Description Research Annotations

Department of Human Genetics
Radboud University Medical Center,
Nijmegen, The Netherlands

June 2018

For questions please contact:
Maartje.vandeVorst@radboudumc.nl

Chromosome
Chromosome where the identified variant is located on. Chromosomes can be chr1 – chr22, chrX, chrY and chrM (mitochondrial chromosome).

Start position
Start position of the identified variant (genomic position on the respective chromosome). Positions are 1 based and we count the bases.

End position
End position of the identified variant (genomic position on the respective chromosome). In case of a substitution or insertion the start and end position are identical, in case of a deletion the end position is the start position + the length of the deletion.

Reference
Reference nucleotide (wt) at the given genomic position (based on HG19/NCBI build 37).

Variant
Variant nucleotide detected at the respective position. If there are two different changes compared to the reference sequence the variant is denoted twice in the file (e.g. chr 5:200 A>C and chr5:200 A>G).

Quality by Depth
Variant call confidence normalized by depth of sample reads supporting a variant.

Variant type
Kind of detected aberration (substitution, deletion, insertion, complex).

SNP id
rs-number as given in dbSNP (dbSNP version is given in the meta info of the variant file).

SNP state
Same SNP: the exact identified variant is reported in dbSNP.
Overlapping location: there is a variant reported in dbSNP at the same position, but the nucleotide change is different. (E.g. the nucleotide change you see in your data is G>A, but the SNP reported in dbSNP is G>T, at the same genomic position).
Blancs: the identified variant is not reported in dbSNP.

SNP reference
Reference (wt) allele of the reported SNP

SNP variant
Variant allele of the reported SNP

SNP Frequency
SNP frequency as given in dbSNP. If no frequency information is available you will see an artificial value of -1.

Causative – Projects
When the identified variant has already been reported to be causative (e.g. by in our in-house database, Alg018 or HGMD) there will be an entry listed in this column. This entry tells us the project for which the variant has been identified to be causative.

NonCausative - Allele Frequency
When the identified variant has so far NOT been reported to be causative (e.g. by in our in-house database, Alg018 or HGMD) there will be a allele frequency value listed in this column. The frequency tells in how many alleles the variant was identified (percentage).

NonCausative – Projects
When the identified variant has so far NOT been reported to be causative (e.g. by in our in-house database, Alg018 or HGMD) there will be an entry listed in this column. This entry tells us all projects in which the variant has been identified.

Gene name
Name of the gene in which the variant was identified.

NC Gene name
Name of the non-coding gene in which the variant was identified.

Gene id
Transcript id (ENST-number) belonging to the respective gene.

NC Gene ids
Transcript id (ENST-number) belonging to the respective non-coding genes.

Strand
Demonstrates whether the gene is located on the plus or on the minus strand.

Total exons
Total number of exons belonging to the gene in which the variant was identified.

Gene component
The gene component tells whether a variant is located within a gene or outside the gene. Possible annotations:
EXON_REGION
CODING_SPLICE_SITE_REGION: synonymous variants located within 1-3 bases of the exon
SA_SITE_CANONICAL: Splice acceptor, positions -1 and -2 (AG)
SD_SITE_CANONICAL: Splice donor, positions +1 and +2 (GT)
NONCODING_SPLICE_SITE_REGION: Splice donor/acceptor, positions within 3-8 bases
INTRON_REGION
5'UTR
3'UTR
Blancs: If this column is empty (no entry for a given variant), then the variant is located outside of a gene.
In case the transcript is non-coding the annotation will have NONCODING_TRANSCRIPT_ in front, for example: NONCODING_TRANSCRIPT_EXON_REGION

Component nr
This number tells us in which component a variant is found (e.g. in exon 2).

Local position string
Text that describes the location of the variant with respect to the nearest start / end of the genomic component (exon/intron) it is located in.
Examples:
Ex2-5:

the variant is located in exon 2, at the position -5 counted from the 3'end of the exon
Ex2 +14:
the variant is located in exon 2, at the position +14, counted from the 5'end of the exon
IVS8 -5:
the variant is located in intron 8 , at position -5, counted from the 3'end of the intron
IVS12 +7:
the variant is located in intron 12, at position +7, counted from the 5'end of the intron
Reason to look at this local position: By this you get an idea of a variant is located rather in the middle of an exon or intron, or somewhere near the exon/intron boundaries. Positions like Ex 5-1 and -2 for example still belong to the splice donor, but are annotated as exonic, and not as splice site. With the help of the local position information you might therefore get an idea of an exonic variant might simultaneously have an effect on a splice site. This might also be useful to check for synonymous changes, that might still have an effect on the splice site.
IVS = intervening sequence = another word for intron.

Total coding size
Length of all coding exons in bps.

Protein Effect
Indicates the effect of the variant on protein level. For the possible annotations see the table of Ensemble (https://www.ensembl.org/info/genome/variation/predicted_data.html#consequences).

All transcripts
Transcript id (ENST-number) of all overlapping transcripts including the protein effect.

Reference Amino Acid
Reference amino acid (wt) at the respective position.

Mutation Amino Acid
Variant amino acid based on the mutation sequence at the respective position. * indicates a stop codon, whereas X indicates a frameshift.

Synonymous
TRUE: the identified variant does NOT lead to an aminoacid change (=synonymous)
FALSE: the identified variant does lead to an aminoacid change (=non-synonymous)

AminoAcid position
Position of the aminoacid (e.g. 487).

Codons
Codon change (e.g. Aca/Gca).

Codon position
Position of the identified variant within the codon. Possible positions:
-1: artificial value, given for variants in introns, UTR, splice sites, outside genes
1: first position in the codon
2: second position in the codon
0: third position in the codon (usually leading to synonymous AA)

Hgvsg
HGVS genomic nomenclature (e.g. 1:g.69511A>G)

Hgvsc
HGVS coding nomenclature (e.g. ENST00000335137.4_2.1:c.421A>G)

Hgvsp
HGVS protein nomenclature (e.g. ENSP00000334393.3:p.Thr141Ala)

mRNA changes
Identified change annotated on mRNA level (e.g 1460G>C).

Protein Id
Uniprot id

Protein variant
Identified aminoacid change (e.g. P>L).

Protein domain
Tells whether the identified variant is located within a known protein domain (from uniprot).

Domain description
Description of the protein domain containing the identified variant.

phyloP
PhyloP score for evolutionary conservation on nucleotide level (Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010; 20(1):110-121.). Score ranges from -7 till +7. Minus values are not conserved. The higher the score, the better the conservation. Score of above 2.5 are considered to be most of interesting for pathogenic mutations (Vissers et al, Nat Genet 2010).

Gerp coding score
A genomic evolutionary rate profiling (GERP) score. How higher the score, how more conserved. (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html)

CADD_RawScore
Combined Annotation Dependent Depletion (CADD) raw score. Best not to use this score for variant prioritization, but use the CADD_PHRED score.

CADD_PHRED
Kircher M, et al, Nat Genet 2014 suggests to put a cutoff somewhere between 10 and 20 (http://cadd.gs.washington.edu/info).

Grantham Score
Grantham scores categorize codon replacements into classes of increasing chemical dissimilarity (Granthan R., Amino acid difference formula to help explain protein evolution. Science 1974 185:862-864). Note that this is an unreliable score and is best not used for variant prioritization.

HOMOLOGS
Number of homologs

GO term
Number for a certain Gene Ontology Biological Process term belonging to the gene in which a variant is found.

GO description

Gene Ontology description for the respective gene in which a variant is found.

GeneCard link

Link to the GeneCards webpage where you can find information about the gene.

OMIM

Tells whether the gene in which a variant was found has an entry in OMIM by giving the link to OMIM, if available. The link can be activated by double click on it.

OMIM_DISEASE

When there is a disease entry in OMIM for the gene in which a variant was found, this columns mentions the corresponding disease.

KEGG Name

KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/).
KEGG names = names of pathways in which the respective gene (in which a variant was found) is involved. (e.g. axon guidance, base excision repair, cell cycle, etc.).

KEGG Class

Class to which the corresponding KEGG pathway is belonging to. (e.g. cellular processes, cell growth and death, metabolism, etc.).

KEGG Link

Link to the KEGG pathway in which the gene in which a variant was found is annotated.

Entrez id

Entrez gene ID

Mouse phenotypes

Phenotypes found in mice

Mouse phenotypes - low level

Phenotypes found in mice in low levels

Protein Accession

Protein id according to UniProtKB

DISEASE

Information on whether the gene in which a variant was found is part of one (or more) of our gene packages. If so, it means that this is a gene known to be involved in the respective disease. In addition, the known inheritance pattern is given (e.g. blindness (AR,AD)).
AR= autosomal recessive
AD= auosomal dominant
XL= X-linked

SPIDEX

Dataset which provides machine-learning prediction on how genetic variants affect RNA splicing (Xiong et al, Science 2015)

EXAC
Data from 60,706 unrelated individuals.

EXAC AF
Allele frequency from 60,706 unrelated individuals.

EXAC AC
Allele count from 60,706 unrelated individuals.

EXAC AN
Allele number from 60,706 unrelated individuals.

EXAC AC HET
Count of heterozygous variants

EXAC AC HOM
Count of homozygous variants

EXAC_nonTCGA
Same as *EXAC*, but without samples from The Cancer Genome Atlas.

EXAC AF_nonTCGA
Same as *EXAC AF*, but without samples from The Cancer Genome Atlas.

EXAC AC_nonTCGA
Same as *EXAC AC*, but without samples from The Cancer Genome Atlas.

EXAC AN_nonTCGA
Same as *EXAC AN*, but without samples from The Cancer Genome Atlas.

EXAC AC HET_nonTCGA
Same as *EXAC AC HET*, but without samples from The Cancer Genome Atlas.

EXAC AC HOM_nonTCGA
Same as *EXAC AC HOM*, but without samples from The Cancer Genome Atlas.

Gene expression
Tissues where the respective gene is expressed (e.g. lung, bone, kidney).

Expressed tissue number
Number of tissues where the respective gene is expressed.

Protein interactions
Interactions with other proteins. Proteins marked as (!) are within the OMIM database.

Protein interaction number
Number of protein-protein interactions.

Disease interaction number
Number of protein-protein interactions where those proteins are within the OMIM database.

Percentage disease interaction
<Disease interaction number> / *<Protein interaction number>* * 100

MAF_ESP
Minor allele frequency of ESP data (http://evs.gs.washington.edu/EVS/).

AF_GoNL
Allele frequency of GoNL data (http://www.nlgenome.nl/?page_id=9).

AF_Wellderly
Allele frequency of wellderly data (data retrieved from Complete Genomics).

Kaviar_AF
Allele frequency from the Kaviar database (http://db.systemsbiology.net/kaviar/).

CG_60_genomes
Allele frequency of 60 Complete Genomics genomes (data retrieved from Complete Genomics).

AF_WGS_1000genomes
Allele frequency retrieved from 1000 Genomes Project phase 3. This dataset contains whole exome and genome sequencing data from 2504 samples and was downloaded from the European Variation Archive.

MAF_WGS_BGI-Shenzhen
Minor allele frequency from BGI-Shenzhen major study of depression in Chinese woman. This dataset contains whole genome sequencing data from 10640 samples and was downloaded from the European Variation Archive.

AF_WGS_decode
Allele frequency of 2636 whole genome samples from deCODE. This dataset was downloaded from the European Variation Archive.

AF_WES_GEUVADIS
Allele frequency from 937 whole exome samples retrieved from GEUVADIS and downloaded from the European Variation Archive.

AF_MAX_WGS_UK10K
Allele frequency from UK10K Avon Longitudinal study. This dataset contains 1928 whole genome samples downloaded from European Variation Archive.

Rare_Ns_Kb
Rate non-synonymous per 1000 bases

Rare_Ns_Ss
Rate non-synonymous versus synonymous

decodeAvg, decodeFemale and decodeMale
Recombination annotation of a mutation

TopologicalDomains
TAD data from mouse and human tissue, downloaded from http://chromosome.sdsc.edu/mouse/hi-c/download.html

VistaEnhancerBrowser
Enhancer data, including the tissue in which the enhancer is expressed, downloaded from http://enhancer.lbl.gov/

CTCFBindingSites
CTCFBinding sites, including the specific cell line, downloaded from http://insulatordb.uthsc.edu/

ENCODE_Region
Regulation regions are annotated (e.g. promoters, enhancer, insulator). Used for the prioritization of non-coding variants.

ENCODE _TFBS
Transcription factor binding sites retrieved from ENCODE. Used for the prioritization of non-coding variants.

SuperEnhancer
The cell line of the super enhancer in which the variant is located is annotated. Data is retrieved from Hnisz et al., Cell, Oct 8, 2013.

SuperEnhancerScore
Score of the super enhancer from *SuperEnhancer*. See Hnisz et al., Cell, Oct 8, 2013.

SimpleRepeats
"Simple repeat" is annotated to the variant when this variant is located in a simple repeat.

Potential Introduced Splice Sites
"Introduced startcodon" is annotated when the variant is located within the 5'UTR and the nucleotide order of a startcodon was introduced.
"Potential introduced ss" is annotated when the variant is located within the intron region and a splice site was introduced based on the sequence.

Nearest gene
Distance in bps of the two nearest genes (e.g MIR4772 (Distance: 10787bp) + SLC9A4 (Distance: 30149bp). If the variant is located within the gene, the gene name is annotated.

FitConsScores
These scores serve as a evolution-based measure and in the paper they show considerably improved prediction power for cis regulatory elements. Data from 4 cellines (HUVEC, I6 multicell, GM12878 and H1-hESC) is downloaded and annotated. The ranges are from approximately 0-0.8. For coding is this > 0.4 and for non-coding 0.05-0.35 (http://compgen.cshl.edu/fitCons/0downloads/tracks/).

mirdmg Name
miRNA names are annotated retrieved from http://mirna.bioinf.be/annotate.php.

mirdmg Impact
Variants annotated with mature, arm, loop, DEL or flank are located in the primary miRNA.

gnomAD-E
Variants having an exact match with a variant from the gnomAD exome database will have an annotation "M" and variants having an overlap with a variant will have an annotation "O". Data is retrieved from http://gnomad.broadinstitute.org/

gnomAD-E AF
Allele frequency of the matching variant from the gnomeAD exome database.

gnomAD-G
Variants having an exact match with a variant from the gnomAD genome database will have an annotation “M” and variants having an overlap with a variant will have an annotation “O”. Data is retrieved from http://gnomad.broadinstitute.org/

gnomAD-G AF
Allele frequency of the matching variant from the gnomeAD genome database.

Located in TAD
Annotates TRUE or FALSE. TRUE means that the variant is located within a Topological Associating Domain (TAD). TAD data was obtained from the databases GEO and ENCODE.

Genes in TAD
Annotates overlapping genes

Disease genes in TAD
Annotates overlapping disease genes from our in-house panels

In TAD boundary
Annotated TRUE or FALSE. TRUE means that the variant was located on the boundary of a TAD.

Enhancer Position
Annotates the coordinates of the overlapping enhancers. Data was obtained from the databases FOCS, GEO and Slidebase.

Promoter Position
Annotates the coordinates of the overlapping promoters. Data was obtained from the databases FOCS and Slidebase.

Genome Pathogenicity Score
Annotates the in-house created score. The score [-22 – 15] is calculated by combining multiple annotations such as CADD, PhyloP, GnomAD-G AF and the TAD annotations above.

Enzyme Tests
Annotates diagnostic enzyme tests

%var by AD
Number of reads that show the variant. This number is always smaller than or equal to the total number of reads.

Annotation retrieved from GATK (e.g. AF, AC, AN, GT, DP, AD, etc.)
DP = number of reads (depth)
AF = allele frequency
AC = allele count
AN = allele number
GT = genotype
AD = allelic depth
See https://www.broadinstitute.org/gatk/guide/tagged?tag=annotation for detailed description and other annotations.

Index calls
Only relevant for trio (de novo) analyses. In that case it represents the individual base calls of the reads at this position.

Paternal calls
Only relevant for trio (de novo) analyses. In that case it represents the individual base calls of the reads at this position.

Maternal calls
Only relevant for trio (de novo) analyses. In that case it represents the individual base calls of the reads at this position.

De novo assessment
Only relevant for trio (de novo) analyses.
MV = variant was called in the mother
PV = variant was called in the father
Maternal = variant was not called in the mother, but inspection of the alignment showed inheritance from the mother
Paternal = variant was not called in the mother, but inspection of the alignment showed inheritance from the father
Shared = variant was not called in the mother or father, but inspection of the alignment showed inheritance from both
Possible de novo = variant was not called in the mother or father and no additional evidence for inheritance was found in the parental alignment files