grexome@TIMC – user manual

Author: Nicolas Thierry-Mieg
Last updated: 24/06/2021

Introduction

This document defines the columns that may not be self-explanatory in the Cohorts, Transcripts, and Samples files produced by the grexome@TIMC pipeline.

Some abbreviations or definitions are used throughout:

- HR, HET, HV: Homozygous Reference, Heterozygous, Homozygous Variant.
- "cohort": used throughout but actually represents a phenotype abbreviation, e.g. MMAF or NOA. These are defined in the metadata files.
- "compatible": designates cohorts that are NOT used as negative controls for each other. These are defined in the pathologies metadata file.
- Multiallelic site: a site where several alternative alleles are observed in our entire dataset. These alleles are listed (comma-separated) in the ALT column, and genotypes in sample-lists match this, e.g. a sample genotyped 0/2 is heterozygous for the REF and the second ALT allele.

Cohorts files – identifying candidate causal variants

One file for each phenotype = $cohort. Each row represents one variant+transcript pair. Therefore there can occasionally be several lines with the same POSITION-REF-ALT:

- if a variant affects several different transcripts – this is rare currently because we only keep PICKed transcripts (usually a single transcript per gene) and we filter out MODIFIER-impact variants;
- if a position is multiallelic (several alleles are listed in the ALT column) – there will be one line for each ALT allele that passes all filters and that is observed in $cohort. This is rare and often a sign that the position is error-prone, but it could be relevant.

Many columns come from VEP and should be mostly self-explanatory, but you can check the documentation or ask me if needed. The other columns are defined as follows.

KNOWN_CANDIDATE_GENE: lists the pathologie(s) that this gene is suspected to be involved in, along with the associated confidence score(s), as indicated in the candidateGenes metada files. The confidence scores typically range from 1 (low confidence candidate) to 5 (causal variants have been identified/published in this gene for this pathology). If a gene is a candidate for several pathologies, they all appear as a colon-separated string.

COUNT_HR: total number of HR samples from any cohort.

COUNT_$cohort_HV, COUNT_$cohort_HET: number of HV or HET samples from this $cohort, excluding those that have a $causalVariant in another gene.

COUNT_$cohort_OTHERCAUSE_HV, COUNT_$cohort_OTHERCAUSE_HET: number of HV or HET samples from $cohort that have a $causalVariant in another gene.

COUNT_COMPAT_HV, COUNT_COMPAT_HET: number of HV or HET samples from a cohort defined as "compatible" with $cohort.

COUNT_NEGCTRL_HV, COUNT_NEGCTRL_HET: number of HV or HET samples from all other cohorts.

COUNT_OTHERGENO: only relevant for multiallelic sites, this is the total number of samples from any cohort whose genotype involves a different ALT allele than the one that this row deals with.

IMPACT: VEP impacts are complemented with MODHIGH, currently representing:

- missense variants predicted as deleterious by at least 3 methods among SIFT, PolyPhen, CADD, mutationTaster, REVEL;
- and splice_region_variants predicted deleterious by both ada_score and rf_score (from dbscSNV).

$cohort_HV, $cohort_HET, $cohort_OTHERCAUSE_HV, $cohort_OTHERCAUSE_HET, COMPAT_HV, COMPAT_HET: genotype ~ list of samples counted in the corresponding COUNT* columns (see above). The genotype is always 1/1 for HV and 0/1 for HET, except at multiallelic sites. Each sample appears as "sampleID(patientID)[DP:AF]", where DP is the sequencing DePth and AF is the Allele Frequency (fraction of reads that support the ALT allele).

NEGCTRL_HV, NEGCTRL_HET, OTHERGENO: lists of NEGCTRL and OTHERGENO samples counted in the corresponding COUNT* columns, these appear at the end of the files (last columns). Format is the same as $cohort_HV, see above.

GTEX_testis_RATIO, GTEX_ovary_RATIO: relative expression in testis or ovary compared to the average expression in all available tissues.

GTEX_*: GTEx v7 expression data – "450 donors and over 9600 RNA-seq samples across 51 tissue sites and 2 cell lines", values are the median TPM in each tissue. Tissues deemed most interesting to us immediately follow the GTEX_*_RATIO columns, all other GTEX columns appear later.

Transcripts files – identifying candidate causal genes

One file per cohort. Where the COHORTS files are variant-centric, these files are transcript-centric: each row represents one transcript (“Feature” column). Genes are also *usually* represented in a single row since we currently keep only PICKed transcripts.

Many columns have the same title and content as in the COHORTS files (see above). Additional columns in SAMPLES files are:

COUNTSAMPLES_HV_HIGH+ : number of samples (in this cohort) with at least one HV HIGH variant.

COUNTSAMPLES_HV_MODHIGH+ : number of samples with at least one HV MODHIGH or HIGH variant.

COUNTSAMPLES_HV_MODER+ : number of samples with at least one HV MODERATE or MODHIGH or HIGH variant.

COUNTSAMPLES_BIALLELIC_HIGH+ : number of samples with at least TWO HET (or one HV) HIGH variants.

COUNTSAMPLES_BIALLELIC_MODHIGH+ : number of samples with at least TWO HET (or one HV) variants whose impact is MODHIGH or HIGH.

COUNTSAMPLES_BIALLELIC_MODER+ : number of samples with at least TWO HET (or one HV) variants whose impact is MODERATE or MODHIGH or HIGH.

COUNTSAMPLES_OTHERCAUSE: same as the 6 COUNTSAMPLES* above, but counting the samples from this cohort that have a "known causal variant" in another gene, all concatenated into a single colon-separated string (with :: between HVs and HETs to make reading it easier).

COUNTSAMPLES_COMPAT: same as COUNTSAMPLES_OTHERCAUSE but counting the samples belonging to “compatible” cohorts.

COUNTSAMPLES_NEGCTRL_HV_HIGH+ and 5 more COUNTSAMPLES_NEGCTRL_* columns: same as COUNTSAMPLES_HV_HIGH+ and friends defined above, but counting the samples belonging to NEGCTRL cohorts (i.e. any cohort that is not “compatible” with this cohort).

HV_HIGH, HV_MODHIGH, HV_MODER, BIALLELIC_HIGH, BIALLELIC_MODHIGH, BIALLELIC_MODER: non-redundant list of samples counted in the corresponding COUNTSAMPLES columns.

OTHERCAUSE_BIALLELIC_MODHIGH+ : all samples that belong to this cohort but have a "known causal variant" in another gene, and that bear in this transcript at least one

HV or two HET variants whose impact is MODHIGH or HIGH. This is exactly the list of samples counted in COUNTSAMPLES_BIALLELIC_MODHIGH+.

COMPAT_BIALLELIC_MODHIGH+ : same as OTHERCAUSE_BIALLELIC_MODHIGH+ but listing samples belonging to COMPATible cohorts.

NEGCTRL_BIALLELIC_MODHIGH+ : same as OTHERCAUSE_BIALLELIC_MODHIGH+ but listing samples belonging to NEGCTRL cohorts.

IMPORTANT NOTE:

- The COUNTSAMPLES columns are **cumulative**, for example HV HIGH samples are counted in all six COUNTSAMPLES values.
- The sample-list columns for $cohort are not: they are **non-redundant**, as specified. For example HV_HIGH samples are **not** listed in the HV_MODHIGH column (despite being counted in COUNTSAMPLES_HV_MODHIGH+). This allows to sort or filter sanely on e.g. COUNTSAMPLES_BIALLELIC_MODHIGH+ while keeping file sizes and legibility reasonable.
- The sample-list columns for OTHERCAUSE, COMPAT and NEGCTRL only list the BIALLELIC_MODHIGH+ samples, but they list them all. Therefore these sample-lists are cumulative, but they are not redundant since there is only one list per category. Samples from these categories that are hit only by MODERATE variants are not listed.

Samples files – diagnosing a specific sample

One file per sample.

Many columns have the same title and content as in the COHORTS files (see above). Additional columns in SAMPLES files are:

- GENOTYPE: HET or HV, at multiallelic sites you can check ALLELE_NUM to know which ALT allele was seen in this sample.
- DP:AF: sequencing DePth and Allele Frequency (fraction of reads that support the ALT allele), COHORTS files had this information in the $cohort_HV etc columns.
- BIALLELIC: says whether both alleles of this gene may be hit, and how severely. Designed to find compound heterozygotes (in addition to HVs). Value is one of:
  - HIGH → at least 1 HV or 2 HET variants whose impact is HIGH affect this gene in this sample;
  - MODHIGH → at least 1 HV or 2 HET variants whose impact is MODHIGH or HIGH affect this gene in this sample;
  - MODERATE → at least 1 HV or 2 HET variants whose impact is MODERATE or MODHIGH or HIGH affect this gene in this sample;
  - LOW → at least 1 HV or 2 HET variants whose impact is at least LOW affect this gene in this sample;
  - NO → this variant is HET and no other variant hits this transcript in this sample.