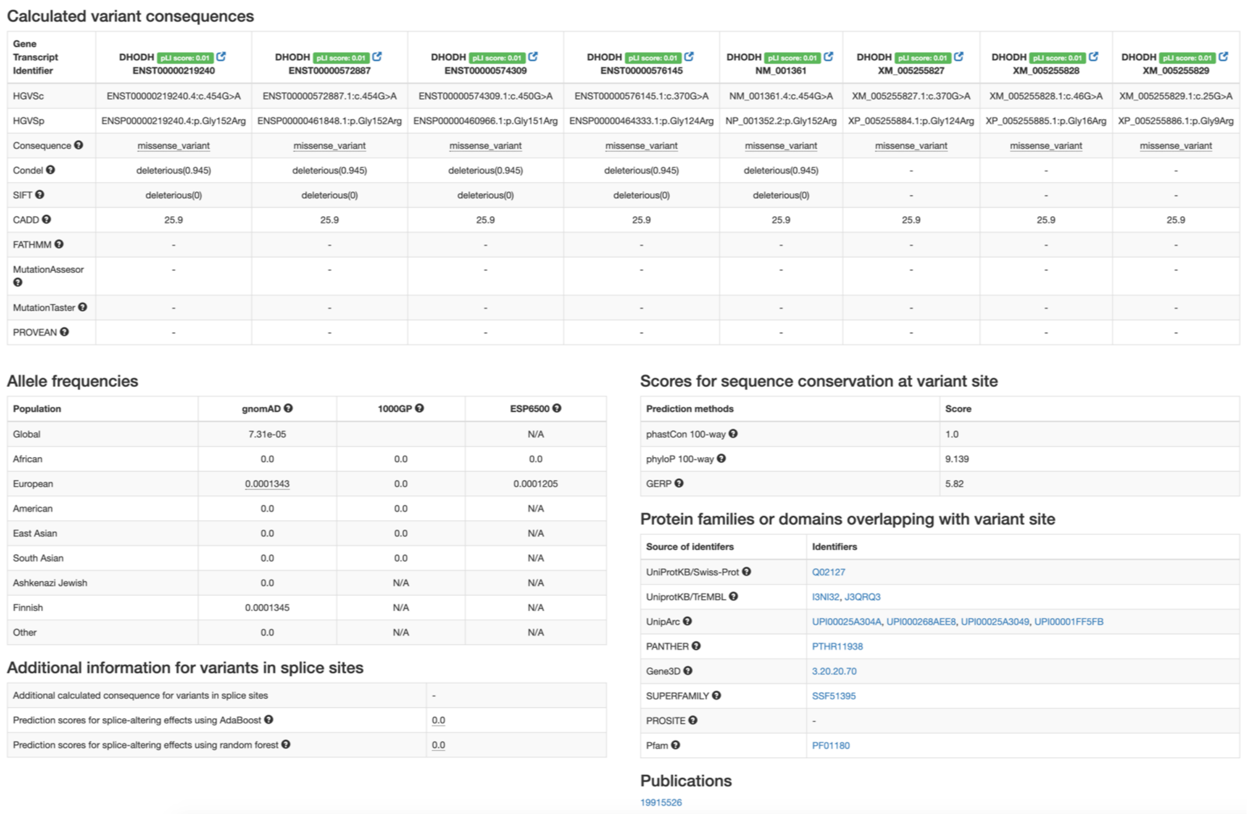
- `Detailed view`: Opens a separate window to show various detailed information about the variant.
- `gnomAD`: Links to a variant page in gnomAD, showing detailed allele frequencies and coverages.
- `Marrvel`: Links to Marrvel, a web application to prioritize human variants for rare diseases. It features ortholog search across model organisms including alignment of protein domains in ortholog proteins.
- `Varsome`: Links to Varsom, a community-based application of variant interpretation. It provides a variety of genetic and clinically relevant information for the variant.
- `Beacon`: Links to GA4GH Beacon Network, a search engine of genetic variants across various institutes and organizations.
- `WEScover`: Links to WEScover to investigate breadth of coverage of a gene over exomes in 1000 Genomes Project. In contrast to `Coverage metric` that provides locus-specific value, it provides a gene-centric value.
- (Restricted) `Orphanet`: Opens a new window with lists of phenotypes associated to the gene.
- `VarSite`: Links to residue report for the variant by VarSite. VarSite features potential effects of the variant on protein 3D structure.

Also, the hidden row shows variant allele frequency in gnomAD exomes for each of 5 population groups (AFR, AMR, EAS, EUR (NFE), and SAS).

Besides the above common columns, the main table may contain additional columns depending on the current tab. The additional columns specific to each tab will be explained in subsections for the tab.

4.2.3 Variant details

Each row on the main table only shows the essential information as well as the calculated consequence of most severity. However, the calculated consequence of a variant can change depending on the transcript used for prediction Also, more information on the variant such as the predicted pathogenicity score, allele frequencies in multiple population-scale data, or protein families or domains affected by the variant can be very useful to interpret the variant. The `Detailed view` on the main table opens a new window containing the following information.



Calculated variant consequences

Gene Transcript Identifier	DHODH pLI score: 0.01 ENST00000219240	DHODH pLI score: 0.01 ENST00000572887	DHODH pLI score: 0.01 ENST00000574309	DHODH pLI score: 0.01 ENST00000576145	DHODH pLI score: 0.01 NM_001361	DHODH pLI score: 0.01 XM_005255827	DHODH pLI score: 0.01 XM_005255828	DHODH pLI score: 0.01 XM_005255829
HGVSc	ENST00000219240.4:c.454G>A	ENST00000572887.1:c.454G>A	ENST00000574309.1:c.450G>A	ENST00000576145.1:c.370G>A	NM_001361.4:c.454G>A	XM_005255827.1:c.370G>A	XM_005255828.1:c.46G>A	XM_005255829.1:c.25G>A
HGVSp	ENSP00000219240.4:p.Gly152Arg	ENSP00000461848.1:p.Gly152Arg	ENSP00000460966.1:p.Gly151Arg	ENSP00000464333.1:p.Gly124Arg	NP_001352.2:p.Gly152Arg	XP_005255884.1:p.Gly124Arg	XP_005255885.1:p.Gly16Arg	XP_005255886.1:p.Gly9Arg
Consequence	missense_variant	missense_variant	missense_variant	missense_variant	missense_variant	missense_variant	missense_variant	missense_variant
Condel	deleterious(0.945)	deleterious(0.945)	deleterious(0.945)	deleterious(0.945)	deleterious(0.945)	-	-	-
SIFT	deleterious(0)	deleterious(0)	deleterious(0)	deleterious(0)	deleterious(0)	-	-	-
CADD	25.9	25.9	25.9	25.9	25.9	25.9	25.9	25.9
FATHMM	-	-	-	-	-	-	-	-
MutationAssesor	-	-	-	-	-	-	-	-
MutationTaster	-	-	-	-	-	-	-	-
PROVEAN	-	-	-	-	-	-	-	-

Allele frequencies

Population	gnomAD	1000GP	ESP6500
Global	7.31e-05		N/A
African	0.0	0.0	0.0
European	0.0001343	0.0	0.0001205
American	0.0	0.0	N/A
East Asian	0.0	0.0	N/A
South Asian	0.0	0.0	N/A
Ashkenazi Jewish	0.0	N/A	N/A
Finnish	0.0001345	N/A	N/A
Other	0.0	N/A	N/A

Additional information for variants in splice sites

Additional calculated consequence for variants in splice sites	-
Prediction scores for splice-altering effects using AdaBoost	0.0
Prediction scores for splice-altering effects using random forest	0.0

Scores for sequence conservation at variant site

Prediction methods	Score
phastCon 100-way	1.0
phyloP 100-way	9.139
GERP	5.82

Protein families or domains overlapping with variant site

Source of identifers	Identifiers
UniProtKB/Swiss-Prot	Q02127
UniprotKB/TrEMBL	I3NI32, J3QRQ3
UnipArc	UPI00025A304A, UPI000268AEE8, UPI00025A3049, UPI00001FF5FB
PANTHER	PTHR11938
Gene3D	3.20.20.70
SUPERFAMILY	SSF51395
PROSITE	-
Pfam	PF01180

Publications

19915526

- `Calculated variant consequences`: variant consequences predicted with VEP and pathogenicity scores calculated by multiple methods are organized by each gene or transcript on the variant's location.
  - Includes a score of gene's tolerance to loss-of-function variant (ExACpLI score **'Re>'_**).
  - For each transcript, the representation of variant for cNDA or protein is shown.
  - Pathogenicity scores calculated by Condel, SIFT, CADD, FATHMM, *MutationAssessor <http://mutationassessor.org/>*, MutationTaster, and PROVEAN.
- `Allele frequencies`: variant allele frequencies from 3 population-scale data (gnomAD, 1000 Genomes Project, and the NHLBI Exome Sequencing Project).
  - Allele frequencies are calculated per each population group.
- `Additional information for variants in splice sites`: prediction scores for splicing-altering effects for variants in splice sites.
- `Scores for sequence conservation at variant site`: scores for sequence conservation from multiple sequence alignment of various species. Scores from **'phastCon <>'_**, **'phyloP <>'_**, and **'GERP <>'_** are provided.
- `Protein families or domains overlapping with variant site`: lists protein or protein domain identifiers that overlaps with variant's position.
- `Publications`: list of PubMed identifiers for publications that cite the variant.

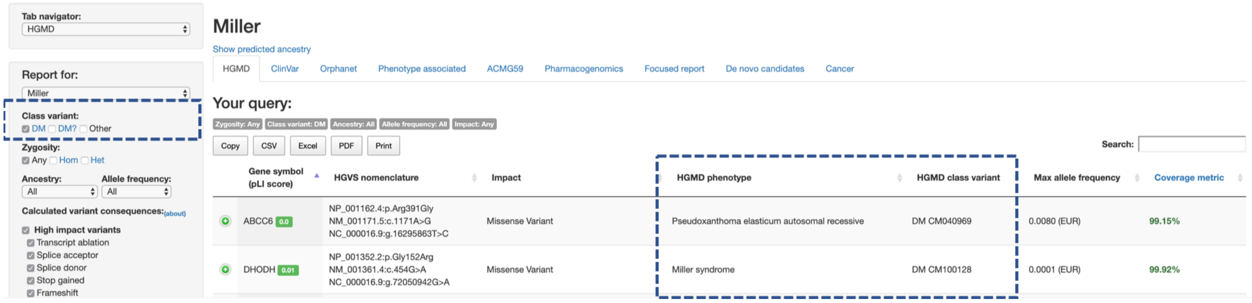
4.3 HGMD variants

> **Warning:** The contents for this section is available only for users licensed from the Human Gene Mutation Database (HGMD). Please note that the registration to CGAR does not automatically grant access to HGMD. Users need to obtain license from HGMD, separately.

The Human Gene Mutation Database (HGMD®) maintains a collection of known (published) genes associated with human inherited disease. HGMD® categorizes its data based on level of association with disease. Among those, variants with the most potential clinical implication would be:

- Disease causing mutation? (*DM?*): variant reported to be disease-causing in the report, yet author indicated some degree of doubt, or subsequence evidence caused the deleteriousness of the variant into question.
- Disease causing mutation (*DM*): pathological variant reported to be disease-causing.

`HGMD` tab in CGAR shows which variants in the current sample have matches from HGMD®.



Tab navigator: HGMD

Report for: Miller

Class variant: ☑ DM □ DM? □ Other

Zygosity: ☑ Any □ Hom □ Het

Ancestry: All

Allele frequency: All

Calculated variant consequences: (about)

☑ High impact variants
- ☑ Transcript ablation
- ☑ Splice acceptor
- ☑ Splice donor
- ☑ Stop gained
- ☑ Frameshift

Your query:

Zygosity: Any | Class variant: DM | Ancestry: All | Allele frequency: All | Impact: Any

Gene symbol (pLI score)	HGVS nomenclature	Impact	HGMD phenotype	HGMD class variant	Max allele frequency	Coverage metric
ABCC6 0.0	NP_001162.4:p.Arg391Gly NM_001171.5:c.1171A>G NC_000016.9:g.16295863T>C	Missense Variant	Pseudoxanthoma elasticum autosomal recessive	DM CM040969	0.0080 (EUR)	99.15%
DHODH 0.01	NP_001352.2:p.Gly152Arg NM_001361.4:c.454G>A NC_000016.9:g.72050942G>A	Missense Variant	Miller syndrome	DM CM100128	0.0001 (EUR)	99.92%

The `Class variant` on the side menu controls matches to which categories in HGMD will be shown (multiple choices are allowed).

- **DM**: show matches to *DM* variants.
- **DM?**: show matches to *DM?* variants.
- **Other**: show matches to other categories than **DM** or **DM?**.

By default, `HGMD` tab shows only variants that are categorized as **DM** in HGMD®, regardless of variant genotype, allele frequency, or variant consequences.

- `Zygosity`: **Any**
- `Class variant`: **DM**
- `Ancestry`: **All**
- `Allele frequency`: **All**
- `Calculated variant consequences`: **Any**

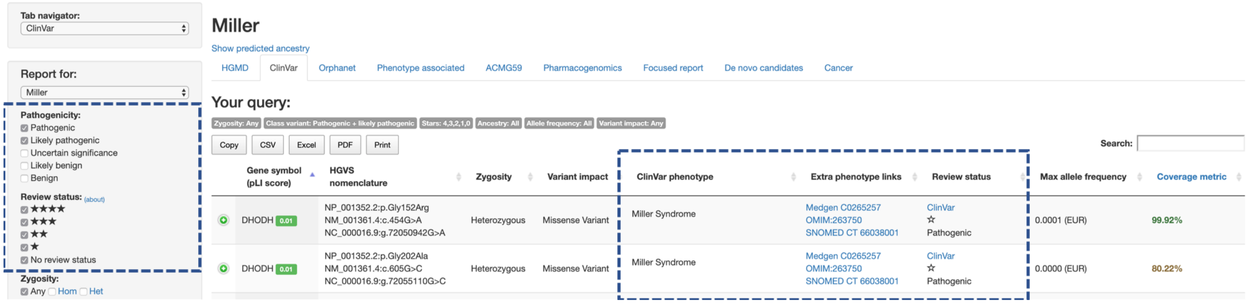
For each variant, the main table on the right shows the disease phenotypes associated (or caused) by the variant (`HGMD phenotype` column) and categories and identifiers in HGMD® (`HGMD class variant`).

4.4 ClinVar variants

> ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication about

> the relationships asserted between human variation and observed health status, and the history of that interpretation. (From ClinVar introduction)

ClinVar is a public database which archives reports on relationship between human genomic variants and phenotypes. Like in HGMD tab, ClinVar tab in CGAR shows variants in the current sample that were previously submitted to ClinVar and their interpretations.



Tab navigator: ClinVar

Report for: Miller

Pathogenicity:
☑ Pathogenic
☑ Likely pathogenic
□ Uncertain significance
□ Likely benign
□ Benign

Review status: (about)
☑ ★★★★
☑ ★★★
☑ ★★
☑ ★
☑ No review status

Zygosity:
☑ Any □ Hom □ Het

Miller

Show predicted ancestry

Your query:

Zygosity: Any | Class variant: Pathogenic + likely pathogenic | Stars: 4,3,2,1,0 | Ancestry: All | Allele frequency: All | Variant impact: Any

Search:

Gene symbol (pLI score)	HGVS nomenclature	Zygosity	Variant impact	ClinVar phenotype	Extra phenotype links	Review status	Max allele frequency	Coverage metric
DHODH 0.01	NP_001352.2:p.Gly152Arg NM_001361.4:c.454G>A NC_000016.9:g.72050942G>A	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0001 (EUR)	99.92%
DHODH 0.01	NP_001352.2:p.Gly202Ala NM_001361.4:c.605G>C NC_000016.9:g.72055110G>C	Heterozygous	Missense Variant	Miller Syndrome	Medgen C0265257 OMIM:263750 SNOMED CT 66038001	ClinVar ☆ Pathogenic	0.0000 (EUR)	80.22%

Users can filter variants by their reported clinical significance in ClinVar (Pathogenicity) and the level of review supporting the clinical significance (Review status). In other words, CGAR can either be set to show variants which had been submitted to ClinVar with multiple evidence of its pathogenicity, or be set to show any variants which had ever been submitted to ClinVar - by controling Pathogenicity and Review status in combination of other controls in side menu.

- Pathogenicity can take any number of values from the 5 levels: **Pathogenic** (most deleterious) / **Likely pathogenic / Uncertain significance / Likely benign / Benign** (least deleterious).
- Review status shows the number of gold starts which corresponds to each of the review statuses in ClinVar (link) (multiple values are allowed).

In the main table, CGAR shows the phenotype reported with the variant (ClinVar phenotype), links to phenotype databases (Extra phenotype links), and brief summary of variant's status in ClinVar (Review status).

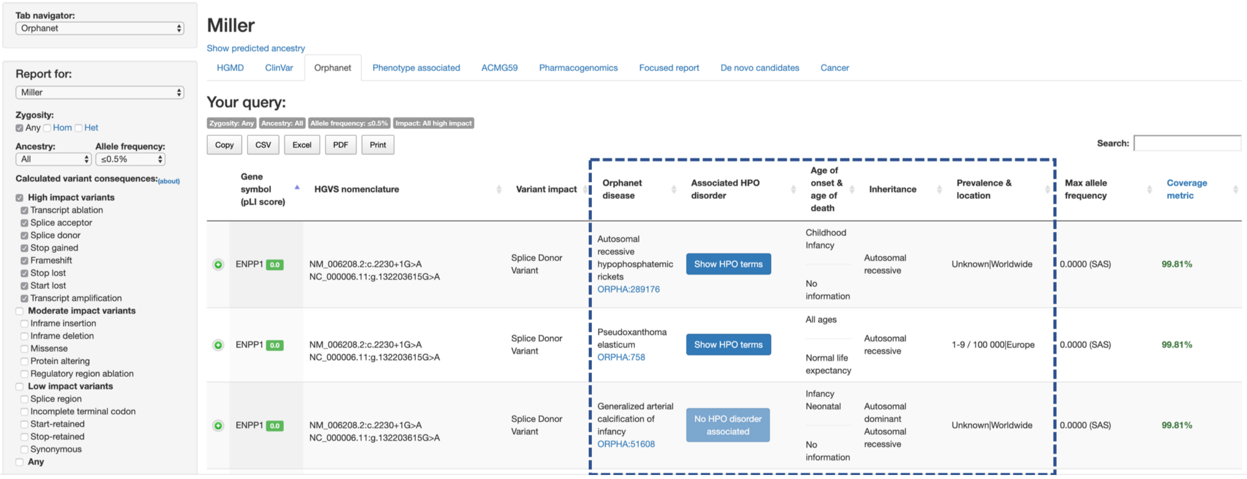
- Extra phenotype links provide links to MedGen, Online Mendelian Inheritance in Men (OMIM), Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), Human Pathology Ontology (HPO), and OrphaNet. If the variant is linked to multiple phenotypes, a set of links is provided for each phenotype.
- Review status column represents slightly different contents from side menu. In the main table, it contains:
  - Graphical representation of review status (stars)
  - Aggregated interpretation of the variant in ClinVar
  - Link to the variant's record in Clinvar, where all submissions to ClinVar regarding the variant can be viewed.

Note: CGAR tries to maintain the most up-to-date data from all resources. However, there can be some discrepancy between what is available in ClinVAR and CGAR.

4.5 Genes associated with rare phenotype

Warning: The contents for this section is available only for users licensed from OrphaData.org. Please note that the registration to CGAR does not automatically grant access to data from OrphaData.org. Users need to obtain license from OrphaData.org, separately.

OrphaData.org provides large data sets related to rare disease such as genes related to rare diseases or epidemiological data. CGAR uses these data to find variants on genes associated with rare disease phenotypes. Here, no additional input from users is required (other than common paramteres such as allele frequency).



Miller

Zygosity: ☑ Any □ Hom □ Het

Ancestry: All

Allele frequency: ≤0.5%

Calculated variant consequences:(about)

- ☑ High impact variants
  - ☑ Transcript ablation
  - ☑ Splice acceptor
  - ☑ Splice donor
  - ☑ Stop gained
  - ☑ Frameshift
  - ☑ Stop lost
  - ☑ Start lost
  - ☑ Transcript amplification
- □ Moderate impact variants
  - □ Inframe insertion
  - □ Inframe deletion
  - □ Missense
  - □ Protein altering
  - □ Regulatory region ablation
- □ Low impact variants
  - □ Splice region
  - □ Incomplete terminal codon
  - □ Start-retained
  - □ Stop-retained
  - □ Synonymous
- □ Any

Your query:

Zygosity: Any | Ancestry: All | Allele frequency: ≤0.5% | Impact: All high impact

	Gene symbol (pLI score)	HGVS nomenclature	Variant impact	Orphanet disease	Associated HPO disorder	Age of onset & age of death	Inheritance	Prevalence & location	Max allele frequency	Coverage metric
⊕	ENPP1 0.0	NM_006208.2:c.2230+1G>A NC_000006.11:g.132203615G>A	Splice Donor Variant	Autosomal recessive hypophosphatemic rickets ORPHA:289176	Show HPO terms	Childhood Infancy No information	Autosomal recessive	Unknown|Worldwide	0.0000 (SAS)	99.81%
⊕	ENPP1 0.0	NM_006208.2:c.2230+1G>A NC_000006.11:g.132203615G>A	Splice Donor Variant	Pseudoxanthoma elasticum ORPHA:758	Show HPO terms	All ages Normal life expectancy	Autosomal recessive	1-9 / 100 000|Europe	0.0000 (SAS)	99.81%
⊕	ENPP1 0.0	NM_006208.2:c.2230+1G>A NC_000006.11:g.132203615G>A	Splice Donor Variant	Generalized arterial calcification of infancy ORPHA:51608	No HPO disorder associated	Infancy Neonatal No information	Autosomal dominant Autosomal recessive	Unknown|Worldwide	0.0000 (SAS)	99.81%

In the main table, 5 columns show informations on rare phenotype associated with the variant (through gene), to enhance variant interpretation.

- `Orphanet disease`: the name and identifier of phenotype used in OrphaNet with link.
- `Associated HPO disorder`: opens a table listing every Human Phenotype Ontology terms associated with the phenotype, a standardized terms for human disease abnomalities. Useful to correlate with other data using HPO terms.
- `Age of onset & age of death`: age of onset and/or death for the phenotype. Could be used to further filter out irrelevant variants. For example, in a sample where the first symptom occured at 20s, phenotypes with age of onset as **Infancy** can be disregarded.
- `Inheritance`: mode of inheritance for the phenotype. Also could be used to filter out irrelevant variants. If a subspected phenotype is **Autosomal recessive**, heterozygous variants can be less likely the target.
- `Prevalence & location`: phenotype frequency by geological locations. Also could be used for further filtering variant, especially in combination with the ancestry of the sample.

4.6 Putative phenotype-associated genes

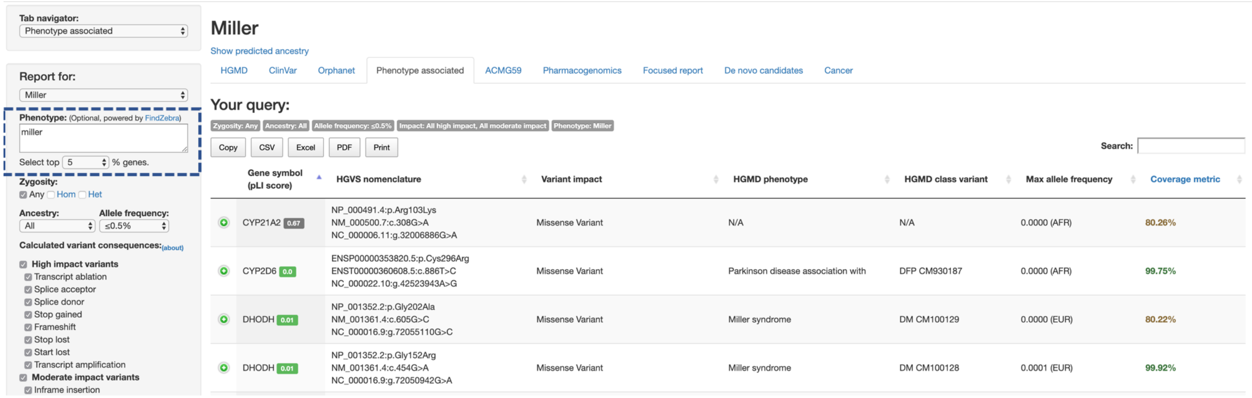
For uses who want to focus on variants on genes associated to specific phenotypes or symptoms, but not yet certain about which genes to look for, CGAR provides a method to find candidate variants with phenotypes.

First, CGAR sends the free-form text input from user describing phenotypes to a machine-learning based text search engine in FindZebra.

Then, it retrieves information from a corpus of documents consisting of more than 36,000 entries from curated sources such as OMIM, Genetic and Rare Diseases Information Center, and OrphaNet, and ranks genes according to gene-specific scores for input phenotype.

From the ranked list of genes, CGAR selects the highest scored genes within user-specified proportion, and shows variants on the genes satisfying conditions on allele frequencies or consequences.

The advantage of using FindZebra is that users are not required to provide standardized ontology terms such as Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) or Human Pathology Ontology (HPO).



Tab navigator: Phenotype associated

Report for: Miller

Phenotype: (Optional, powered by FindZebra)
miller
Select top 5 % genes.

Zygosity:
☑ Any □ Hom □ Het

Ancestry: All
Allele frequency: ≤0.5%

Calculated variant consequences: (about)

☑ High impact variants
☑ Transcript ablation
☑ Splice acceptor
☑ Splice donor
☑ Stop gained
☑ Frameshift
☑ Stop lost
☑ Start lost
☑ Transcript amplification
☑ Moderate impact variants
☑ Inframe insertion

Miller

Your query:
Zygosity: Any | Ancestry: All | Allele frequency: ≤0.5% | Impact: All high impact, All moderate impact | Phenotype: Miller

Search:

Gene symbol (pLI score)	HGVS nomenclature	Variant impact	HGMD phenotype	HGMD class variant	Max allele frequency	Coverage metric
CYP21A2 0.67	NP_000491.4:p.Arg103Lys NM_000500.7:c.308G>A NC_000006.11:g.32006886G>A	Missense Variant	N/A	N/A	0.0000 (AFR)	80.26%
CYP2D6 0.0	ENSP00000353820.5:p.Cys296Arg ENST00000360608.5:c.886T>C NC_000022.10:g.42523943A>G	Missense Variant	Parkinson disease association with	DFP CM930187	0.0000 (AFR)	99.75%
DHODH 0.01	NP_001352.2:p.Gly202Ala NM_001361.4:c.605G>C NC_000016.9:g.72055110G>C	Missense Variant	Miller syndrome	DM CM100129	0.0000 (EUR)	80.22%
DHODH 0.01	NP_001352.2:p.Gly152Arg NM_001361.4:c.454G>A NC_000016.9:g.72050942G>A	Missense Variant	Miller syndrome	DM CM100128	0.0001 (EUR)	99.92%

Here, users are only required to provide list of phenotypes (or descriptions) in the text box (`Phenotype`), and specify the percentage of high-scored genes you want to use from values **1** (use genes in top 1%), **5**, **10**, **25**, **50**, **100** (use all genes).

4.7 Secondary findings

> **Caution:** The American College of Medical Genetics and Genomics (ACMG) discourages the use of genes in ACMG SF v2.0 for purposes other than reporting incidental findings after *clinical sequencing*. Therefore the use of variants in this section outside of *clinical sequencing* is not recommended. (based on ACMG statement on the use of ACMG secondary findings recommendations for general population screening)

The ACMG has published recommendations for reporting incidental findings in the exons of certain genes (ACMG SF v2.0). The recommendation specifies, for clinical sequencing of each phenotype, genes and variants to report as secondary findings.

CGAR list variants on any of genes listed in ACMG SF v2.0, along with which phenotype the gene was recommended for (`ACMG phenotype` column for phenotype and `MIM ID` for MIM identifyer of the phenotype). Given the lack of the knowledge on phenotype or the purpose of sequencing, CGAR will always list variants on any of genes listed in ACMG SF v2.0. However, users are recommended to use this information only in a way compliant to ACMG recommendation and statement.

Additionally, CGAR shows phenotype and variant category from HGMD® (requires license) (columns `HGMD phenotype` and `HGMD class variant`).

4.8 Pharmacogenomic variants

The `Pharmacogenomics` tab in CGAR reports variants or genes that could modulate drug metabolism or interfere with responses to drugs, summarized as in the following three categories:

1. Variants specified in the dosage guidelines by the Clinical Pharmacogenetics Implementation Consortium (CPIC).

CPIC publishes guidelines to help clinicians to optimize drug therapy using genetic test results. The guidelines are published for pairs of drugs and genes, detailing drug therapy recommendations based on genotypes. CGAR lists all variants that were used in any of the guidelines, as a reminder to users (especially, clinicians) that some adjustment on drug prescription may be required.

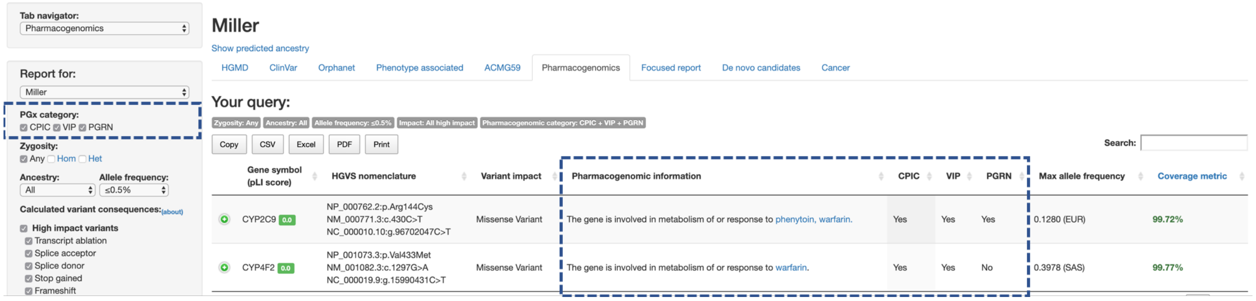
> **Caution:** The guidelines are based on allele types of genes, which may be defined by combination of genotypes in multiple loci. CGAR currently does not take multiple variants into consideration, and depending on quality of variant data, genotype of some loci might not be available, preventing the accurate application of guidelines. Thus, CGAR results in this section is best regarded as a reminder that some adjustments may be required. The actual adjustment should be dune after careful evaluation of alleles of the gene.

2. Variants on genes from curated collection of genes Very Important Pharmacogene (VIP), playing important roles in drug metabolism or response.

The Pharmacogenomics Knowledge Base (PharmGKB) is a knowledge resource for the impact of human genetic variation on drug responses. One of most interesting feature in PharmGKB is the list of *Very Important Pharmacogenes* (VIPs), genes involved in metabolism of or response to drug(s). Thus, variants on these genes could have pharmacogenomic impact.

3. Variants on genes selected for the investigation of drug response and genetic variation by the Phamacogenomics Research Network (PGRN).

PGRN developed a custom-capture panel of 84 genes associated with pharmacophenotypes, PGRNseq. Along with VIPs, these genes could be used to identify variants of pharmacogenomic impact.



PGx category:
☑ CPIC ☑ VIP ☑ PGRN

Zygosity:
☑ Any □ Hom □ Het

Ancestry: All

Allele frequency: ≤0.5%

Calculated variant consequences: (about)

☑ High impact variants
☑ Transcript ablation
☑ Splice acceptor
☑ Splice donor
☑ Stop gained
☑ Frameshift

Miller

Your query:

Zygosity: Any | Ancestry: All | Allele frequency: ≤0.5% | Impact: All high impact | Pharmacogenomic category: CPIC + VIP + PGRN

Gene symbol (pLI score)	HGVS nomenclature	Variant impact	Pharmacogenomic information	CPIC	VIP	PGRN	Max allele frequency	Coverage metric
CYP2C9 0.0	NP_000762.2:p.Arg144Cys NM_000771.3:c.430C>T NC_000010.10:g.96702047C>T	Missense Variant	The gene is involved in metabolism of or response to phenytoin, warfarin.	Yes	Yes	Yes	0.1280 (EUR)	99.72%
CYP4F2 0.0	NP_001073.3:p.Val433Met NM_001082.3:c.1297G>A NC_000019.9:g.15990431C>T	Missense Variant	The gene is involved in metabolism of or response to warfarin.	Yes	Yes	No	0.3978 (SAS)	99.77%

The PGx category on the side menu allows users to select which set of variants or genes would be used to identify variants with potential pharmacogenomic impact (multiple choices are allowed).

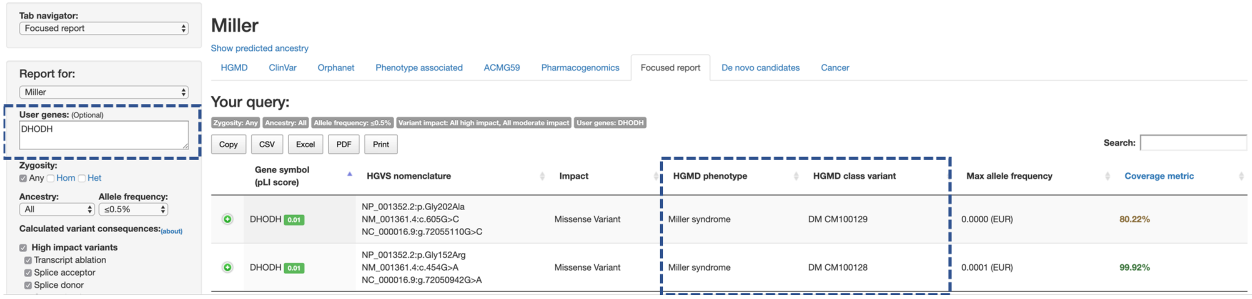
- **CPIC**: shows variants that matches to the variants in the CPIC guidelines.
- **VIP**: shows variants on genes listed as VIPs.
- **PGRN**: shows variants on genes included in PGRNseq panel.

By default, CGAR is set to use all of the above. For **VIP** and **PGRN**, rare and high impact variants are selected.

- Zygosity: **Any**
- Ancestry: **All**
- Allele frequency: **<0.5%**
- Calculated variant consequences: **High impact variants**
- Allele frequency based on: **Max**

The columns CPIC, VIP, and PGRN in the main table shows the inclusion criteria for the variant. For variants used in CPIC guideline, Phamacogenomic information columns shows list of drugs along with links to the CPIC guidelines. For variants on genes in VIPs or PGRNseq, users are recommended to check repective documents.

4.9 User-defined set of genes



Tab navigator: Focused report

Report for: Miller

User genes: (Optional) DHODH

Zygosity: ☑ Any ☐ Hom ☐ Het

Ancestry: All Allele frequency: ≤0.5%

Calculated variant consequences:(about)

☑ High impact variants
☑ Transcript ablation
☑ Splice acceptor
☑ Splice donor

Miller

Your query: Zygosity: Any | Ancestry: All | Allele frequency: ≤0.5% | Variant impact: All high impact, All moderate impact | User genes: DHODH

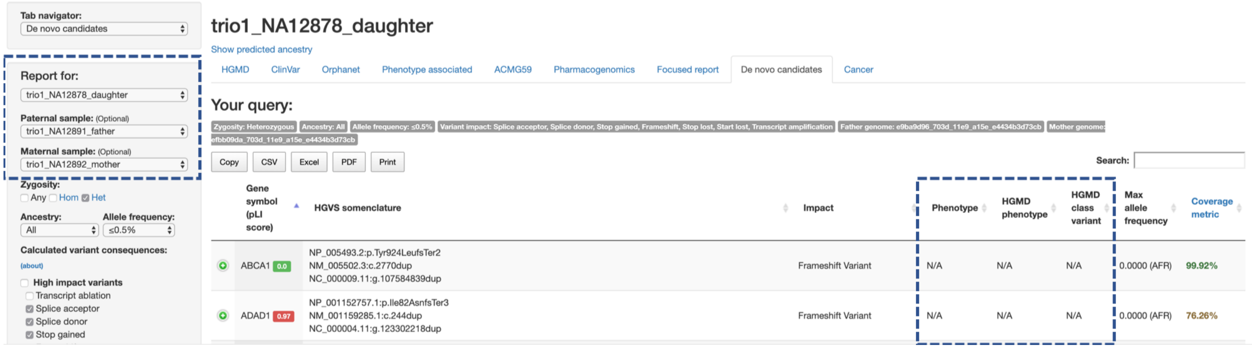
Search:

	Gene symbol (pLI score)	HGVS nomenclature	Impact	HGMD phenotype	HGMD class variant	Max allele frequency	Coverage metric
⊕	DHODH 0.01	NP_001352.2:p.Gly202Ala NM_001361.4:c.605G>C NC_000016.9:g.72055110G>C	Missense Variant	Miller syndrome	DM CM100129	0.0000 (EUR)	80.22%
⊕	DHODH 0.01	NP_001352.2:p.Gly152Arg NM_001361.4:c.454G>A NC_000016.9:g.72050942G>A	Missense Variant	Miller syndrome	DM CM100128	0.0001 (EUR)	99.92%

In cases where specific gene(s) need to be investigated, users can simply type in gene symbo(s) into text box (`User genes`). CGAR will show any variants on the gene(s), provided that the variant matches other conditions as well. Additionally, CGAR shows phenotype and variant category from HGMD® (requires license) (columns `HGMD phenotype` and `HGMD class variant`).

4.10 De novo variant candidates

In cases where variants from both parents are available with an index case, CGAR supports simple *de novo* variant analysis by identifying heterozygous variants from index case that are not found in both parents.



Tab navigator: De novo candidates

Report for: trio1_NA12878_daughter

Paternal sample: (Optional) trio1_NA12891_father

Maternal sample: (Optional) trio1_NA12892_mother

Zygosity: ☐ Any ☐ Hom ☑ Het

Ancestry: All Allele frequency: ≤0.5%

Calculated variant consequences: (about)

☐ High impact variants
☐ Transcript ablation
☑ Splice acceptor
☑ Splice donor
☑ Stop gained

trio1_NA12878_daughter

Your query: Zygosity: Heterozygous | Ancestry: All | Allele frequency: ≤0.5% | Variant impact: Splice acceptor, Splice donor, Stop gained, Frameshift, Stop lost, Start lost, Transcript amplification | Father genome: e9ba9d96_703d_11e9_a15e_e4434b3d73cb | Mother genome: efbb09da_703d_11e9_a15e_e4434b3d73cb

Search:

	Gene symbol (pLI score)	HGVS somenclature	Impact	Phenotype	HGMD phenotype	HGMD class variant	Max allele frequency	Coverage metric
⊕	ABCA1 0.0	NP_005493.2:p.Tyr924LeufsTer2 NM_005502.3:c.2770dup NC_000009.11:g.107584839dup	Frameshift Variant	N/A	N/A	N/A	0.0000 (AFR)	99.92%
⊕	ADAD1 0.97	NP_001152757.1:p.Ile82AsnfsTer3 NM_001159285.1:c.244dup NC_000004.11:g.123302218dup	Frameshift Variant	N/A	N/A	N/A	0.0000 (AFR)	76.26%

On the side menu, index case (if different from current one) can be selected from `Report for:` dropdown list. Parental samples are specified by `Paternal sample` and `Maternal sample`, respectively.

On the resulting table, if the *de novo* variant candidate is also found in denovo-db, a collection of germline *de novo* variants identified in human, the phenotype from denovo-db (`Phenotype`, the phenotype for the original family study) is also shown. Additionally, CGAR shows phenotype and variant category from HGMD® (requires license) (columns `HGMD phenotype` and `HGMD class variant`).

4.11 Cancer-associated genes and variants

The `Cancer` tab in CGAR reports variants that could have implications in cancer, based on the following categories:

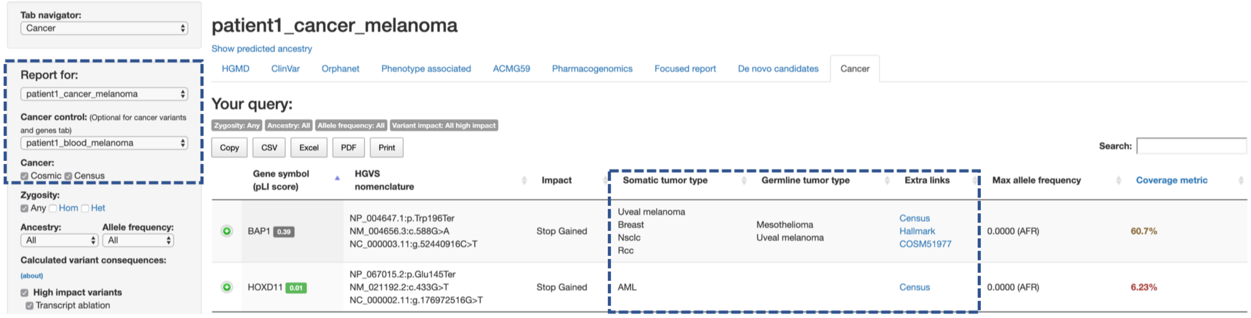
1. Variants previously found in the Catalog of Somatic Mutations in Cancer (COSMIC), an expert-curated database of somatic mutations.

COSMIC maintains large collection of expert-curated somatic mutation information relating to human cancers. From user sample, CGAR lists all variants that had matches in COSMIC. If the sample is from cancer case, the variants previously found in COSMIC may worth further investigation, especially when the cancer types in COSMIC matches with the sample.

2. Variants on genes listed in Cancer Gene Census (CGC), genes with mutations causally implicated in cancer.

CGC catalogs genes with mutations that have been causally implicated in cancer. Genes in CGC are divided into two groups based on the level of documented evidence for oncogenic functionality of genes. CGAR list variants on genes which are grouped as tier 1 in CGC: genes with documented activity relevant to cancer, evidence of mutations in cancer which can change the activity of gene product towards promoting oncogenic transformation.

Along with the above categories, if a matched control sample is avaialble, any variants that were found in both control and the case will be excluded.



	Gene symbol (pLI score)	HGVS nomenclature	Impact	Somatic tumor type	Germline tumor type	Extra links	Max allele frequency	Coverage metric
⊕	BAP1 0.39	NP_004647.1:p.Trp196Ter NM_004656.3:c.588G>A NC_000003.11:g.52440916C>T	Stop Gained	Uveal melanoma Breast Nsclc Rcc	Mesothelioma Uveal melanoma	Census Hallmark COSM51977	0.0000 (AFR)	60.7%
⊕	HOXD11 0.01	NP_067015.2:p.Glu145Ter NM_021192.2:c.433G>T NC_000002.11:g.176972516G>T	Stop Gained	AML		Census	0.0000 (AFR)	6.23%

Checkboxes under `Cancer` on the side menu allows users to select how to identify variants with potential association with cancer (multiple choices are allowed).

The `Cancer control` dropdown list under the main sample `Report for:` allows users to specify the matched control sample.

CHAPTER 5

Running CGAR locally

5.1 Requirements

- Docker: any versions of Docker that support docker compose file version 2 and `docker-compose`.
- CGAR distribution: contains Docker image construction files and source code. Plase contact us for getting the file.
- Free disk space of 700MB

5.2 Installation

First, unpack CGAR distribution file (`cgar-dist.tar.gz`):

```
$ cd /path/to/distribution/file
$ tar -xzvf [filename]
$ cd cgar-dist
```

Next, build Docker images for CGAR (two images for CGAR database and web application, respectively):

```
# build all images with docker-compose
$ cd /path/to/docker-compose.yaml
$ docker-compose build
# see if images appear in the list of available Docker images
$ docker images
```

When successful, Docker images `cgar-dist_cgar_web` and `cgar-dist_cgar_db` will be created.

Then, adjust application settings for local environment in `docker-compose.yml`:

```
version: '2'
services:
  cgar_db:
```

(continues on next page)

(continued from previous page)

```
    build: ./cgar_db
    volumes:
      - ./cgar_web/db:/docker-entrypoint-initdb.d
      - /path/to/save/mysqldata:/var/lib/mysql ## [ UPDATE LOCAL PATH ]
      - /path/to/save/dbkey:/var/lib/mysql-keyring:rw ## [ UPDATE LOCAL PATH ]
    environment:
      - MYSQL_ROOT_PASSWORD=*your_password* ## [ UPDATE THIS ]
      ...
      - MYSQL_PASSWORD=*temp* ## [ UPDATE THIS ]
    command: --early-plugin-load=keyring_file.so
  cgar_web:
    build: ./cgar_web
    volumes:
      - /path/to/vep/cache:/cache:rw ## [ UPDATE LOCAL PATH ]
    ports:
      - "CGAR_PORT:80" ## [ CHANGE PORT NUMBER ]
    expose:
      - "CGAR_PORT" ## [ CHANGE PORT NUMBER ]
    depends_on:
      - cgar_db
    environment:
      ...
      - CGAR_DB_PASSWORD=*temp*    ## [ UPDATE THIS ]
      - VEP_CACHE_DIR=/cache
      - LOW_MEM=1         ## [ UPDATE THIS ]
      - VEP_CACHE_VER=92  ## [ UPDATE THIS ]
      - COSMIC_USER=some.email@institute.edu ## [ UPDATE THIS ]
      - COSMIC_PW=password ## [ UPDATE THIS ]
...
```

Settings to be customized in `cgar_db` section:

- `/path/to/save/mysqldata`: persistent storage for mySQL database tables that will be used in CGAR. Should point to local folder where Docker container can read and write, and have enough disk space for database files (grows over time). If the folder does not exist already, `dockder-compose` will create it for you.
- `/path/to/save/dbkey`: persistent location for mySQL encryption key. Should point to **an existing** local folder where Docker container can read and **write**.
- `MYSQL_ROOT_PASSWORD`: this is administrator password for the entire mySQL database in CGAR. Not directly used by CGAR.
- `MYSQL_PASSWORD`: mySQL password for CGAR application only. Should also be the same as `CGAR_DB_PASSWORD` under `cgar_web` section.

Settings to be customized in `cgar_web` section:

- `/path/to/vep/cache`: persistent location to store annotation files for VEP. Should have enough space for VEP cache files (~15GB).
- `CGAR_PORT`: the port number for CGAR web application.
- `CGAR_DB_PASSWORD`: mySQL password for CGAR application only. Should also be the same as `MYSQL_PASSWORD` under `cgar_db` section.
- `LOW_MEM`: keep the line if the installed memory (or memory available by Docker) < 40GB. Otherwise delete or comment out this line.
- `VEP_CACHE_VER`: the version of VEP annotation to be used in CGAR. Change this to use the latest VEP version.

- COSMIC_USER and COSMIC_PW: for cancer-associated genes and variants, CGAR uses data files from COSMIC (requires COSMIC account). CGAR will use this COSMIC account information to get data fiiles (downloaded only once during the first time of running).

Lastly, start and initalize CGAR application.

```
# start CGAR container for the first time
$ cd /path/to/docker-compose.yml
$ docker-compose up -d
# check if containers are up and running
$ docker container ls
...
# initialize VEP cache and other internal resources
# (these will take a lot of time...)
# 1. download and build VEP annotations
$ docker -it exec (cgar_web image name) /bin/bash /cgar/build_cache.sh
# 2. prepare ClinVar annotations
$ docker -it exec (cgar_web image name) /bin/bash /cgar/prep_ref_table_clinvar.py
# 3. prepare COSMIC annotations
$ docker -it exec (cgar_web image name) /bin/bash /cgar/prep_ref_table_census.py
$ docker -it exec (cgar_web image name) /bin/bash /cgar/prep_ref_table_cosmic.py
# 4. prepare denovoDB annotations
$ docker -it exec (cgar_web image name) /bin/bash /cgar/prep_ref_table_denovodb.py
```

Note: The first time of running CGAR images could take a long time due to the amount of time to build internal database.

Afterwards, if necessary, stop and start CGAR containers as follows.

```
$ cd /path/to/docker-compose.yml
$ docker-compose stop
$ docker-compose start
```

5.3 Access local CGAR

Assuming the port number 8081 was used for CGAR_PORT, open url http://localhost:8081.

5.4 Source Code

- Interested users can find the source code for CGAR is available from here.

CHAPTER 6

Contact

General inquiries may be addressed to Sek-Won Kong.

Specific questions regarding the use of CGAR and local installation, please send an email to gNOME staff.

Bibliography

[EIGMIX] X. Zheng and B. S. Weir. Eigenanalysis of SNP data with an identity by descent interpretation. doi:10.1016/j.tpb.2015.09.004.