Summarize Cohort Mutations

Variant information is stored on a per sample basis, but it can be informative to view variants in the context of recurrent variants identified within the project's sample cohort to identify both the frequency of variants and the samples that share a particular variant. The *Summarize cohort mutations* task can be invoked from any *Variants* or *Annotated variants* data node to generate a report of shared variants identified from detection against a reference sequence or among paired samples.

Summarize cohort mutations dialog

*T*he *Summarize cohort mutations* task, user needs to specify *Minimum coverage for genotype calls*. In general, it is likely that if a variant is not called in a sample at a particular locus then the sample has a homozygous reference genotype. Yet this may not always be the case as factors such as insufficient depth or low quality bases at that locus may lead to an inability of the variant caller to identify any genotype at that locus. As such, setting a minimum coverage will make the assumption that the sample contains a homozygous reference genotype if the depth requirement is met. This is done for the purpose of generating genotype calls for all samples (even reference homozygotes) at all variant loci within the project.

For paired variant caller report, if *Merge pairs* check button is unselected, pairs will be analyzed separately. If it is selected, all samples will be analyzed together.

Cohort mutation summary report

The **Cohort mutation summary report** provides a row in the table for all variant sites, either SNVs or INDELs, identified in the project (Figure 1). Hovering over a column header will provide a brief description of the column data. Columns presented in the table include the following information:

- *View* provides a link to Chromosome View by selecting the chromosome icon .
- *Chr* represents chromosome from the reference assembly
- *Position* represents the base position in the chromosome
- *Mutation type* is the category of variant (Substitution for SNVs and Insertion or Deletion for INDELs)
- *Reference allele* is the base(s) in the reference assembly sequence
- *Case genotypes* are the genotypes of the samples with a variant at the locus
- *HWE* is p-value of Exact test on Hardy-Weinberg Equilibrium[1], small p-value indicates deviation from HWE, which might be the consequence of e.g. genotyping error.
- *Variant frequency* represents the frequency of the variant site in the sample cohort
- *Sample count* is the fraction of samples in the cohort with the variant
- *Samples* are the names of the samples that contain the variant.

Download

Optional columns

	View	Chr	Position	Mutation Type	Reference allele	Case genotypes	HWE	Variant frequency	Sample count	Samples
1		1	13125	Substitution	C	CT	1	0.25	1/4	ERR1619385
2		1	13273	Substitution	G	GC	1	0.50	2/4	ERR1619381, ERR1619...
3		1	13417	Insertion	C	REF/ALT	1	0.50	2/4	ERR1619381, ERR1619...
4		1	14653	Substitution	C	CT	1	0.25	1/4	ERR1619385
5		1	14677	Substitution	G	GA	1	0.25	1/4	ERR1619385
6		1	16257	Substitution	G	CC,GC	0.428571	0.50	2/4	ERR1619381, ERR1619...
7		1	16298	Substitution	C	TT	0.142857	0.25	1/4	ERR1619078

Figure 1. Example of the Cohort mutations summary table (truncated) for variants detected against a reference sequence

The *Summarize cohort mutations* task is not available for variants detected by LoFreq as no genotypes are produced from the caller. If variant detection was performed on paired samples in Samtools (Figure 2), There are some extra columns:

- *GT Change* presents the possible change in zygosity between cases and controls at the variant locus
- *Control Genotypes* are the genotypes of the designated control samples in the pairs
- *Case Genotypes* are the genotypes of the cases in the pairs
- *Control/reference match* is yes or no in the column

Additional columns can be added to the *Cohort mutation summary report* table by selecting *Optional columns*. The optional columns are dependent upon the information present in the underlying vcf file and include variant and sample metrics from variant detection and information from the annotation. Hovering over a term in the list will provide a brief description of the data contained in that column. *Optional columns* can also be used to exclude default columns in the table.

Download

Optional columns

	View	Chr	Position	MutationType	Reference allele	GT Change	Case genotypes	HWE	Variant frequency	Sample count	Samples
1		chr1	49402	Substitution	A	HOM->HOM	TT	1	0.17	1/6	09-128-N
2		chr1	54751	Substitution	T	HOM->HOM	GG	0.333333	0.17	1/6	09-287-N
3		chr1	55500	Substitution	G	HOM->HOM	AA	1	0.17	1/6	09-287-T
4		chr1	55588	Substitution	T	HOM->HOM	CC	1	0.17	1/6	09-287-T
5		chr1	56888	Substitution	C	HOM->HOM	AA	1	0.17	1/6	09-287-T

Figure 2. Additional columns added to the Cohort mutation summary report for variants detected by Samtools paired analysis

Below each data column header in the *Cohort mutation summary report*, the *Search...* section allows for filtering of the table. The search can be useful for limiting the list of variants to those of interest when large numbers of variants are present in the table. For columns with numbers, exact values or ranges using either ">" or "<" can be utilized in the search. For columns with letters or words, and exact string of characters must be entered in order to obtain a match. In the case of table cells with multiple entries, there must be an exact match between the query and 1 entry to retain the table row.

If the *Summarize cohort mutations* task is performed upon an *Annotated variants* data node, additional information can be presented in the *Cohort mutation summary report* table. Click on Optional columns to select more fields to add to the table (Figure 3)

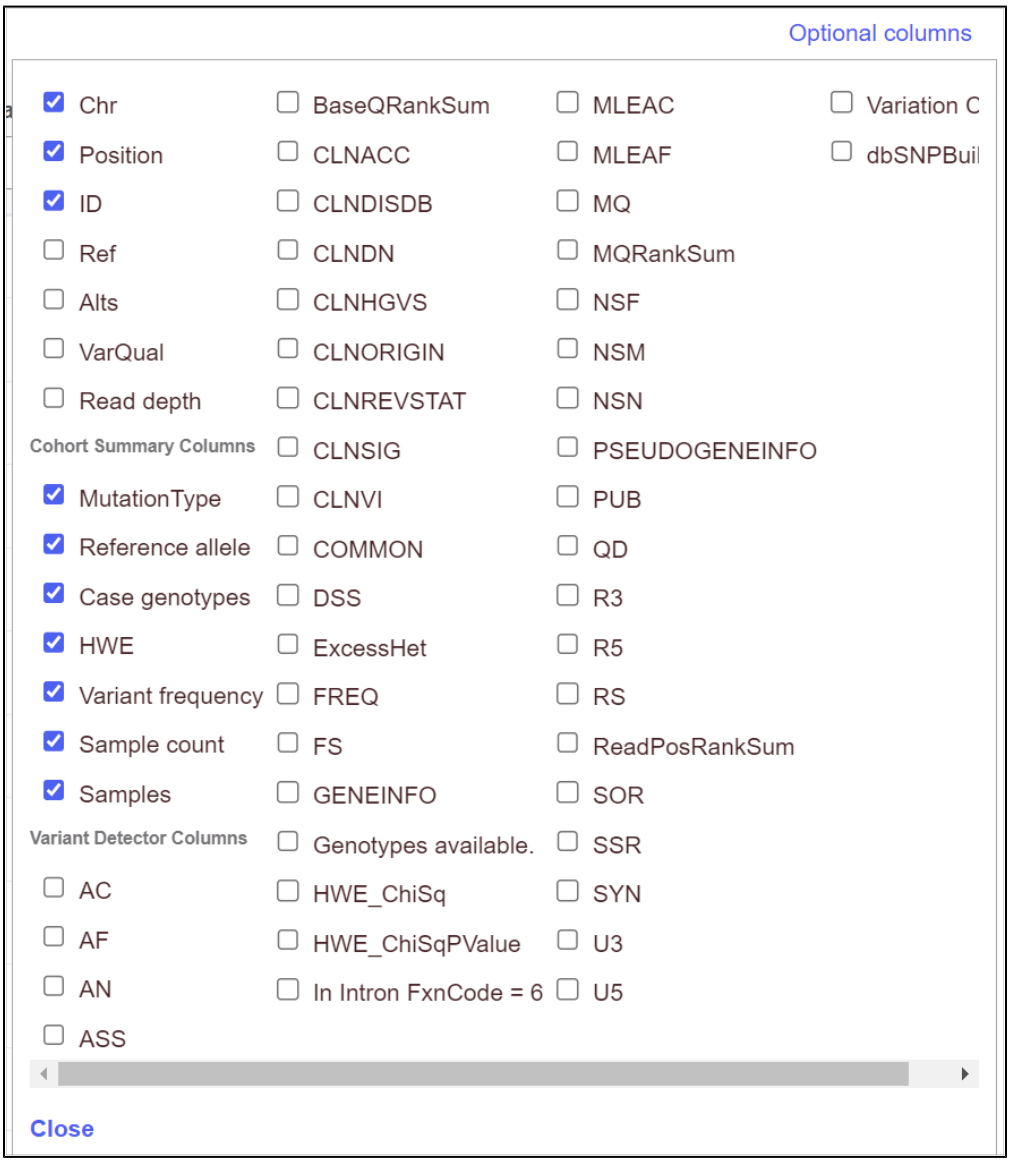

Figure 3. Optional columns for gene/feature annotation in the Cohort mutation summary report

At any point, information in the *Cohort mutation summary report* table can be saved in text or vcf format by selecting *Download* at the bottom right corner of the table. If the table is exported in text format, the visible table will be appended with additional columns for all samples in the project. These columns specify the genotype call for each variant locus in the project. In instances where no variant was detected within a sample, the value specified by *Minimum coverage for genotype calls* in the task dialog will be used to call either a homozygous reference genotype if above the specified threshold or no genotype if below the specified threshold.

References

1. Janis E, et al. A Note on Exact Tests of Hardy-Weinberg Equilibrium. PMC1199378

Additional Assistance

If you need additional assistance, please visit our support page to submit a help ticket or find phone numbers for regional support.

Your Rating:

Results: 39 rates