Genome Trax™ User Manual

Revision 2014.3

Table of Contents

General format for track descriptions

Track Properties

Each track is described following the same template, with information on the following:

Version: The release number or release date of the dataset from which the track was created. Genome Trax™ provides an up-to-date collection of both proprietary and public datasets.

Track Description: A detailed, sometimes technical description of the contents of the track, with background information about the data source and processing.

Benefit: A short description of the main benefit of the track.

Track Name: A short base name for the track, used in naming annotation files. 15 characters maximum.

Annotation Fields: A table defining the annotation fields for the track's records, and describing the nature of their content. Each track has annotation records, typically thousands or millions. Each record contains annotation stored in track-specific annotation fields. The description can contain the following elements in addition to a general explanation:

Cardinality: 1 means that each record will contain exactly one value in this field, and the field always will contain a value. 0..1 means that the value is optional, but if present there can be only one value. 0..* means that the field can be empty or contain multiple values (a list of values), and 1..* means that there always must be at least one value, but the field also can contain a list of values.

MySQL Type: The database type of the column in the MySQL relational version of Genome Trax.

Accepted Values: Certain fields only allow values from a controlled vocabulary as values. The permissible possible values are listed here.

Examples: This element presents one or more example values for the field, sometimes with further explanations about their format.

Default Annotations Fields:

The Annotation Fields that follow are present in all tracks, to ease integration of results from multiple tracks.

NAME (LEGACY NAME)	DESCRIPTION
ID	This Field is only provided in the GFF3 flat-files, in compliance with the GFF3 specification. Unique integers are assigned to this field. Cardinality: 1
accession	Identifiers in the data source (external) Please note that the exact format of this field is **under development** and likely to change in future releases. MySQL Type: TEXT Cardinality: 1
brief (feature)	A snapshot of the most relevant information associated with the record. The value usually is a mashup of information extracted from some of the remaining annotation fields. The exact formula used to generate the value is track specific and documented together with the remaining annotation fields of each track. For being a 'synthetic annotation field', it might not be availabe in some serialized forms (eg. relational). Please note that the exact format of this field is **under development** and likely to change in future releases. MySQL Type: TEXT Cardinality: 0..1
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. MySQL Type: TEXT Cardinality: 0..*
	External identifier. Entrez gene ID for the gene

entrez_gene_id (entrez)	MySQL Type: BIGINT Cardinality: 0..*
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. MySQL Type: TEXT Cardinality: 0..*
hyperlink	Link to a report or web-page with more detailed information. MySQL Type: TEXT Cardinality: 0..1
pmid	Pubmed ID of the reference from which the information was taken. MySQL Type: BIGINT Cardinality: 0..*
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. MySQL Type: TEXT Cardinality: 0..*

Track Descriptions

Mutations and Variants

"Mutations and Variants" correspond to variation that has been associated with human disease. It includes human disease mutation data from HGMD Professional®,PGMD™, ClinVar, GWAS Central polymorhphisms, COSMIC somatic variants, Exome Variant Server (EVS), dbNSFP and dbSNP or Ensembl entries overlapping with other BIOBASE features.

HGMD® inherited disease mutations

Version: HGMD® professional 2014.3

Track Description: This track contains germ-line mutations which have been manually curated from the scientific literature. The track contains all disease-associated mutations from the Human Gene Mutation Database, HGMD® Professional, for which chromosomal coordinates are available, including polymorphisms with functional implications.

For a full description of the curation policies and mutation types, see the online documentation for HGMD®. This documentation also contains a detailed description of the polymorphism inclusion criteria in HGMD®.

Benefit: These are exclusive, manually curated, germ-line specific mutations from the scientific literature, not available elsewhere in such a comprehensive manner. HGMD®, has been the gold standard for numerous analyses of whole genome sequencing efforts, starting from James Watson and Craig Venter, all the way through to clinical and diagnostic applications.

Track Name: hgmd

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	HGMD® mutation accession. Cardinality: 1 MySQL Type: TEXT Examples: `CM960042`
alt	Alternative base Cardinality: 1 MySQL Type: TEXT
aminoacid_change	Aminoacid change represented in amino acid code Cardinality: 0..1 MySQL Type: TEXT
brief (feature)	The gene, disease and mutational change. Cardinality: 0..1 MySQL Type: TEXT Examples: `BRCA1;Breast_Cancer;134C>T`

citation_type	The nature of the information source for the mutation.
	Cardinality: 1..*
	MySQL Type: VARCHAR (11)
	Accepted Values:
	`Primary`
	All reported data is obtained from this paper.
	`LSDB`
	Some mutations are imported from LSDB, hence do not have a pmid.
	`LSDB Report`
	Some mutations are imported from LSDB, hence do not have a pmid.
	`APR`
	Additional phenotype report.
	`FCR`
	Functional characterization report.
	`MCR`
	Molecular characterization report.
	`SAR`
	Simple additional report.
	`FAPR`
	Functional characterization additional phenotype report.
	`ACR`
	Additional literature report.

codon_change	The codon changes effected because of the mutation. Cardinality: 0..1 MySQL Type: TEXT
	The affected codon in the gene sequence. This is also the position of the amino acid residue in the protein.

codon_number	Cardinality: 0..1 MySQL Type: BIGINT

comments	Any additional observations noted by the curators.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`aka G20566T`
	`aka IVS-II-654/c.316-197`

confidence	The strength of the evidence for the mutation/disease relationship.
	Cardinality: 1
	MySQL Type: VARCHAR (4)
	Accepted Values:
	`High`
	If curators had strong evidence.
	`Low`
	If curators had some reservation about the strength of the evidence.

disease	The associated disease or phenotype.
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`17-alpha-hydroxylase/17,20-lyase deficiency`
	`Mucopolysaccharidosis II`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT

genomic_sequence	30 residues upstream and downstream flanking region of the mutation described.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`tcatgttcatacctcttatcttcctcccac(a/g)gCTCCTGGGCAACGTGCTGGTCTGTGTGCT`
	`GATGCCAAACCAGGTCAATTCCCTTGGCAG(g/t)tactttatactgatggtgtgtcaaaactgg`
	`agaataacagtgataatttctgggttaagg(c/t)aatagcaatatctctgcatataaatatttc`

hgmd_acc (hgmdAcc)	HGMD ID Cardinality: 1 MySQL Type: TEXT Examples:

	`CS850009`
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hgvs	The complete hgvs description for the mutation. Cardinality: 0..1 MySQL Type: TEXT Examples: `NM_000518.4: c.316-2A>G`
hyperlink	A static version of HGMD® professional mutation report (Full reports can be reached from this page. To get access to full HGMD® professional functionality and content, a separate license is required). Cardinality: 0..1 MySQL Type: TEXT
lsdb_source	The locus specific database (LSDB) in which the variant was reported. A small percentage of the mutations in HGMD does not originate from litereature reports, but from well documented reports in LSDBs. In those cases, the lsdb_source field will contain the URL of the LSDB, and pmid_citation_type will be LSDB Report. Cardinality: 0..1 MySQL Type: TEXT Examples: `http://www.genet.sickkids.on.ca/cftr/`

mutation_type (mutationType)	A one-letter code determining which class (and table in HGMD) the mutation belongs to. If there are several mutations of different type they are given as a comma-delimted list, such as *D,I,M* for a gene with at least one deletion, insertion and point mutation.
	Cardinality: 1..*
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	Deletion.
	`E`
	Amplet.
	`G`
	Gross deletion - refers to lesions covering more than 20 nucleotides.
	`I`
	Insertion.
	`M`
	Mutation (mis-sense or non-sense single nucleotide).
	`N`
	Gross Insertion/Deletion.
	`P`
	Complex Rearrangement.
	`R`
	Promoter mutation.

	`S`
	Splice site mutation.
	
	`X`
	Indel.

nucleotide_change (nucleotideChange)	A description of the nucleotide change in HGVS nomenclature. Cardinality: 0..1 MySQL Type: TEXT Examples: `316-2A>G`
omim	Accession number of the corresponding OMIM entry. Cardinality: 0..1 MySQL Type: BIGINT Examples: `300746`
pmid	Pubmed ID Cardinality: 0..* MySQL Type: BIGINT Primary Field: citation_type
pmid_notes	Any additional observations noted by the curators for a specific paper. This list of notes has a one to one correspondence with the list of pmids given in the pmid key. Cardinality: 0..* MySQL Type: TEXT Primary Field: citation_type
ref	Reference base Cardinality: 1 MySQL Type: TEXT
rsid	rsid of the SNP entry in dbSNP, corresponding to the mutation, where available. Cardinality: 0..1 MySQL Type: TEXT Examples: `rs33914668`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT
	The severity category of the variant. Cardinality: 1..* MySQL Type: VARCHAR (3) Accepted Values: `DP`

variant_type (variantType)	Disease-associated polymorphism - A polymorphism reported to be in significant association with a disease/phenotype (p<0.05) that is assumed to be functional (e.g. as a consequence of location, evolutionary conservation, replication studies etc), although there may as yet be no direct evidence (e.g. from an expression study) of a functional effect. DFP Disease-associated polymorphism with additional supporting functional evidence - A polymorphism reported to be in significant association with disease (p<0.05) that has evidence of being of direct functional importance (e.g. as a consequence of altered expression, mRNA studies etc). FP In vitro/laboratory or in vivo functional polymorphism - A polymorphism reported to affect the structure, function or expression of the gene (or gene product), but with no disease association reported as yet. FTV Frameshift or truncating variant - A polymorphic or rare variant reported in the literature (e.g. detected in the process of whole genome/exome screening) that is predicted to truncate or otherwise alter the gene product (i.e. a nonsense or frameshift variant) but with no disease association reported as yet. Please note that any variant affecting the obligate donor/acceptor splice site of a gene will not be included in this category unless there is evidence for an effect on the splicing phenotype. Variants occurring in pseudogenes will also be excluded unless evidence for a functional effect is present for both the pseudogene itself and the variant in question. CNV Copy number variations are DNA segments >1 kb in length that present with variable numbers of copies in a given population. These variants are being reported in the literature with an ever increasing frequency. CNVs are potentially functionally significant and should therefore in principle be treated by HGMD in a similar manner to any other polymorphism. DM Disease causing mutation - Pathological mutation reported to be disease causing in the corresponding report. DM? Disease causing mutation (report questionable) - mutation reported to be disease causing in the corresponding report, but where the author has indicated that there may be some degree of doubt, the curator had doubts about the validity of the claim given the data presented or subsequent evidence has come to light in the literature, calling the deleterious nature of the variant into question. R Removed - mutations that were removed from the database, for example because the report was erroneous or has been retracted. To allow users to track these changes, the records were not actually removed, but flagged as R, retaining all their other characteristics. These variants should not be used for annotation purposes.

HGMD® imputed inherited disease mutations

Version: HGMD® professional 2014.3

Track Description: This is derived from HGMD® the track. From original HGMD® mutations, all alternative possible codon changes that result in the same amino acid change reported in the actual HGMD® record are calculated. To each of these imputed mutations, the information from the corresponding mutation from HGMD® is transferred.

Benefit: In cases where the phenotype is caused by an amino acid change, you will be able to find novel mutations that have not been reported in the literature as disease causing, but lead to the same change. The assumption here is that codon changes influence phenotype through the resulting protein change. This need not be so in all cases, however. It is also possible that the phenotype that is observed results from nucleotide changes affecting splicing, and an alternate nucleotide that leads to the same amino-acid change need not lead to the same alternate splicing event. To limit this risk, we did not calculate alternates for neighboring positions in triplets that span exon boundaries.

Track Name: hgmdimputed

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	HGMD® mutation accession. Cardinality: 1 MySQL Type: TEXT Examples: `CM960042`
alt	Alternative base Cardinality: 1 MySQL Type: TEXT
aminoacid_change	Aminoacid change represented in amino acid code Cardinality: 0..1 MySQL Type: TEXT
brief (feature)	The gene, disease and mutational change. Cardinality: 0..1 MySQL Type: TEXT Examples: `BRCA1;Breast_Cancer;134C>T`

citation_type	The nature of the information source for the mutation.
	Cardinality: 1..*
	MySQL Type: VARCHAR (11)
	Accepted Values:
	`Primary`
	All reported data is obtained from this paper.
	`LSDB`
	Some mutations are imported from LSDB, hence do not have a pmid.
	`LSDB Report`
	Some mutations are imported from LSDB, hence do not have a pmid.
	`APR`
	Additional phenotype report.
	`FCR`
	Functional characterization report.
	`MCR`
	Molecular characterization report.
	`SAR`
	Simple additional report.
	`FAPR`
	Functional characterization additional phenotype report.
	`ACR`
	Additional literature report.

codon_change	The codon changes effected because of the mutation. Cardinality: 0..1 MySQL Type: TEXT
	The affected codon in the gene sequence. This is also the position of the amino acid residue in the protein.

codon_number	Cardinality: 0..1 MySQL Type: BIGINT

comments	Any additional observations noted by the curators.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`aka G20566T`
	`aka IVS-II-654/c.316-197`

confidence	The strength of the evidence for the mutation/disease relationship.
	Cardinality: 1
	MySQL Type: VARCHAR (4)
	Accepted Values:
	`High`
	If curators had strong evidence.
	`Low`
	If curators had some reservation about the strength of the evidence.

disease	The associated disease or phenotype.
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`17-alpha-hydroxylase/17,20-lyase deficiency`
	`Mucopolysaccharidosis II`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT

genomic_sequence	30 residues upstream and downstream flanking region of the mutation described.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`tcatgttcatacctcttatcttcctcccac(a/g)gCTCCTGGGCAACGTGCTGGTCTGTGTGCT`
	`GATGCCAAACCAGGTCAATTCCCTTGGCAG(g/t)tactttatactgatggtgtgtcaaaactgg`
	`agaataacagtgataatttctgggttaagg(c/t)aatagcaatatctctgcatataaatatttc`

hgmd_acc (hgmdAcc)	HGMD ID Cardinality: 1 MySQL Type: TEXT Examples: `CS850009`

hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hgvs	The complete hgvs description for the mutation. Cardinality: 0..1 MySQL Type: TEXT Examples: `NM_000518.4: c.316-2A>G`
hyperlink	A static version of HGMD® professional mutation report (Full reports can be reached from this page. To get access to full HGMD® professional functionality and content, a separate license is required). Cardinality: 0..1 MySQL Type: TEXT
lsdb_source	The locus specific database (LSDB) in which the variant was reported. A small percentage of the mutations in HGMD does not originate from litereature reports, but from well documented reports in LSDBs. In those cases, the lsdb_source field will contain the URL of the LSDB, and pmid_citation_type will be LSDB Report. Cardinality: 0..1 MySQL Type: TEXT Examples: `http://www.genet.sickkids.on.ca/cftr/`

mutation_type (mutationType)	A one-letter code determining which class (and table in HGMD) the mutation belongs to. If there are several mutations of different type they are given as a comma-delimted list, such as *D,I,M* for a gene with at least one deletion, insertion and point mutation.
	Cardinality: 1..*
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	Deletion.
	`E`
	Amplet.
	`G`
	Gross deletion - refers to lesions covering more than 20 nucleotides.
	`I`
	Insertion.
	`M`
	Mutation (mis-sense or non-sense single nucleotide).
	`N`
	Gross Insertion/Deletion.
	`P`
	Complex Rearrangement.
	`R`
	Promoter mutation.
	`S`

	Splice site mutation. `X` Indel.
nucleotide_change (nucleotideChange)	A description of the nucleotide change in HGVS nomenclature. Cardinality: 0..1 MySQL Type: TEXT Examples: `316-2A>G`
omim	Accession number of the corresponding OMIM entry. Cardinality: 0..1 MySQL Type: BIGINT Examples: `300746`
pmid	Pubmed ID Cardinality: 0..* MySQL Type: BIGINT Primary Field: citation_type
pmid_notes	Any additional observations noted by the curators for a specific paper. This list of notes has a one to one correspondence with the list of pmids given in the pmid key. Cardinality: 0..* MySQL Type: TEXT Primary Field: citation_type
ref	Reference base Cardinality: 1 MySQL Type: TEXT
rsid	rsid of the SNP entry in dbSNP, corresponding to the mutation, where available. Cardinality: 0..1 MySQL Type: TEXT Examples: `rs33914668`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT
	The severity category of the variant. Cardinality: 1..* MySQL Type: VARCHAR (3) Accepted Values: `DP` Disease-associated polymorphism - A polymorphism reported to be in significant

variant_type (variantType)	association with a disease/phenotype (p<0.05) that is assumed to be functional (e.g. as a consequence of location, evolutionary conservation, replication studies etc), although there may as yet be no direct evidence (e.g. from an expression study) of a functional effect.
	
	`DFP`
	Disease-associated polymorphism with additional supporting functional evidence - A polymorphism reported to be in significant association with disease (p<0.05) that has evidence of being of direct functional importance (e.g. as a consequence of altered expression, mRNA studies etc).
	
	`FP`
	In vitro/laboratory or in vivo functional polymorphism - A polymorphism reported to affect the structure, function or expression of the gene (or gene product), but with no disease association reported as yet.
	
	`FTV`
	Frameshift or truncating variant - A polymorphic or rare variant reported in the literature (e.g. detected in the process of whole genome/exome screening) that is predicted to truncate or otherwise alter the gene product (i.e. a nonsense or frameshift variant) but with no disease association reported as yet. Please note that any variant affecting the obligate donor/acceptor splice site of a gene will not be included in this category unless there is evidence for an effect on the splicing phenotype. Variants occurring in pseudogenes will also be excluded unless evidence for a functional effect is present for both the pseudogene itself and the variant in question.
	
	`CNV`
	Copy number variations are DNA segments >1 kb in length that present with variable numbers of copies in a given population. These variants are being reported in the literature with an ever increasing frequency. CNVs are potentially functionally significant and should therefore in principle be treated by HGMD in a similar manner to any other polymorphism.
	
	`DM`
	Disease causing mutation - Pathological mutation reported to be disease causing in the corresponding report.
	
	`DM?`
	Disease causing mutation (report questionable) - mutation reported to be disease causing in the corresponding report, but where the author has indicated that there may be some degree of doubt, the curator had doubts about the validity of the claim given the data presented or subsequent evidence has come to light in the literature, calling the deleterious nature of the variant into question.
	
	`R`
	Removed - mutations that were removed from the database, for example because the report was erroneous or has been retracted. To allow users to track these changes, the records were not actually removed, but flagged as R, retaining all their other characteristics. These variants should not be used for annotation purposes.

PharmacoGenomic Mutation Database (Beta)

Version: 2014.3 (beta)

Track Description: This track contains variants that have been shown to exhibit a pharmacogenomic effect on patients. The data has been manually curated from the medical and research literature, and from official drug label information.

Variants can be single nucleotide polymorphisms, Insertions, Deletions, Indels, VNTRs (Variable Number Tandem Repeats), or entire haplotypes. We provide the phenotype and associated data such as dosage effects or ethnicity of the study population. We also provide supporting evidence such as number of cases and controls, and statistical significance of the correlation. In cases where a haplotype is associated with an effect, Genome Trax™ will report a hit, if your input data matched all of the variants that make up the haplotype.

Let us know if you have suggestions for improvement.

Benefit: These pharmacogenomic variants help you to identify variants that may influence how an individual reacts to certain drugs, and what dosage of drugs might be advisable. Please note that Genome Trax™ is NOT A DIAGNOSTIC TOOL, and these variants are only provided for research purposes. You should always consult your M.D. in regard to treatment options. DO NOT MAKE ANY MEDICAL DECISION BASED ON THE DATA PROVIDED HERE.

Track Name: pgmd

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION

accession	An accession number that is unique for this variant in its pharmacogenomic context
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`GV000003898`
	GV (Genome variation) identifiers for single location variants.
	`HP000000518`
	HP (Haplotype) identifiers for haplotypes and diplotypes.

age (pgmd_age)	Describes the age of the case group
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`42`
	`32-58`
	Age can be represented as a range, lower age - upper age of the group
	`>35`
	Age can be represented with qualifiers like '<', '>', '<=', '>=', etc

amino_acid	Amino acid change, called by snpEff. From snpEff documentation - Amino acid change: old_AA AA_position/new_AA Cardinality: 0..* MySQL Type: TEXT Examples: `E30K`
baseline_genotype_ind (pgmd_baseline_genotype_ind)	In genetic association studies, one allele/genotype/haplotype/diplotype is often considered as a baseline against which the others are compared. For example, the genotype G/G is the baseline against which the genotypes G/C and C/C show "Decreased clearance of metabolite". The baseline is not necessarily the most common case in the population under study, or the one listed in the reference build of the human genome, although this often may be the case. If there was no group for this observation, then the field is empty. Cardinality: 0..1 MySQL Type: VARCHAR (4) Accepted Values: `TRUE` The allele/genotype/haplotype/diplotype acts as the baseline in a group of observations.

	`N/A` The allele/genotype/haplotype/diplotype does not act as the baseline in a group of observations.

brief
(feature)
	Genotype:Phenotype. See description of those fields for more detail.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`C/G or C/C:Decreased risk of drug-induced extrapyramidal symptoms`
	`Short/Short:Decreased risk of drug-induced weight gain`

cases (pgmd_cases)	The total number of cases studied for this particular observation. Cardinality: 0..1 MySQL Type: BIGINT Examples: `18`
comments (pgmd_comments)	Free text annotation about sample details, statistical tests, corrections used, etc, that did not fit any of the other categories. The comments are prefixed by a classification tag as follows: • Primary statistical information • Secondary statistical information • Details about replication • Additional sample details • Variation details • Genotype, haplotype or diplotype details • Additional details • External reference Cardinality: 0..* MySQL Type: TEXT Examples: `Primary statistical information: Significance is based on two-tailed Fisher's exact test`

confidence_interval
(pgmd_confidence_interval)
	Confidence interval for the OR (95%) for a particular genotype.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`1-6.09`
	`0.01-0.66`

controls (pgmd_controls)	The total number of controls studied for this particular observation. Cardinality: 0..1 MySQL Type: BIGINT Examples: `18`

disease	Describes the associated disease, if any, in individuals from the case population, by MeSH term (if there is a matching MeSH term). Cardinality: 0..* MySQL Type: TEXT Examples: `Schizophrenia`

disease_mesh_id	MeSH-id(s) for the disease term(s) from disease.
	Cardinality: 0..*
	MySQL Type: TEXT
	Primary Field: disease
	Examples:
	`D012559`
	`D004827`

drug	The name of the compound, drug, substance, drug class or treatment therapy under investigation. Names for actual drugs are taken in order of preference from Drugbank, Pubchem compound, Pubchem substance, and MeSH. The identifiers given in the fields drugbank_id, pubchem_cid, and drug_mesh_id correspond to the names given here, in the same order. This field contains all drugs given to patients, even if the study does not indicate a specific association with each drug. For a list of only the drugs that were under investigation and shown to affect the observed phenotype, see focus_drug. MeSH drug classes are derived from the therapeutics tree of MESH. Drugs, drug classes or treatments that were applied to different study groups are separated by the pipe symbol (|), names within such a group are separated by semicolon (;). This means the field can contain several lists of drugs, drug classes or treatments if there were different patient study groups. Cardinality: 1..* MySQL Type: TEXT Examples: `Risperidone`
drug_mesh_id (pgmd_drug_mesh_id)	The MeSH ID(s) from the Chemicals and Drugs and the Therapeutics sections of MeSH for drug class or drug classes, and treatment or treatments applied. The order is the same as the order of names in drug, with empty positions where no ID is known. Cardinality: 0..* MySQL Type: BIGINT Primary Field: drug Examples: `5073`
	The drugbank ID(s) for the compound or

drugbank_id	compounds applied. The order is the same as the order of names in drug, with empty positions where no ID is known. Cardinality: 0..* MySQL Type: TEXT Primary Field: drug Examples: `DB00333`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
ethnicity (pgmd_ethnicity)	Describes the ethnicity of the cases by MeSH term. Cardinality: 0..* MySQL Type: TEXT Examples: `European Continental Ancestry Group`
ethnicity_mesh_id (pgmd_ethnicity_mesh_id)	MeSH-id(s) for the ethnicity term(s) from ethnicity. Cardinality: 0..* MySQL Type: TEXT Primary Field: ethnicity Examples: `D044465`

evidence (pgmd_evidence)	Describes the level of evidence, to aid users as a filter. The classification follows a scheme recommended by PharmGKB.
	Cardinality: 1
	MySQL Type: VARCHAR (40)
	Accepted Values:
	`Pharmacokinetics`
	Genetic variation in processes involved in the absorption, distribution, metabolism, or elimination of a drug can result in changes in drug availability.
	`Pharmacodynamics and Drug Response`
	Genetic variation in drug targets can cause measurable differences in the response of an organism to a drug.
	`Molecular and Cellular Functional Assays`
	Genetic variation can alter results of molecular and cellular functional assays, and this may correlate with

	variations in the organism's drug response. `Clinical Outcome` Genetic variations in the response to drugs can cause measurable differences in clinical endpoints such as rates of cure, morbidity, side effects, and death.
focus_disease (pgmd_focus_disease)	The disease that was the focus of the study. The difference between the focus_disease and disease field is that the disease field contains an exhaustive list of all diseases that the patients may have, regardless of whether or not the drug in question was targeting all of those diseases. Cardinality: 0..* MySQL Type: TEXT Examples: `HIV Infections, Hepatitis C`
focus_drug (pgmd_focus_drug)	All the drugs, drug classes or treatments that were the focus of the study, irrespective of co-medications used. This is a subset of the list in drug, consisting only of those which were shown to affect observed phenotypes. Cardinality: 0..* MySQL Type: TEXT Examples: `Risperidone`

genetic_model (pgmd_genetic_model)	Describes the genetic model used to analyze genotype-phenotype information from association studies.
	Cardinality: 0..1
	MySQL Type: VARCHAR (21)
	Accepted Values:
	`Allelic model`
	`General genetic model`
	`Dominant model`
	`Recessive model`
	`Co-dominant model`
	`Multiplicative model`
	`Additive model`
	`Haplotype`
	`Combined effect`
	`Not Applicable`
	`Global`
	`Site-specific effect`
	`Trend`
	`Combined genotype`

	The associated allele, genotype, haplotype or diplotype. If specific genotypes are not specified they are given as structured text. Note that if strictly the position itself has been associated with a phenotype, this field may be empty. A haplotype given here may result in multiple records. See also haplotype_id, site_genotype, non_carrier_ind, het_only_ind.

genotype
(pgmd_genotype)

Because of the many different ways to describe genotypes, please refer to the examples section for more detail. Square brackets are given just for readability and can be ignored.

Cardinality: 0..1

MySQL Type: TEXT

Examples:

`G`
If a single allele was implicated in a drug response, it will be represented as such, in a haploid manner. (variant_type is SNP).

`A/C`
Genotypes for a single nucleotide variant are represented by separating the allele by slash, always in alphabetical order. (variant_type is SNP).

`A/C or C/C`
Two or more genotypes that were classified as having the same effect are separated by an or. (variant_type is SNP).

`Ins{TTCAC}`
Insertion of a given sequence. (variant_type is Indel).

`Del{TC}`
Deletion of a given sequence. (variant_type is Indel).

`Del{T}Ins{GGC}`
An overlapping Indel, in which a sequence was deleted, while simultaneously another sequence was inserted. (variant_type is Indel).

`het[Del{TC}]`
The het tag indicates that the specified variation must occur in a heterozygous manner; the second copy must not necessarily be reference matching but it must not be the specified variation. (can be any variant_type).

`non[Del{TC}]`
The non tag indicates that anything other than the specified variation should result in a match. (can be any variant_type).

`Del{CFTR}`
Deletion of the CFTR gene. (variant_type is Gene deletion).

`Dup{CFTR}2`
Duplication of the CFTR gene. (variant_type is Gene duplication).

`Mul{KRAS}7`
Seven-fold multiplication of the CFTR gene. (variant_type is Gene multiplication).

`3R{CTTCCA}`
VNTR with three copies of the sequence CTTCCA. (variant_type is VNTR).

`>3R{CTTCCA}`
VNTR with more than 3 copies of the sequence CTTCCA. (variant_type is VNTR).

`<3R{CTTCCA}`
VNTR with less than three copies of the sequence CTTCCA. (variant_type is VNTR).

`19R`
VNTR with nineteen copies of undefined sequence. (variant_type is VNTR).

`3R{CTT}_5G`
A VNTR in which a SNP occurred as well. This example

	is equivalent to 'CTTCGTCTT' where the 5th base is a T>G change. (variant_type is VNTR).
	
	`Ref/Ref`
	Genotype at the position was homozygous reference. (Applies to many variant_types).
	
	`G-C-A-T-T`
	Haplotypes are represented by separating the allele of each variation by a hyphen. (can be any variant_type but will have a non-null haplotype_id).
	
	`0 matches (G-C-A-T-T)`
	Haplotypes in which you must match a certain subset of the sites within that haplotype. In this case, you must not be a the match for any of the given sites. (can be any variant_type but will have a non-null haplotype_id).
	
	`>1 match (G-C-A-T-T)`
	Haplotypes in which you must match a certain subset of the sites within that haplotype. In this case, you must match at least one of the sites. (can be any variant_type but will have a non-null haplotype_id).
	
	`2 to 4 matches (G-C-A-T-T)`
	Haplotypes in which you must match a certain subset of the sites within that haplotype. In this case, you must match at least 2, but no more than 4 sites. (can be any variant_type but will have a non-null haplotype_id).
	
	`<5 matches (G-C-A-T-T)`
	Haplotypes in which you must match a certain subset of the sites within that haplotype. In this case, you must match less than five of the given sites. (can be any variant_type but will have a non-null haplotype_id).
	
	`G/G-C/C`
	Diplotypes are represented by separating the genotype of each variation by a hyphen. This does not necessarily imply phasing. If the diplotypes are phased, | is used instead of /. (can be any variant_type but will have a non-null haplotype_id).

genotyping_source (pgmd_genotyping_source)	Describes the source tissue/cell used for genotyping. Cardinality: 0..1 MySQL Type: TEXT Examples: `Peripheral blood`
geography (pgmd_geography)	Describes the geographical provenance of the cases by MeSH term. When the case group includes individuals of more than one geographical region, each term is separated by a semicolon. Cardinality: 0..* MySQL Type: TEXT Examples: `Spain`
geography_mesh_id (pgmd_geography_mesh_id)	Describes the geographical provenance of the cases by MeSH term. When the case group includes individuals of more than one geographical region, each id is separated by a semicolon. Cardinality: 0..* MySQL Type: TEXT

	Primary Field: geography Examples: `D013030`

group_id
(pgmd_group_id)
	A number unique within a particular study that is used to group one or more genotypes that were observed for a particular drug response. Each studied variation will belong to one or more observation groups. E.g. Variation G>C leads to three possible genotypes. G/G, G/C, and C/C, studied for a phenotype (e.g. Increased drug toxicity) leading to these three observations being grouped together with a unique group id. If these same 3 genotypes were also studied for another phenotype (e.g. Response rate), those records would be assigned a new group id.
	Cardinality: 1
	MySQL Type: BIGINT
	Examples:
	`1`
	G/G versus G/C versus C/C were compared against each other for drug toxicity impact in one paper, so they are group ID 1.
	`2`
	G/G versus G/C versus C/C were also compared against each other for with respect to response rate in the same paper, so these observations will all fall into group 2.

haplotype_id
(pgmd_haplotype_id)
	If variants co-occuring at multiple sites were shown to lead to a certain phenotype, then those sites would be grouped as a haplotype (we include diplotypes here). A haplotype of the form "A-T-T-G" or a diplotype of the form "A/T-G/C" would be split into individual records per site, and would each share the same haplotype group ID in order to resolve back to the original haplotype.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`HP000000015-001`
	The 3 sites of rs2032582, rs1045642, and rs1128503 that constitute the C-A-C haplotype would each share this ID.
	`HP000000015-002`
	The 3 sites of rs2032582, rs1045642, and rs1128503 that constitute the C-G-T haplotype would each share this ID.

hazard_ratio (pgmd_hazard_ratio)	A measure of how often a particular event happens in one group compared to how often it happens in another group, over time. Often used in clinical trials to measure survival at any point in time in a group of patients who have been given a specific treatment compared to a control group given another treatment treatment or a placebo. Cardinality: 0..1 MySQL Type: INT

het_only_ind
(pgmd_het_only_ind)
	If this indicator is set to true, it signifies that the value in the Genotype field must be heterozygous in a subject in order for there to be a match. This flag will only be set for records that have been entered at the allele level. (e.g. if Genotype field has a value of "G" and het_only_ind is "TRUE", then a subject who had a "G/T" genotype would be a match, but a subject with a "G/G" genotype would not).
	Cardinality: 0..1
	MySQL Type: VARCHAR (4)
	Accepted Values:
	`TRUE`
	`N/A`

hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT

hgvs
(pgmd_hgvs)
	The Human Genome Variation Society (HGVS) description of the variation, or if not available, free text, following HGVS rules. Sometimes chromosomal coordinates must be specified instead. For more on HGVS nomenclature, see http://www.hgvs.org/mutnomen/. An observation can include more than one variation (e.g. for haplotypes and diplotypes).
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`NT_010783.15:g.10634882T>C`
	The default contig is NT
	`genedeletion{BRCA1}`
	CNVs/SVs are described as accurately as possible
	`chr10:96826971delT`
	Absolute coordinate representation

hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT
max_haplotype_matches (pgmd_max_haplotype_matches)	When haplotype_id is populated (the variation is a haplotype), this field may be populated. A populated value means that of the sites in a haplotype, y number of these sites must be a positive match to be a match for the given observation. See also min_haplotype_matches. Cardinality: 0..1 MySQL Type: BIGINT

	Examples: `3` In a haplotype of 5 sites (e.g. G-T-C-A-T), if a subject matches 3 of those sites (e.g. G-T-C-G-G) then they will be a match for this observation, but if they match 4 or all of those sites, they will not.

metabolizer
(pgmd_metabolizer)
	Metabolizer status of the haplotype or diplotype.
	Cardinality: 0..1
	MySQL Type: VARCHAR (25)
	Accepted Values:
	`Slow Metabolizers`
	`Poor Metabolizers`
	`Intermediate Metabolizers`
	`Moderate Metabolizers`
	`Rapid Metabolizers`
	`Extensive Metabolizers`
	`Ultra-rapid Metabolizers`
	`Normal Metabolizers`

min_haplotype_matches
(pgmd_min_haplotype_matches)
	When haplotype_id is populated (the variation is a haplotype), this field may be populated. A populated value means that of the sites in a haplotype, x number of these sites must be a positive match to be a match for the given observation. See also max_haplotype_matches.
	Cardinality: 0..1
	MySQL Type: BIGINT
	Examples:
	`0`
	A value of 0 means that if a subject has no matches for the given haplotype, they are a match for the given observation.
	`2`
	In a haplotype of 3 sites (e.g. A-T-T), if a subject has a match for only 1 or 0 of those sites, then they will not be a match here.

named_variation (pgmd_named_variation)	Regularly known as "Star alleles". In named variants, genotypes are represented with gene names followed by allele designation e.g. CYP2C9*1/CYP2C9*1. We have resolved these star alleles to the actual site(s) that make them up. In cases where a publication has specified the sites that they considered for a given named variation, we use just those sites. In cases where a publication has simply referred to a variation by its star allele nomenclature, we have resolved that variation to the sites for that named variation that differ from the wild-type (*1) allele. In cases that were reported as heterozygous (e.g. CYP2C9*1/*3), rather than considering all sites that have been observed for *1, we only consider the sites where *3 has varied. For homozygous reference, (*1/*1), we consider all sites that have seen variation within the set of named

	variations. To see what the genotypes were resolved to, see also genotype. Cardinality: 0..1 MySQL Type: TEXT Examples: `CYP2C9*1/CYP2C9*6`

nearby_genes (pgmd_nearby_genes)	In addition to the single HGNC gene symbol that gets assigned in the HGNC field, this contains a more complete description of gene symbols that are associated with the variant, based on the variant's position in the genome. Contains a single HGNC symbol, if the variant overlaps a single gene. When a variant overlaps multiple genes, each gene symbol is separated by a semicolon (and the strand of the gene may be given as (+) or(-). If the site is intergenic, it will contain the 4 nearest genes (5' and 3' on the positive and negative strands, and a signed distance from each gene, - meaning towards smaller genomic coordinates, + towards larger ones, separated by comma.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`SLCO1B3`
	A single gene overlaps the variant.
	`SLCO1B3 (+), SLCO1B4 (-)`
	Two genes, on opposite strands overlap the variant.
	`SLCO1B3 (+) -17000, SLCO1B4 (+) +42000, SLCO1B5 (-) -10000, SLCO1B6 (-) +27000`
	The variant is intergenic, the four neighboring genes are given with strand and distance.

non_carrier_ind (pgmd_non_carrier_ind)	If this indicator is set to true, it signifies that the respective record applies only to subjects that do not carry the allele, genotype, haplotype, or diplotype specified in the genotype field. (e.g. if Genotype field has a value of "A" and non_carrier is "true", then a subject who had a "T/T" genotype would be a match, but a subject with a "T/A" genotype would not). See also genotype.
	Cardinality: 0..1
	MySQL Type: VARCHAR (4)
	Accepted Values:
	`TRUE`
	`N/A`

obsid (pgmd_obsid)	A globally unique identifier for each observation that has been curated. This is assigned based on the site/allele/genotype/haplotype/diplotype that has been associated with a specific effect. An observation may span number of genomic sites depending on the number of variants acting together as a haplotype.

	Cardinality: 1 MySQL Type: BIGINT Examples: `3361673`

odds_ratio
(pgmd_odds_ratio)
	The odds that an individual with this genetic profile will actually exhibit this phenotype.
	Cardinality: 0..1
	MySQL Type: INT
	Examples:
	`0.29`
	`2.47`

p_value
(pgmd_p_value)
	P-value for a particular genotype as given in the reference.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`0.05`
	`<0.01`
	If the reference states a p-value as being lower than a given value, then the value is prefixed by the 'less than' sign.

phenotype (pgmd_phenotype)	A qualitative description of the impact of genetic variation on drug response. Cardinality: 1 MySQL Type: TEXT Examples: `Decreased risk of drug-induced extrapyramidal symptoms.`
phenotype_category (pgmd_phenotype_category)	The general category of drug response, chosen from our controlled vocabulary of phenotypes. Cardinality: 0..1 MySQL Type: TEXT Examples: `extrapyramidal symptoms`
phenotype_detail (pgmd_phenotype_detail)	A quantitative description of the impact of genetic variation on drug response, typically detailing the fraction of subjects with the specified genetic profile that exhibited the given response, and further detail on that response that would be given in the Phenotype field. Cardinality: 0..1 MySQL Type: TEXT Examples: `Oral clearance of Verapamil for G/G-C/C diplotype is 452.2 +/- 188.6l/hr`
	Pubmed ID of the reference from which the information was taken.

pmid	Cardinality: 0..* MySQL Type: BIGINT
pubchem_cid	The PubChem compound CID(s) for the compound or compounds administered. The order is the same as the order of names in drug, with empty positions where no ID is known. Cardinality: 0..* MySQL Type: BIGINT Primary Field: drug Examples: `5073`

ref_id	The reference from where this curation came. *See also ref_type.
	Cardinality: 1
	MySQL Type: BIGINT
	Primary Field: ref_type
	Examples:
	`11586955`
	A PubMed Identifier
	`Abacavir-03/04/14`
	The abacavir FDA drug label, curated March 4, 2014

ref_type	The type of reference that ref_id refers to.
	Cardinality: 1
	MySQL Type: VARCHAR (9)
	Accepted Values:
	`pubmedid`
	Source is from PubMed (http://www.ncbi.nlm.nih.gov/pubmed/).
	`fda_label`
	Source is an FDA drug label (http://www.accessdata.fda.gov/scripts/cder/drugsatfda/).

reference_allele (pgmd_reference_allele)	The allele found in the corresponding human reference assembly. Cardinality: 0..1 MySQL Type: TEXT Examples: `T`
registry_identifiers (pgmd_registry_identifiers)	Official approval or recommendation like a clinicaltrials.gov number. Currently included are ClinnicalTrials.gov, EudraCT, NCCTG, EDCTP, ACTG, ACTR, Chinese Clinical Trial Registry Number, ISRCTN Register, UMIN-CTR registration, and Netherlands trial registry. Cardinality: 0..* MySQL Type: TEXT Examples: `ClinicalTrials.gov Identifier:NCT00006206`

	`ACTR Number:12610000270011`
relative_risk (pgmd_relative_risk)	The likelihood that an individual with this genetic profile will exhibit this phenotype versus the likelihood that someone who does not have this genetic profile will exhibit this phenotype. Cardinality: 0..1 MySQL Type: INT
rsid	dbSNP ID number, if available. Cardinality: 0..* MySQL Type: TEXT Examples: `rs6280`
sample_size (pgmd_sample_size)	Describes the total sample size in the study. Sum of cases and controls across all genotypes. Cardinality: 0..1 MySQL Type: BIGINT Examples: `111`

sex
(pgmd_sex)
	Describes the gender of the individuals in the study.
	Cardinality: 1
	MySQL Type: VARCHAR (11)
	Accepted Values:
	`Female`
	`Male`
	`Mixed`
	`Unspecified`

	The allele or genotype for a single position that has been derived from the genotype column. In the case of haplotypes, genotype will represent the full haplotype for an observation (e.g. T-A-G), which will then be split into multiple records in site_genotype, linked through the obsid and haplotype group id to resolve the overall haplotype. In the case of multiple genotypes that were all associated with the same drug response, genotype will represent the grouped genotypes (e.g. "A/A or A/G"), which will then be split into multiple records, but will share the same observation id (obsid). Cardinality: 0..1 MySQL Type: TEXT Examples: `T` If Genotype is simply 'T', site_genotype will be the same. If Genotype is 'T or C', this observation will be split into 2 records, having site_genotype of 'T' for entry 1, and 'C' for 2. If Genotype is 'T-G-A' (a haplotype of 3 sites), this

site_genotype (pgmd_site_genotype)	observation will be split into 3 records, having site_genotype of 'T' for the first site, 'G' for the second, and 'A' for the third.
	
	`A/C`
	If Genotype is simply 'A/C', site_genotype will be the same. If Genotype is 'A/C or C/C', this observation will be split into 2 records, having site_genotype of 'A/C' for entry 1, and 'C/C' for 2. If Genotype is 'A/C-G/G-A/A' (a diplotype of 3 sites), this observation will be split into 3 records, having site_genotype of 'A/C' for the first site, 'G/G' for the second, and 'A/A' for the third.
	
	`>16R{CA}/>16{CA}`
	VNTRs are maintained as-is due to the complexity of expanding VNTRs in which a range of repeats was specified.
	
	`16R{CA}`
	VNTRs are maintained as-is due to the complexity of expanding VNTRs in which a range of repeats was specified.
	
	`<16R{CA}`
	VNTRs are maintained as-is due to the complexity of expanding VNTRs in which a range of repeats was specified.
	
	`->TTCAC`
	A Genotype of 'Ins{TTCAC}' (insertion of TTCAC sequence) will be represented as such.
	
	`TC>-`
	A Genotype of 'Del{TC}' (deletion of TC sequence) will be represented as such.
	
	`T>GGC`
	A Genotype of 'Del{T}Ins{GGC}' (overlapping deletion of T and insertion of GGC at the same site) will be represented as such.

study_design (pgmd_study_design)	Describes the study type. In the case of several possible terms, the most specific one that indicates the highest predictive power will be used. For example, for randomized, controlled clinical trials, if the reference mentions several stages, then the highest stage will be used.
	
	Cardinality: 1
	
	MySQL Type: VARCHAR (64)
	
	Accepted Values:
	`Randomized Controlled Clinical Trial`
	`Randomized Controlled Clinical Trial (Clinical Trial, phase I)`
	`Randomized Controlled Clinical Trial (Clinical Trial, phase II)`
	`Randomized Controlled Clinical Trial (Clinical Trial, phase III)`
	`Randomized Controlled Clinical Trial (Clinical Trial, phase IV)`
	`Clinical Trial (general, phases unknown)`
	`Clinical Trial, phase I`
	`Clinical Trial, phase II`
	`Clinical Trial, phase III`
	`Clinical Trial, phase IV`
	`Case-Control Study`
	`Case Series`
	`Genome-Wide Association Study`
	`Intervention Study`
	`Meta-Analysis`

	`Preclinical Study`
	`Prospective Study`
	`Replication Study`
	`Retrospective Study`
	`Retrospective study`

treatment_detail (pgmd_treatment_detail)	Dose, duration and the route of administration of the compounds used for treatment in the case group. Cardinality: 0..1 MySQL Type: TEXT Examples: `Methotrexate was given to all patients for at least 3 months, with an initial dose of 4-5 mg/2 months/week and then up-to maximal dosage of 10 mg/2 months/week by oral mode. Prednisolone was also given to 89 patients and nine patients were supplemented with Folic acid`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT
variant_class (pgmd_variant_class)	Class of variant, as predicted by snpEff, based on the "canonical" transcript. Values include missense, nonsense, synonymous, and frameshift; please refer to snpEff for a full vocabulary. Cardinality: 0..* MySQL Type: TEXT

variant_type
(pgmd_variant_type)
	Variant type.
	Cardinality: 0..*
	MySQL Type: VARCHAR (19)
	Accepted Values:
	`SNP`
	Single Nucleotide Polymorphism
	`Indel`
	Insertion, Deletion, or Overlapping Insertion and Deletion
	`VNTR`
	Variable Number of Tandem Repeats
	`Gene deletion`
	Deletion of entire gene
	`Gene duplication`
	Duplication of gene
	`Gene amplification`
	Amplification event
	`Gene multiplication`
	Multiplication event
	`Polymorphism`
	Applies to polymorphisms of unknown type.

ClinVar Variants

Version: Clinvar-2014-09

Track Description: This track contains data from ClinVar. ClinVar is a public archive of reports that lists relationship between human variations and phenotypes with supporting evidence. Thus ClinVar facilitates access to and communication about the relationships asserted between human variation and observed health status, and how interpretation of variation may change over time. ClinVar collects reports of variants found in patient samples, assertions made regarding their clinical significance, information about the submitter, and other supporting data. The alleles described in the submissions are mapped to reference sequences, and reported according to the HGVS standard.

Benefit: This data set contains experimentally observed, clinically significant variants that are reviewed by experts.

Track Name: clinvar

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession (clinvar_Accession)	The accession as given in ClinVar for a reviewed assertion. Cardinality: 1 MySQL Type: TEXT Examples: `RCV000032585`

age_of_onset (clinvar_AgeOfOnset)	The age of onset of the disease as observed in the samples analyzed.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`Adulthood`
	`Childhood`

allele_id	The allele ID, or ID of the overall ClinVarSet, which packages the ReferenceClinvarAssertion where most of the data is from and its supporting submitted ClinVarAssertions. Cardinality: 1 MySQL Type: BIGINT Examples: `93236`
alt	Alternative base Cardinality: 0..* MySQL Type: TEXT
brief (feature)	HGVS description and the phenotype, colon-separated (Note that the HGVS description also does contain a colon. Cardinality: 0..1 MySQL Type: TEXT Examples:

	`NT_011109.15:g.14128514A>G:Diaphyseal dysplasia` `NG_016363.1:g.5096_5097delCT:Dyskeratosis congenita autosomal dominant`

clinical_significance (clinvar_ClinicalSignificance)	The clinical significance as observed in the study. The values of clinical signficance ClinVar represents are provided only by the submitter, and used to calculate conflicts in interpretation when all submissions about the same variation and disorder are aggregated. The list of significance terms is maintained here. Data submitted by OMIM does not include an interpretation of clinical significance. These submissions are mapped into clinical significance values according to the rules described here. Conflicts are reported when the submitted severities on the 3-point scale of Pathogenic/Likely Pathogenic - Unknown significance - Likely Benign/Benign differ. Clinvar is not consistent in their naming. Below we also provide alternative names they use, in case you want to cross compare.
	Cardinality: 0..*
	MySQL Type: VARCHAR (32)
	Accepted Values:
	`pathogenic`
	top on 5-grade scale of pathogenicity. Also reported as '5 - pathogenic' (lowercase)
	`likely pathogenic`
	runner-up on 5-grade scale of pathogenicity. Also reported as '4 - probable-pathogenic'.
	`uncertain significance`
	middling on 5-grade scale of pathogenicty. Also reported as '0 - unknown'. Often seen as the acronym VUS (variant of unknown significance) in literature.
	`likely benign`
	second-to-last on 5-grade scale of pathogenicty. Also reproted as '3 - probable-non-pathogenic'.
	`benign`
	bottom on 5-grade scale of pathogenicty. Also reported as '2 - non-pathogenic', or no known pathogenicity.
	`drug response`
	Also reported as '6 - drug-response' (with hyphen)
	`association`
	Also reported as '255 - other'
	`risk factor`
	Also reported as '255 - other'
	`protective`
	Also reported as '255 - other'
	`confers sensitivity`
	Also reported as '255 - other'
	`other`
	Also reported as '255 - other'
	`not provided`
	Also reported as '1 - untested'.
	`conflicting data from submitters`
	Also reported as '1 - untested'.

date_last_evaluated	Datestamp of the last evaluation of the variant. Cardinality: 0..1 MySQL Type: TEXT Examples: `2012-12-21`

disease (clinvar_DiseaseName)	The associated phenotype or disease, sometimes with additional information on the form or inheritance mode.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`Dyskeratosis congenita autosomal dominant`
	`Idiopathic fibrosing alveolitis, chronic form`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
gene_reviews	Accession of a GeneReviews record for the variant/gene. Cardinality: 0..* MySQL Type: TEXT Examples: `NBK1298`
guideline	Cardinality: 0..* MySQL Type: TEXT Examples: `ACMG`
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
	Cardinality: 0..* MySQL Type: TEXT

hgvs	Examples:
	`NG_016363.1:g.5096_5097delCT`
	`NG_016363.1:g.5098G>A`

hyperlink	An individual variant report in ClinVar site at NCBI. Cardinality: 0..1 MySQL Type: TEXT

measure_type	Kind of lesion. Note that variants will not be represented in the data, if they do not have adequate, unique genomic coordinates.
	Cardinality: 0..1
	MySQL Type: VARCHAR (25)
	Accepted Values:
	`copy number gain`
	`copy number loss`
	`deletion`
	`duplication`
	`fusion`
	`indel`
	`insertion`
	`inversion`
	`microsatellite`
	`protein only`
	`single nucleotide variant`
	`structural variant`
	`variation`
	If nothing more specific is known

medgen	Connects the observed phenotype to the MedGen disease if present. Cardinality: 0..* MySQL Type: TEXT Examples: `CN181336`

	The impact of the variation on the sequence as calculated per transcript by NCBI. Confirms to the sequence ontology terminology.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`2KB_upstream_variant`
	`3_prime_UTR_variant`
	`500B_downstream_variant`
	`5_prime_UTR_variant`
	`exon_loss`
	`frameshift_variant`
	`inframe_variant`
	`intergenic_variant`
	`intron_variant`
	`missense_variant`
	`nc_transcript_variant`

molecular_consequence (clinvar_MolecularConsequence)	`splice_acceptor_variant`
	`splice_donor_variant`
	`stop_gained`
	`stop_lost`
	`synonymous_variant`
	`Splice Site`
	`nearGene-3`
	`intron`
	`frameshift`
	`missense`
	`STOP-GAIN`
	`nearGene-5`
	`UTR-3`
	`UTR-5`
	`cds-synon`
	`cds-indel`
	`ncRNA`
	`splice-3`
	`STOP-LOSS`
	`splice-5`
	`Frameshift`
	`nonsense`
	`Missense`

number_submitters	Number of submissions with this variant. Cardinality: 0..1 MySQL Type: BIGINT Examples: `3`
omim	Connects the observed phenotype to the OMIM disease if present. Cardinality: 0..* MySQL Type: BIGINT Examples: `131300`

origin (clinvar_Origin)	A list of alleleic origins for this variant.
	Cardinality: 0..*
	MySQL Type: VARCHAR (12)
	Accepted Values:
	`germline`
	Encompasses inherited or de-novo
	`inherited`
	Encompasses paternal and maternal
	`paternal`
	`maternal`
	`de novo`
	`somatic`
	`uncertain`
	`not provided`

orpha	Orphanet id for the disease. Cardinality: 0..* MySQL Type: BIGINT

	Examples: `1328`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT

prevalence (clinvar_Prevalence)	The Prevalence of the variation.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`1-5/10000`
	`<1/1000000`

ref	Reference base Cardinality: 0..1 MySQL Type: TEXT

review_status (clinvar_ReviewStatus)	Indicates the level of confidence for an assertion. The conflicts are calculated by NCBI if there are multiple submissions for the same phenotype/allele relationship.
	Cardinality: 0..*
	MySQL Type: VARCHAR (33)
	Accepted Values:
	`reviewed by expert panel`
	`reviewed by professional society`
	`classified by single submitter`
	`classified by multiple submitters`
	`not classified by submitter`

rsid	dbSNP rsid for the variant. Cardinality: 0..1 MySQL Type: TEXT Examples: `rs1800469`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

GWAS Catalogue

Version: downloaded on 11th September 2014

Track Description: This track contains data from the GWAS Catalogue [1]. These are literature derived disease associations for polymorphisms from GWAS studies that assayed at least 100,000 single nucleotide polymorphisms, associations listed are limited to those with p-values < 1.0×10^{-5}. The dataset provides Odds Ratios for common variants that can be used to calculate increased or decreased risk for the

disease. A detailed description of the methods to assemble the dataset can be found in Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. Potentialetiologic and functional implications of genome-wide association loci for human diseases and traits.

Proc Natl Acad Sci USA. May 27, 2009., available http://www.genome.gov/pages/about/od/newsandfeatures/pnasgwasonlinecatalog.pdf, and at the GWAS Catalogue at www.genome.gov/gwastudies.

Benefit: These disease association data are manually curated, experimentally determined associations from the scientific literature, mapped to coordinates. They allow you to identify common SNPs that influence the risk for common diseases.

Track Name: gwas

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	dbSNP rsid Cardinality: 1 MySQL Type: TEXT
alt	Alternative base Cardinality: 0..* MySQL Type: TEXT
brief (feature)	The disease, risk allele, and odds-ratio or beta (denoted by OR or beta). Cardinality: 0..1 MySQL Type: TEXT Examples: `Parkinson's disease:rs7702187-?:1.74`

ci_95 (gwas_95pct_CI)	Reported 95% confidence interval associated with best SNP risk allele.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`- 7.90 [NR] msec difference between homozygotes`
	`[1.36-2.24]`

cnv (gwas_CNV)	Study of copy number variation.
	Cardinality: 1
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`N`
	`Y`

	SNP functional class.
	Cardinality: 0..*
	MySQL Type: VARCHAR (10)
	Accepted Values:
	`Intergenic`
	`UTR-3`

context
(gwas_context)
	`intron`
	`UTR-5`
	`cds-synon`
	`nearGene-3`
	`missense`
	`nearGene-5`
	`ncRNA`
	`STOP-GAIN`
	`frameshift`
	`splice-3`
	`splice-5`

disease
(gwas_disease)
	Disease or trait examined in study.
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`Age-related macular degeneration`
	`Parkinson's disease`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	dbSNP record. As the GWAS catalog does not provide reports for the individual SNPs, we link to dbSNP instead. Cardinality: 0..1 MySQL Type: TEXT

initial_sample_size
(gwas_initial_sample_size)
	Sample size for Stage 1 of GWAS.
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`443 sib pairs`
	`96 European ancestry cases,50 European ancestry controls`

or_or_beta (gwas_OR_or_beta)	Reported odds ratio or beta coefficient associated with best SNP risk allele Cardinality: 0..1 MySQL Type: INT Examples: `1.74`
p_value (gwas_p_value)	Cardinality: 0..1 MySQL Type: TEXT Examples: `8E-6`
p_value_context (gwas_p_value_context)	Information describing context of p-value. Cardinality: 0..1 MySQL Type: TEXT Examples: `females, smokers`

platform
(gwas_platform)
	Genotyping platform manufacturer used in Stage 1; also includes notation of pooled DNA study design or imputation of SNPs, where applicable.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`Affymetrix[103,611]`
	`Perlegen [198,345]`

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
ref	Reference base Cardinality: 0..1 MySQL Type: TEXT
region (gwas_region)	Cytogenetic region associated with rs number Cardinality: 0..1 MySQL Type: TEXT Examples: `5p15.31`

replication_sample_size
(gwas_replication_sample_size)
	Sample size for subsequent replication(s).
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`200 > 85th pct,200 < 15th pct,7,817 cohort members`
	`332 cases,332 controls`

	Gene(s) reported by author.

reported_gene (gwas_reported_gene)	Cardinality: 0..* MySQL Type: TEXT Examples: `NOS1AP`
risk_allele	The strongest allele associated with the trait. Cardinality: 0..1 MySQL Type: TEXT

risk_allele_frequency
(gwas_risk_allele_frequency)
	Reported risk allele frequency associated with best SNP.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`0.36`
	`0.70 (HapMap CEU)`

snps
(gwas_snps)
	Best SNP
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`rs380390`
	`rs10494366`

uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

References:

1. Hindorff LA, Junkins HA, Hall PN, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: http://www.genome.gov/gwastudies.

COSMIC somatic disease mutations

Version: 70

Track Description: This track contains data from the Catalogue of Somatic Mutations in Cancer (COSMIC)[1].

COSMIC contains somatic mutation information relating to human cancers. The mutation data and associated information is extracted from the primary literature and entered into the COSMIC database. In order to provide a consistent view of the data a histology and tissue ontology has been created and all mutations are mapped to a single version of each gene. A central aim of COSMIC is to provide somatic mutation frequencies. This track contains SNPs, insertions and deletions from COSMIC.

We include COSMIC mutations for which a chromosomal position can be determined. The percentage of mutations with position is approximately 75%.

Benefit: These somatic mutations complement the set of germ-line mutations from HGMD to allow for a more comprehensive assessment of prior knowledge about observed mutations.

Track Name: cosmic

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
aa_mutation (aaMutation)	The protein sequence mutation as given in COSMIC in HGVS nomenclature. Cardinality: 0..1 MySQL Type: TEXT Examples: `p.K153N`
accession	COSMIC Mutation ID. Cardinality: 1 MySQL Type: TEXT
alt	Alternative base Cardinality: 0..* MySQL Type: TEXT
brief (feature)	The histology and mutational change. Cardinality: 0..1 MySQL Type: TEXT Examples: `carcinoma:c.775G>T`
cds_mutation (cdsMutation)	The coding sequence mutation as given in COSMIC in HGVS nomenclature. Cardinality: 1 MySQL Type: TEXT Examples: `c.459G>C`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
	The associated tumor histology, as given in COSMIC. Cardinality: 0..*

histology	MySQL Type: TEXT Primary Field: sample_name Examples: `carcinoma`
histology_subtype (histologySubtype)	The associated tumor histology subtype, as given in COSMIC. Cardinality: 0..* MySQL Type: TEXT Primary Field: sample_name Examples: `squamous_cell_carcinoma`
hyperlink	An individual mutation report in COSMIC site at the Welcome Trust Sanger Institute. Cardinality: 0..1 MySQL Type: TEXT

mutation_effect	Type of mutation, and its effect on transcription.
	Cardinality: 0..1
	MySQL Type: VARCHAR (31)
	Accepted Values:
	`Complex`
	`Complex - compound substitution`
	`Complex - deletion inframe`
	`Complex - frameshift`
	`Complex - insertion inframe`
	`Deletion - Frameshift`
	`Deletion - In frame`
	`Frameshift`
	`Insertion - Frameshift`
	`Insertion - In frame`
	`Mutation Description`
	`No detectable mRNA/protein`
	`Nonstop extension`
	`Substitution - coding silent`
	`Substitution - Missense`
	`Substitution - Nonsense`
	`Unknown`
	`Whole gene deletion`

mutation_status	Information on whether the sample was reported to be Confirmed Somatic, Previously Reported or Variant of unknown origin.
	Cardinality: 0..1
	MySQL Type: VARCHAR (44)
	Accepted Values:
	`Confirmed germline variant`
	If the mutation has been confimed to be germline in the experiment by sequencing both the tumour and a matched normal from the same patient.
	`Confirmed somatic variant`
	If the mutation has been confimed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient.
	`Not specified`
	`Reported in another cancer sample as somatic`

	Mutation has been reported as somatic in a previous paper but not in the current paper.
	`Reported in another sample as germline`
	Mutation has been reported as germline in a previous paper but not in the current paper
	`Variant of unknown origin`
	When the mutation is known to be somatic but the tumour was sequenced without a matched normal.

mutation_zygosity	Information on whether the mutation was reported to be homozygous , heterozygous or unknown within the sample.
	Cardinality: 0..1
	MySQL Type: VARCHAR (3)
	Accepted Values:
	`hom`
	Homozygous
	`het`
	Heterozygous

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT Primary Field: sample_name
primary_site (primarySite)	The associated tumor site, as given in COSMIC. Cardinality: 0..* MySQL Type: TEXT Primary Field: sample_name Examples: `lung`
ref	Reference base Cardinality: 0..1 MySQL Type: TEXT
sample_name	The sample name as given in COSMIC. The sample name gives a better description about the sample than the internal sample accessions COSMIC assigns. Cardinality: 1..* MySQL Type: TEXT
site_subtype (siteSubtype)	The associated tumor site subtype, as given in COSMIC. Cardinality: 0..* MySQL Type: TEXT Primary Field: sample_name Examples: `rib`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..*

	MySQL Type: TEXT

References:

1. The mutation data was obtained from the Sanger Institute Catalogue Of Somatic Mutations In Cancer web site, http://www.sanger.ac.uk/cosmic.

2. Bamford *et al* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*, 91, 355-358.

EVS Exome Variations

Version: ESP6500

Track Description: The EVS annotation source contains exome sequencing variants retrieved from the Exome Variant Server (EVS) for NHLBI Exome Sequencing Project (ESP) [1]. In the EVS data release ESP6500, the dataset comprised of a set of 2203 African-Americans and 4300 European-Americans unrelated individuals, totaling 6503 samples (13,006 chromosomes).

All data were simultaneously analyzed for exome variants at the University of Michigan (Abecasis Laboratory). The methods used for analysis is explained in detail at http://evs.gs.washington.edu/EVS/

Benefit: EVS provides the population based genotype, allele counts and MAF scores for the variations observed in exome regions.

Track Name: evs

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	a uniqe number identifying the EVS record. Cardinality: 1 MySQL Type: TEXT Examples: `EVS2265387`
african_american_allele_count (evs_AfricanAmericanAlleleCount)	The observed allele counts for the listed alleles in African American population. (delimited by /). Cardinality: 1 MySQL Type: TEXT Examples: `A=1/G=4405`
african_american_genotype_count (evs_AfricanAmericanGenotypeCount)	The observed genotype counts for the listed genotypes in African American population. (delimited by /). Cardinality: 1 MySQL Type: TEXT Examples: `TT=0/TC=308/CC=1895`
all_allele_count (evs_AllAlleleCount)	The observed allele counts for the listed alleles in all populations. (delimited by /). Cardinality: 1

	MySQL Type: TEXT Examples: `A=1/G=12997`
all_genotype_count (evs_AllGenotypeCount)	The observed genotype counts for the listed alleles in all populations. (delimited by /). Cardinality: 1 MySQL Type: TEXT Examples: `AA=0/AG=9/GG=5947`
alleles (evs_Alleles)	The observed alleles (delimited by /). Cardinality: 1 MySQL Type: TEXT Examples: `A/G`
alt	Alternative base Cardinality: 0..* MySQL Type: TEXT
avg_sample_read_depth (evs_AvgSampleReadDepth)	The average sample read depth. Cardinality: 0..1 MySQL Type: BIGINT Examples: `103`
brief (feature)	The allele of the NCBI human reference sequence (and hg19). Cardinality: 0..1 MySQL Type: TEXT Examples: `C`
cdna_pos (evs_cDNAPos)	The corresponding cDNA position for a SNP. Cardinality: 0..1 MySQL Type: INT Examples: `559`

chimp_allele
(evs_ChimpAllele)
	Chimp alleles are acquired from UCSC human/chimp alignment files.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`A`
	`unknown`
	The variation does not fall within an alignment block, or if it's an indel.
	`-`

	The variation falls within a gap in the alignment.
clinical_info (evs_ClinicalInfo)	The potential clinical implications associated with a SNP (limited). Cardinality: 0..1 MySQL Type: TEXT Examples: `http://www.ncbi.nlm.nih.gov/pubmed?term=21459883`
conservation_score_phast_cons (evs_ConservationScorePhastCons)	A number between 0 and 1 that describes the degree of sequence conservation among 17 vertebrate species; these numbers are downloaded from the UCSC Genome site and are defined as the "posterior probability that the corresponding alignment column was generated by the conserved state of the phylo-HMM, given the model parameters and the multiple alignment". Cardinality: 0..1 MySQL Type: INT Examples: `0.3`
dbsnp_version (evs_dbSNPVersion)	dbSNP version which established the rs_id. Cardinality: 0..1 MySQL Type: TEXT Examples: `dbSNP_134`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
european_american_allele_count (evs_EuropeanAmericanAlleleCount)	The observed allele counts for the listed alleles in African American population. (delimited by /). Cardinality: 1 MySQL Type: TEXT

	Examples: `T=11/C=8573`
european_american_genotype_count (evs_EuropeanAmericanGenotypeCount)	The observed genotype counts for the listed genotypes in European American population. (delimited by /). Cardinality: 1 MySQL Type: TEXT Examples: `TT=0/TC=11/CC=4281`
evs_conservation_score_gerp (evs_ConservationScoreGERP)	The rejected-substitution score from the program GERP, a number between -11.6 and 5.82 that describes the degree of sequence conservation among 34 mammalian species, with 5.82 being the most conserved; these scores were provided by Gregory M. Cooper of the University of Washington Department of Genome Sciences to the EVS project. Cardinality: 0..1 MySQL Type: INT
filter_status (evs_FilterStatus)	A machine-learning technique called support vector machine (SVM) classification was applied for variant filtering. After the initial SNP calls were generated, we re-examined the BAM files to collect additional information about each variant site. Based on the information, variants are initially filtered by individual thresholds. For example, variants with posterior probability <99% (glfMultiples SNP quality <20), were <5bp away from an indel detected in the 1000 Genomes Pilot Project, had total depth across samples of >5,379 or >5,379,000 reads (~1-1000 reads per sample), having >65% of reads as heterozygotes carrying the variant allele or where the absolute squared correlation between allele variant or reference) and strand (forward or reverse) was >0.15 were marked as problematic SNPs. Sites failed 3 or more criteria are used as negative examples to train SVM classifier. HapMap and OMNI polymorphic sites were used as positive examples. The SVM classifier produces scores for each site, and we marked ~8.5% of sites at threshold 0.3 as SVM filter-failed. The unfiltered set had Ti/Tv = 2.63, and the filtered set had Ti/Tv = 2.78. Cardinality: 0..1 MySQL Type: VARCHAR (4) Accepted Values:

	`PASS` `FAIL`

function_gvs
(evs_FunctionGVS)
	The GVS functions are calculated by the Exome Variant Server; they are based on the alleles for all populations and individuals; the bases in the coding region are divided into codons (if a multiple of 3), and the resulting amino acids are examined.
	Cardinality: 1
	MySQL Type: VARCHAR (29)
	Accepted Values:
	`alternate`
	`coding`
	`codingComplex`
	`coding-notMod3`
	`coding-notMod3-near-splice`
	`coding-synonymous`
	`coding-synonymous-near-splice`
	`frameshift`
	`intergenic`
	`intron`
	`missense`
	`missense-near-splice`
	`near-gene-3`
	`near-gene-5`
	`splice-3`
	`splice-5`
	`stop-gained`
	`stop-gained-near-splice`
	`stop-lost`
	`stop-lost-near-splice`
	`utr-3`
	`utr-5`

gene_accession (evs_GeneAccession)	NCBI mRNA transcripts accession number. Cardinality: 0..1 MySQL Type: TEXT Examples: `NM_177987.2`
grantham_score (evs_GranthamScore)	Grantham Scores categorize codon replacements into classes of increasing chemical dissimilarity based on the publication by Granthan R.in 1974, Amino acid difference formula to help explain protein evolution. Science 1974 185:862-864. Cardinality: 0..1 MySQL Type: INT Examples: `22`
	HGNC gene symbol for the gene. If the track describes features that are not

hgnc	directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT
maf_in_percent_aa (evs_maf_in_percent_aa)	The minor-allele frequency in percent for African American populations. Cardinality: 0..1 MySQL Type: INT Examples: `0.0`
maf_in_percent_all (evs_maf_in_percent_all)	The minor-allele frequency in percent for all populations. Cardinality: 0..1 MySQL Type: INT Examples: `0.0`
maf_in_percent_ea (evs_maf_in_percent_ea)	The minor-allele frequency in percent for European American populations. Cardinality: 0..1 MySQL Type: INT Examples: `0.0`

on_illumina_human_exome_chip
(evs_OnIlluminaHumanExomeChip)
	Whether a SNP is on the Illumina HumanExome Chip.
	Cardinality: 0..1
	MySQL Type: VARCHAR (3)
	Accepted Values:
	`yes`
	`no`

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
polyphen (evs_Polyphen)	Prediction of possible impact of an amino acid substitution on protein structure and function based on PolyPhen program. Cardinality: 0..1 MySQL Type: VARCHAR (17)

	Accepted Values:
	`unknown`
	`probably-damaging`
	`benign`
	`possibly-damaging`

protein_pos (evs_ProteinPos)	The corresponding amino acid position in a protein Cardinality: 0..1 MySQL Type: INT Examples: `315`
ref_base_ncbi37 (evs_RefBaseNCBI37)	Cardinality: 0..1 MySQL Type: TEXT Examples: `rs141675066:TUBB8`
rsid	The dbSNP id (rsid) for the variation. Cardinality: 0..1 MySQL Type: TEXT Examples: `rs2170774`

rsid_mapping
(evs_rsid_mapping)
	Few genomic locations are not accurately mapped to the the rsID by SeattleSeqAnnotation137. This key indicates if the mapping to the rsid is accurate or not. If marked approximate, it should be considered as a suggestion rather than an accurate mapping to the existing records in dbSNP.
	Cardinality: 1
	MySQL Type: VARCHAR (11)
	Accepted Values:
	`accurate`
	`approximate`

uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

References:

1. Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS) [December 15, 2011]

dbNSFP Nonsynonymous functional predictions

Version: version:v2.4

Track Description: This track contains data from dbNSFP(Database for Non-

synonymous SNPs Functional Predictions) [1]. dbNSFP is an integrated database of functional predictions from multiple algorithms for the comprehensive collection of human non-synonymous SNPs (NSs). It compiles prediction scores from four new and popular algorithms (SIFT, Polyphen2, LRT, and MutationTaster), along with a conservation score (PhyloP) and other related information, for every potential NS SNP in the human genome. More details about the methods of prediction is available at http://www.ncbi.nlm.nih.gov/pubmed/21520341

Benefit: This track also provides a calculated consensus prediction based on the results from different prediction algorithms from dbNSFP data. The prediction of each NS is accreted according to its deleterious tendency ("Likely Pathogenic", "Uncertain Significance", "Likely Not Pathogenic","Not Pathogenic").

Track Name: dbnsfp

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
aa_altref (dbNSFP_aa_altref)	Alternative amino acid. Cardinality: 0..1 MySQL Type: TEXT Examples: `L`
aa_pos (dbNSFP_aa_pos)	Amino acid position as to the protein. Cardinality: 0..* MySQL Type: BIGINT Examples: `1`
aa_ref (dbNSFP_aa_ref)	Reference amino acid. Cardinality: 0..1 MySQL Type: TEXT Examples: `M`
accession	Gene ID Cardinality: 1 MySQL Type: TEXT Examples: `85440`
altref (dbNSFP_altref)	Alternative nucleotide allele (as on the + strand) Cardinality: 0..1 MySQL Type: TEXT Examples: `C`
ancestral_allele (dbNSFP_Ancestral_allele)	Ancestral allele (based on 1000 genomes reference data) Cardinality: 0..1 MySQL Type: TEXT Examples:

	`A`
brief (feature)	Aminoacid reference base > Aminoacid alternate reference base: Consensus prediction. Cardinality: 0..1 MySQL Type: TEXT Examples: `M>L:Probably Deleterious 50%`
cadd_score (dbNSFP_cadd_score)	CADD raw score for funtional prediction of a SNP. The larger the score the more likely the SNP has damaging effect Cardinality: 0..1 MySQL Type: INT Examples: `0.997162`
ccds_id (dbNSFP_CCDSid)	CCDS ID. Cardinality: 0..1 MySQL Type: TEXT Examples: `CCDS2.2`
consensus_prediction (dbNSFP_Consensus_Prediction)	The prediction of each non-synonymous SNP is accreted according to its deleterious tendency ("Likely Pathogenic", "Uncertain Significance","Likely Not Pathogenic", "Not Pathogenic"). Cardinality: 0..1 MySQL Type: TEXT Examples: `Probably deleterious 25%` 2 out of 4 tools reports the prediction to be possibly deleterious.
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT

fathmm_pred
(dbNSFP_FATHMM_pred)
	FATHMM prediction.
	Cardinality: 0..*
	MySQL Type: VARCHAR (1)
	Primary Field: fathmm_score
	Accepted Values:
	`D`
	Deleterious
	`T`
	Tolerated

fathmm_score (dbNSFP_FATHMM_score)	FATHMM score Cardinality: 0..* MySQL Type: INT Examples: `0.123`
gerp_nr (dbNSFP_GERP_NR)	GERP++ neutral rate. Cardinality: 0..1 MySQL Type: INT Examples: `4.17`
gerp_rs (dbNSFP_GERP_RS)	GERP++ RS score. Cardinality: 0..1 MySQL Type: INT Examples: `4.17`
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT

lr_pred
(dbNSFP_LR_pred)
	Prediction of LR based ensemble prediction score, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5
	Cardinality: 0..1
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	Damaging
	`T`

	Tolerated
lr_score (dbNSFP_LR_score)	Logistic regression (LR) based ensemble prediction score, which incorporated 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from 0 to 1. Cardinality: 0..1 MySQL Type: INT Examples: `0.997162`

lrt_pred
(dbNSFP_LRT_pred)
	LRT prediction.
	Cardinality: 0..1
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	Deleterious
	`N`
	Neutral
	`U`
	Unknown

lrt_score (dbNSFP_LRT_score)	LRT two-sided p-value (LRTori), ranges from 0 to 1. Cardinality: 0..1 MySQL Type: INT Examples: `0.999996`

mutation_assessor_pred
(dbNSFP_MutationAssessor_pred)
	MutationAssessor prediction.
	Cardinality: 0..1
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	Deleterious
	`T`
	Tolerated

mutation_assessor_score (dbNSFP_MutationAssessor_score)	MutationAssessor functional impact combined score Cardinality: 0..1 MySQL Type: INT Primary Field: mutation_assessor_pred Examples: `0.123`

mutation_taster_pred
(dbNSFP_MutTaster_pred)
	MutationTaster prediction.
	Cardinality: 0..1
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`A`
	disease_causing_automatic
	`D`
	disease_causing
	`N`
	polymorphism
	`P`
	polymorphism automatic

mutation_taster_score (dbNSFP_MutTaster_score)	MutationTaster score. Cardinality: 0..1 MySQL Type: INT Primary Field: mutation_taster_pred Examples: `0.781789`
phastcons100way_vertebrate (dbNSFP_phastCons100way_vertebrate)	phastCons conservation score based on the multiple alignments of 100 primate genomes (including human). The larger the score, the more conserved the site. Cardinality: 0..1 MySQL Type: INT Examples: `0.9`
phastcons46way_placental (dbNSFP_phastCons46way_placental)	phastCons conservation score based on the multiple alignments of 33 placental mammal genomes (including human). The larger the score, the more conserved the site. Cardinality: 0..1 MySQL Type: INT Examples: `0.857`
phastcons46way_primate (dbNSFP_phastCons46way_primate)	phastCons conservation score based on the multiple alignments of 10 primate genomes (including human). The larger the score, the more conserved the site. Cardinality: 0..1 MySQL Type: INT Examples: `0.567`
	phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 100 primate

phylop100way_vertebrate (dbNSFP_phyloP100way_vertebrate)	genomes (including human). The larger the score, the more conserved the site. Cardinality: 0..1 MySQL Type: INT Examples: `0.9`
phylop46way_placental (dbNSFP_phyloP46way_placental)	phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 33 placental mammal genomes (including human). The larger the score, the more conserved the site. Cardinality: 0..1 MySQL Type: INT Examples: `0.857`
phylop46way_primate (dbNSFP_phyloP46way_primate)	phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 10 primate genomes (including human). The larger the score, the more conserved the site. Cardinality: 0..1 MySQL Type: INT Examples: `0.567`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT

polyphen2_pred (dbNSFP_Polyphen2_pred)	Polyphen2 prediction.
	Cardinality: 0..*
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	probably damaging
	`P`
	possibly damaging
	`B`
	benign

polyphen2_score (dbNSFP_Polyphen2_score)	Polyphen2 score, i.e. pph2_prob. Cardinality: 0..* MySQL Type: INT Primary Field: polyphen2_pred Examples: `0.997`

radial_svm_pred
(dbNSFP_radialSVM_pred)
	Prediction of SVM based ensemble prediction score, "T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.
	Cardinality: 0..1
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	Damaging
	`T`
	Tolerated

radial_svm_score (dbNSFP_radialSVM_score)	Support vector machine (SVM) based ensemble prediction score, which incorporates 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from -2 to 3 in dbNSFP. Cardinality: 0..1 MySQL Type: INT Examples: `0.997162`
ref (dbNSFP_ref)	Reference nucleotide allele (as on the + strand). Cardinality: 0..1 MySQL Type: TEXT Examples: `A`
reliability_index (dbNSFP_reliability_index)	Number of observed component scores (except the maximum frequency in the 1000 genomes populations) for RadialSVM and LR. Ranges from 1 to 10. As RadialSVM and LR scores are calculated based on imputed data, the less missing component scores, the higher the reliability of the scores and predictions. Cardinality: 0..1 MySQL Type: INT Examples: `0.997162`
rsid	dbSNP id if known Cardinality: 0..* MySQL Type: TEXT Examples: `rs1234`

sift_pred
(dbNSFP_SIFT_pred)
	SIFT prediction.
	Cardinality: 0..*
	MySQL Type: VARCHAR (1)
	Primary Field: sift_score
	Accepted Values:
	`D`
	Damaging
	`T`
	Tolerated

sift_score (dbNSFP_SIFT_score)	SIFT score. Scores range from 0 to 1. The smaller the score the more likely the SNP has damaging effect. Cardinality: 0..* MySQL Type: INT Examples: `1.0`
siphy_29way_log_odds (dbNSFP_SiPhy_29way_logOdds)	SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site. Cardinality: 0..1 MySQL Type: INT Examples: `1.23`
siphy_29way_pi (dbNSFP_SiPhy_29way_pi)	The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes. Cardinality: 0..1 MySQL Type: TEXT Examples: `1.23`
slr_test_statistic (dbNSFP_SLR_test_statistic)	Sitewise likelihood statistic for testing natural selection on codons from the Genome-wide survey of sitewise selective pressures in mammals A negative value indicates negative selection, and a positive value indicates positive selection. Larger magnitude of the value suggests stronger evidence. Cardinality: 0..1 MySQL Type: INT Examples: `1.45`
transcript (dbNSFP_transcript)	ENSEMBL transcript id Cardinality: 0..* MySQL Type: TEXT Examples:

	ENST00000400754
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT
variant (dbNSFP_Variant)	Reference amino acid > Alternative amino acid Cardinality: 0..1 MySQL Type: TEXT Examples: M>V

References:

1. Liu X, Jian X, and Boerwinkle E. 2011. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Human Mutation. 32:894-899.

dbSNP

Version: dbSNP138

Track Description: Data on polymorhpisms from the NCBI dbSNP database. The track has further been enriched with a HGMD® flag, indicating if a SNP is a known disease causing mutation or disease associated polymorphism in HGMD®.

Benefit: This track will allow one to assess if a variant in your file is already known in dbSNP and its allele frequency. This is often useful to assess if the variant is potentially pathogenic, as variants that are common in the population are less likely to be so. The flagging of records was part of HGMD® allows you to really focus only on the putatively harmless variants, by ignoring the flagged ones.

Track Name: fulldbsnp

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	dbSNP rsid. Cardinality: 1 MySQL Type: TEXT
allele_freq_count (DbSNP_alleleFreqCount)	Number of observed alleles with frequency data Cardinality: 0..1 MySQL Type: BIGINT Examples: 1 4
allele_freqs (DbSNP_alleleFreqs)	Allele frequencies Cardinality: 0..* MySQL Type: TEXT Examples: A=0.439781

	C=0.003200,T=0.996800
allele_ns (DbSNP_alleleNs)	Count of chromosomes (2N) on which each allele was observed. Note: this is extrapolated by dbSNP from submitted frequencies and total sample 2N, and is not always an integer. Cardinality: 0..* MySQL Type: TEXT Examples: A=2174.000000 C=2084.000000,T=100.000000 A=799.000000,T=1389.000000
alleles (DbSNP_alleles)	Observed alleles for which frequency data are available. Cardinality: 0..* MySQL Type: TEXT Examples: A,G CA,TG -,AAAC
av_het (DbSNP_avHet)	Average heterozygosity from all observations. Note: may be computed on small number of samples. Cardinality: 0..1 MySQL Type: INT Examples: 0.049878 0.5 0.009099
av_het_se (DbSNP_avHetSE)	Standard Error for the average heterozygosity. Cardinality: 0..1 MySQL Type: INT Examples: 0.066833 0.222222 0.000076
bin (DbSNP_bin)	Indexing field to speed chromosome range queries. Cardinality: 0..1 MySQL Type: INT Examples: 586
bitfields (DbSNP_bitfields)	SNP attributes extracted from dbSNP's SNP_bitfield table. Cardinality: 0..* MySQL Type: VARCHAR (33) Accepted Values: clinically-assoc maf-5-some-pop maf-5-all-pops has-omim-omia microattr-tpa submitted-by-lsdb

	`genotype-conflict`
	`rs-cluster-nonoverlapping-alleles`
	`observed-mismatch`

brief (feature)	The rsid accession and the variant. Cardinality: 0..1 MySQL Type: TEXT Examples: `rs201565223;C/G`

class
(DbSNP_class)
	The type of variant.
	Cardinality: 1
	MySQL Type: VARCHAR (14)
	Accepted Values:
	`single`
	`in-del`
	`het`
	`microsatellite`
	`named`
	`mixed`
	`mnp`
	`insertion`
	`deletion`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT

exceptions
(DbSNP_exceptions)
	Unusual conditions noted by UCSC that may indicate a problem with the data.
	Cardinality: 0..*
	MySQL Type: VARCHAR (26)
	Accepted Values:
	`RefAlleleMismatch`
	`RefAlleleRevComp`
	`DuplicateObserved`
	`MixedObserved`
	`FlankMismatchGenomeLonger`
	`FlankMismatchGenomeEqual`
	`FlankMismatchGenomeShorter`
	`NamedDeletionZeroSpan`
	`NamedInsertionNonzeroSpan`
	`SingleClassLongerSpan`
	`SingleClassZeroSpan`
	`SingleClassTriAllelic`
	`SingleClassQuadAllelic`
	`ObservedWrongFormat`

	`ObservedTooLong`
	`ObservedContainsIupac`
	`ObservedMismatch`
	`MultipleAlignments`
	`NonIntegerChromCount`
	`AlleleFreqSumNot1`
	`SingleAlleleFreq`
	`InconsistentAlleles`

func
(DbSNP_func)
	Functional category of the SNP.
	Cardinality: 1..*
	MySQL Type: VARCHAR (14)
	Accepted Values:
	`unknown`
	`coding-synon`
	`intron`
	`near-gene-3`
	`near-gene-5`
	`ncRNA`
	`nonsense`
	`missense`
	`stop-loss`
	`frameshift`
	`cds-indel`
	`untranslated-3`
	`untranslated-5`
	`splice-3`
	`splice-5`

hgmd	HGMD accessions for HGMD entries falling on the same (1 nuc.) position as the SNP. This field is useful as an indicator of the association of diseases (as reported by HGMD) with the SNP. Cardinality: 0..* MySQL Type: TEXT Examples: `CM960042`
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	dbSNP record. Cardinality: 0..1 MySQL Type: TEXT

loc_type
(DbSNP_locType)
	Type of mapping inferred from size on reference; may not agree with variant type.
	Cardinality: 1
	MySQL Type: VARCHAR (17)
	Accepted Values:
	`range`
	`exact`

	`between`
	`rangeInsertion`
	`rangeSubstitution`
	`rangeDeletion`
	`fuzzy`

maf (DbSNP_MAF)	dbSNP reports the minor allele frequency for each rsid included in a default global population. Since this is being provided to distinguish common polymorphism from rare variants, the MAF the second most frequent allele value. In other words, if there are 3 alleles, with frequencies of 0.50, 0.49, and 0.01, the MAF will be reported as 0.49. This is calculated as the ratio of the number of times the allele is found on a chromosome and the total sample size. Cardinality: 0..1 MySQL Type: INT Examples: `0.439781`

mol_type
(DbSNP_molType)
	Sample type from exemplar submitted SNPs (ss)
	Cardinality: 1
	MySQL Type: VARCHAR (7)
	Accepted Values:
	`unknown`
	`genomic`
	`cDNA`

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT

ref_ncbi
(DbSNP_refNCBI)
	Reference genomic sequence from dbSNP.
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`C`
	`-`
	`AACCCCTAACCCTAACCCTAAACCCTA`

ref_ucsc
(DbSNP_refUCSC)
	Reference genomic sequence from UCSC lookup.
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`C`
	`-`
	`AACCCCTAACCCTAACCCTAAACCCTA`

score (DbSNP_score)	Not used. Cardinality: 0..1 MySQL Type: INT
submitter_count	Number of distinct submitter handles for submitted SNPs for this ref SNP. Cardinality: 1

(DbSNP_submitterCount)	MySQL Type: INT Examples: `2`

submitters
(DbSNP_submitters)
	List of submitter handles.
	Cardinality: 1..*
	MySQL Type: TEXT
	Examples:
	`1000GENOMES`
	`1000GENOMES,ABI,BCMHGSC_JDW,BCM_SSAHASNP,BUSHMAN,ENSEMBL,GMI,HGSV,HUMANGENOME_JCVI,PJP,SC_SNP,WI_SSAHASNP`

uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

valid
(DbSNP_valid)
	Validation status of the SNP.
	Cardinality: 1..*
	MySQL Type: VARCHAR (15)
	Accepted Values:
	`unknown`
	`by-cluster`
	`by-frequency`
	`by-submitter`
	`by-2hit-2allele`
	`by-hapmap`
	`by-1000genomes`

variant
(variation)
	The observed alleles, separated by /. Deletions are represented as a hypehn character (-).
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`C/G`
	`-/A/C/T`

weight
(DbSNP_weight)
	The quality of the alignment.
	Cardinality: 1
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`1`
	unique mapping
	`2`
	non-unique
	`3`
	many matches

Allele Frequencies (Beta)

Version: dbSNP137

Track Description: Data on population specific polymorhpisms from the 1000 Genomes.

Benefit: This track will allow one to assess if a variant in the file is already known in dbSNP and its allele frequency in a particular population. It is seen that the prevalance of SNPs differs between populations, thus knowing the prevalance of an SNP in the population is an important piece of information while dternining treatment strategies.

Track Name: ethnicsnp

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	dbSNP rsid. Cardinality: 1 MySQL Type: TEXT

af_ceu_pilot (G1000_af_ceu_pilot)	Allele frequencies for the European CEU population taken from the pilot dataset.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`A=0.439781`
	`C=0.003200,T=0.996800`

af_jptchb_pilot (G1000_af_jptchb_pilot)	Allele frequencies for the Asian JPT/CHB population taken from the pilot dataset.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`A=0.439781`
	`C=0.003200,T=0.996800`

af_yri_pilot (G1000_af_yri_pilot)	Allele frequencies for the African YRI population taken from the pilot dataset.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`A=0.439781`
	`C=0.003200,T=0.996800`

alleles (G1000_alleles)	Observed alleles for which frequency data are available.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`A,G`
	`CA,TG`
	`-,AAAC`

brief (feature)	The rsid accession and the variant. Cardinality: 0..1 MySQL Type: TEXT Examples:

	`rs201565223:C/G`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT

evidence (G1000_evidence)	The data source for this variation.
	Cardinality: 0..*
	MySQL Type: VARCHAR (21)
	Accepted Values:
	`Multiple_observations`
	`Frequency`
	`HapMap`
	`1000Genomes`
	`Cited`

hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	dbSNP record. Cardinality: 0..1 MySQL Type: TEXT
maf (G1000_MAF)	dbSNP reports the minor allele frequency for each rsid included in a default global population. Since this is being provided to distinguish common polymorphism from rare variants, the MAF the second most frequent allele value. In other words, if there are 3 alleles, with frequencies of 0.50, 0.49, and 0.01, the MAF will be reported as 0.49. This is calculated as the ratio of the number of times the allele is found on a chromosome and the total sample size. Cardinality: 0..1 MySQL Type: INT Examples: `0.439781`
	The minor allele Cardinality: 0..1

minor_allele
(G1000_minor_allele)
	MySQL Type: TEXT
	Examples:
	`A`
	`G`

minor_allele_count
(G1000_minor_allele_count)
	Number of minor alleles with frequency data
	Cardinality: 0..1
	MySQL Type: BIGINT
	Examples:
	`1`
	`4`

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

validation_status
(G1000_validation_status)
	Validation status of the SNP.
	Cardinality: 0..*
	MySQL Type: VARCHAR (10)
	Accepted Values:
	`cluster`
	`frequency`
	`submitter`
	`doublehit`
	`hapmap`
	`1000Genome`
	`freq`
	`precious`

variant
(variation)
	The observed alleles, separated by /. Deletions are represented as a hypehn character (-).
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`C/G`
	`-/A/C/T`

Regulatory Features

"Regulatory Features" correspond to transcription factor binding sites, predicted binding sites within ChIP-chip, ChIP-seq and DNAse sensitivity fragments, CpG islands, microsatellite repeats, virtual transcription start sites (TSSs) from TRANSFAC®, and post-translational modifications from PROTEOME™.

TRANSFAC® experimentally verified TFBS

Version: TRANSFAC® 2014.3

Track Description: This track contains literature-curated transcription factor binding sites (TFBS) and miRNA target sites from the TRANSFAC® database. These are experimentally demonstrated sites. Sites are labeled with a unique TRANSFAC® site accession and are linked to the corresponding Site Report.

If you also have subscription to TRANSFAC®: in some cases binding sites may be reported in TRANSFAC® that do not have corresponding entries in Genome Trax. This can happen, if it is not possible to unambiguously resolve the genomic coordinates of the site. For example it can happen when the location in the original literature is relative to the translation start without indicating the underlying reference sequence build.

Benefit: Manually curated, experimentally determined sites from the scientific literature might lead to deleterious effects in gene regulation when disrupted by mutations.

Track Name: transfac_sites

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	A site accession from TRANSFAC®. Cardinality: 1 MySQL Type: TEXT Examples: `R14194`

binding_factor (bindingFactor)	A descriptive name of the binding factor and its TRANSFAC® accession number, enclosed in angle brackets. Some sites bind multiple factors, so a list of factors is possible. If the species of the factor used was not human, it is appended in UPPERCASE. There are cases where several factors have the same name, for example if they represent different versions of a complex that uses this general name, or if they represent the factor in a species specific or general manner. These detail informaton can be obtained from TRANSFAC® using the accession number. There are also rare cases where no binding factor has been unambiguouly identified in the paper, but binding of a factor in genral has been demonstrated. In these cases the field is empty.
	Cardinality: 1..*
	MySQL Type: TEXT
	Examples:
	`CREB1 <T08562>`
	A human transcription factor
	`GKLF-isoform1 <T14242>`
	A specific Isoform of the factor GKLF
	`HSF1_MOUSE <T00384>`
	A HSF1 facor from mouse
	`Sp1 <T00759>`
	A named complex
	`p50:NF-AT2 <T17767>`
	A complex made up from two individual factors
	`hsa-miR-7-5p <T09931>`
	A miRNA
	`Egr-1_MAMM <T10531>`
	A Egr-1 factor from a not further specified mammalian species

	The accession number of the binding site in TRANSFAC®, and,

brief (feature)	if available, descriptive name(s) and accession numbers of the binding factor(s).
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`R37260`
	A binding site without factors.
	`R56373:hsa-miR-365a-3p <T26559>`
	A miRNA target site, and the targeting miRNA.
	`R56877:NF-kappaB <T00590>,RelA-p65 <T08711>,c-Rel <T09254>`
	A transcription factor binding site, shown to bind three factors

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	A site report specifically created for Genome Trax™. (Full reports can be reached from this page. To get access to full TRANSFAC professional functionality and content, a separate license is required.) Cardinality: 0..1 MySQL Type: TEXT
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
site_acc (siteAcc)	Cardinality: 1 MySQL Type: TEXT Examples: `R18913`
	Cardinality: 1 MySQL Type: TEXT Examples:

site_type	R miRNA target site D DNA transcription factor binding site
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Predicted ChIP-Seq TFBS

Version: TRANSFAC® 2014.2

Track Description: This track contains experimental binding sites which are refined by prediction.

A transcription factor is identified as binding the ChIP-chip or ChIP-seq fragment experimentally. TRANSFAC® positional weight matrices (PWMs) for this factor are then used for the analysis, and the best scoring sites are calculated with the Match algorithm executed with option to return one best hit in the whole sequence. The most conserved relevant matrix was used for the calculations.

Tech Note: ChIP-Seq fragments are typically hundreds of nucleotides long, and it is known which factor binds them, but not exactly where in the sequence the factor binds. We use our knowledge about the structure of the binding sites to identify the actual binding site within the ChIP-seq sequence. The site structure comes from our manually curated binding sites for this factor, and is captured in the form of Positional Weight Matrices. In some cases, the consensus sequence can also be inferred by finding a conserved sequence motive in a large number of ChIP-seq sequences for the same factor. By limiting the site prediction to a predefined transcription factor and a short ChIP-seq fragment, there is low risk of identifying false-positive binding sites in this process.

Benefit: This track is a value add over publicly available ChIP-Seq fragments. It takes advantage of our manually curated binding sites and their derived PWMs.

Track Name: chip

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	Fragment Accession Number from the BIOBASE Knowledge Library. Cardinality: 1 MySQL Type: TEXT Examples: FR000017430
binding_factor (bindingFactor)	A descriptive name of the binding factor. Cardinality: 1 MySQL Type: TEXT Examples: c-Myc A human transcription factor
	A descriptive name of the binding factor

brief (feature)	Cardinality: 0..1 MySQL Type: TEXT Examples: `Sp1` A human transcription factor
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene that might be regulated by the predicted chip site. This gene is defined as the one having the closest distance to the site containing fragment. This could be an overlapping gene or an upstream or an downstream gene relatvie to the location of the fragment as a result of the unknown fragment strand. Cardinality: 0..* MySQL Type: TEXT Examples: `ENSG00000160185`
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
fragment_end	genomic end coordinate of the fragment Cardinality: 0..1 MySQL Type: BIGINT Examples: `7592725`
fragment_start	genomic start coordinate of the fragment Cardinality: 0..1 MySQL Type: BIGINT Examples: `7589439`
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	A fragment report specifically created for Genome Trax™. (Full reports can be reached from this page. To get access to full TRANSFAC® professional functionality and content, a separate license is required.) Cardinality: 0..1 MySQL Type: TEXT
matrix_id	Matrix ID from TRANSFAC®. Cardinality: 0..1 MySQL Type: TEXT Examples: `V$SP1_Q6`

matrix_score (matrixScore)	Matrix similarity score (calculated by Match algorithm) of the highest scoring match in the fragment sequence. This score is normalized between 0 and 1, with 1 being the highest. Cardinality: 0..1 MySQL Type: INT Examples: `0.902`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Predicted TFBSs in DNAse hypersensitivity regions

Version: TRANSFAC® 2014.3

Track Description: This track contains predicted binding sites at the DNAse hypersensitive regions.

Hypersensitivity to DNAse correlates with regulatory elements in the neighborhood of genes. We collected DNAse fragments from 142 ENCODE[1] data sets, and predict potential transcription factor binding sites on the DNAse fragment sequences by running MATCH[2] with stringent criteria that allow only one site per fragment per matrix.

Tech Note: DNAse sensitive fragments are typically hundreds of nucleotides long, and it is not known which factor(s) binds them, or where. We use knowledge about the structure of the binding sites to identify actual binding site within the DNAse sensitivity fragment. The site structure comes from our manually curated binding sites, and is captured in the form of Positional Weight Matrices. MATCH was run using a non-redundant set of 148 high quality matrices from vertebrates using the "unique" option and the "non-redundant_minFP" matrix-specific cut-off which minimizes false-positive matches, to generate only one high scoring site if such a site was found, for each sequence and matrix.

Benefit: This track is a value add over publicly available DNAse hypersensitive fragments. It takes advantage of our manually curated binding sites and their derived PWMs.

Track Name: dnase

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	Matrix accession number from TRANSFAC®. Cardinality: 1 MySQL Type: TEXT Examples: `M00129`
	The matrix identifier, matrix similarity score from MATCH® and matched sequence in uppercase with

brief (feature)	short flanking sequences in lowercase, separated by colons. The matrix similarity is a score that describes the quality of a match between a matrix and the input sequences. This Score is normalized between 0 and 1, 1 indicating the best possible score. Cardinality: 0..1 MySQL Type: TEXT Examples: `V%24HFH1_01:0.969:acaTAAACattg`
core_similarity_score (dnase_css)	The core similarity score from MATCH®. The core similarity denotes the quality of a match between the core sequence of a matrix (the five most conserved positions within a matrix) and an arbitrary part of the input sequence. Score is normalized between 0 and 1, 1 indicating the best score. Cardinality: 1 MySQL Type: INT Examples: `0.857`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT

fragment_acc
(fragmentAcc)
	Fragment accession number from TRANSFAC®.
	Cardinality: 1..*
	MySQL Type: TEXT
	Examples:
	`FR022189646`
	`FR006042912,FR005215467,FR005070392`

hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	A fragment report specifically created for Genome Trax™. (Full reports can be reached from this page. To get access to full TRANSFAC® professional functionality and content, a separate license is required.) Cardinality: 0..1 MySQL Type: TEXT

matrix_acc (dnase_matrix_acc)	An accession number for the matrix that yields the predicted binding site, from TRANSFAC®. Cardinality: 1 MySQL Type: TEXT Examples: `M00123`
matrix_id (dnase_matrix_id)	An encoded id for the matrix that yields the predicted binding site. It is encoded for the taxonomy, potential binding factor and the evidence of the experimental support for the binding matrix. E.g. V$AP2_Q6_01. V stands for vertebrate, the term after the $ is the factor name, this also can be the trivial name for a complex of several genes. Evidences are the values behind Q, and range from 1 to 5, with one being the highest. As a matrix is build from multiple binding sites, the evidence score is conservatively taken to be the lowest for any of the individual site binding evidences used. Q6 means not classified. Matrices that were directly taken from literature do not have a Q value. The final two digits are a running number in case there are several different matrices for the same factor. Cardinality: 1 MySQL Type: TEXT Examples: `V%24HFH1_01`
matrix_similarity_score (dnase_mss)	The similarity score of the binding site defined by the factor's matrix, and the sequence, ranging from 0 (no similarity) to 1 (perfect match). Cardinality: 1 MySQL Type: INT Examples: `0.848`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT

sequence_site
(dnase_sequence_site)
	Identifies the matching sequence. Capital letters indicate the positions in the sequence that match with the core sequence of the matrix, while the lower case letters refer to positions which match to other parts of the matrix. Note: The matrix can also contain highly conserved positions outside the core of the matrix.
	Cardinality: 0..1
	MySQL Type: TEXT
	Examples:
	`acaTAAACattg`
	`CAGGAacttcc`

uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..*

	MySQL Type: TEXT

References:

1. The ENCODE project and UCSC.
2. Kel, A. E.; Goessling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O. V.; Wingender, E. Match(TM) : A tool for searching transcription factor binding sites in DNA sequences, Nucleic Acids Res. 31, 3576-3579 2003.

CpG Islands

Version: TRANSFAC® 2014.3

Track Description: This track contains computationally determined CpG islands from TRANSFAC®. There is no linking BED file for this track, only a descriptive BED file. This file contains CpG islands across the human genome computed using the algorithm of (Wang and Leung), and these features are strand independent.

Scientific background note: CpG islands are genomic regions that contain a high frequency of CpG sites, where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length. Cytosines in CpG dinucleotides can be methylated. CpG sites in the CpG islands of promoters are unmethylated if genes are expressed. This observation led to the speculation that methylation of CpG sites in the promoter of a gene may inhibit the expression of a gene. In mammals, methylating the cytosine within a gene can also turn the gene off.

Benefit: This is calculated data, provided as a convenience track.

Track Name: cpg_islands

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	TRANSFAC® Promoter accession number Cardinality: 1 MySQL Type: TEXT Examples: `PM000042212`
brief (feature)	CpGs=X:Percent=Y Cardinality: 0..1 MySQL Type: TEXT Examples: `CpGs=133:Percent=71`
cpg_count (cpgCount)	Number of CpGs in island. Cardinality: 1 MySQL Type: BIGINT Examples: `130`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT

entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
gc_percent (gcPercent)	Percentage number describing the GC content. Cardinality: 1 MySQL Type: INT Examples: `71`
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	TRANSFAC® Promoter Report (a TRANSFAC® license is required to access this data). Cardinality: 0..1 MySQL Type: TEXT
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Microsatellites

Version: TRANSFAC® 2014.3

Track Description: This track contains computationally determined microsatellite repeats and their patterns from TRANSFAC.

Scientific background note: Microsatellites, also known as Short Tandem Repeats (STRs), are short sequences of DNA, often of just two to four nucleotides, repeated ten to a hundred times.

Microsatellites are typically neutral and co-dominant and exhibit an increased rate of mutation. Consequently can be multiple alleles for a microsatellite locus, making them useful as molecular markers in genetics, for kinship and population studies. Length changes of microsatellites within promoters and other cis-regulatory regions can change gene expression. Microsatellites within introns also influence phenotype, through means that are not currently understood.

Benefit: This is calculated data, provided as a convenience track.

Track Name: microsatellites

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION

accession	Serial number (not necessarily stable between releases). Cardinality: 1 MySQL Type: TEXT
brief (feature)	Motif of the microsatellite Pattern as DNA nucleotide sequence. Cardinality: 0..1 MySQL Type: TEXT
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT
pattern	The entire DNA motif repeated in the microsatellite. Cardinality: 1 MySQL Type: TEXT Examples: `TGTGCATGTATGTATGTG`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Virtual Transcription Start Sites (TSSs)

Version: TRANSFAC® 2014.3

Track Description: This track contains the location of calculated transcription start

sites (TSSs).

The calculation of "virtual TSSs" as reference points is based on a collection of TSSs for a given gene. TSSs are taken from EnsEMBL.

EnsEMBL TSSs are assumed to be the first nucleotide of the most 5' exon of an EnsEMBL mRNA model. Thus, collected TSSs for a given gene are located on a sequence fragment which sometimes spans several thousand nucleotides, in some cases far more than 100 kb. They are frequently not located in tight clusters of only a few dozen nucleotides length, but are often widespread throughout the sequence.

In order to define a reasonable number of "virtual TSSs" for a given gene from this data collection, an algorithm was designed which applies a set of rules to the data collection in order to find "clusters" of TSSs. A window of 3000 nt length is slid along the entire sequence fragment. A "clustering score" is calculated by summing up weighted contributions from each TSS in the window. Each TSS derived file is scored with 5 evidence points, The weights of evidence points are additionally multiplied by a distance score: the central position is multiplied by 1, the outer positions are multiplied by 0, and all positions in between by a value taken from a cosine function, according to the distance from the center of the window. The peaks of the resulting clustering score are regarded as potential "virtual TSSs".

For some of the genes only a handful of evidence points are available, thus resulting in multiple "virtual TSSs", each consisting of only a few evidence points. Therefore, for all those genes where less than 19 evidence points are available only the most 5' "virtual TSS" is accepted. For all other genes those peaks are accepted as "virtual TSSs" for which the respective sequence window contains at least 8% of all evidence points. However, there are genes, for which - although the coverage with data is pretty good - the annotated TSSs are so equally distributed along the sequence, that no prominent peaks occur, and therefore - according to the above mentioned rules - no peak would be accepted. In this case the most prominent peaks are accepted. If there are more than two peaks for which these conditions are true, the most 5' "virtual TSS" is accepted.

Benefit: This is calculated data, provided as a convenience track. The proprietary calculation algorithm provides higher confidence TSSs.

Track Name: tss

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	Promoter accession from the BIOBASE Knowledge Library Cardinality: 1 MySQL Type: TEXT Examples: `PM000000055`
brief (feature)	HGNC symbol of gene for which this is a TSS. Cardinality: 0..1 MySQL Type: TEXT Examples: `MED14`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT

entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	BIOBASE Knowledge Library Promoter Report (a TRANSFAC® license is required to access this data). Cardinality: 0..1 MySQL Type: TEXT
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Post translational modifications

Version: PROTEOME™ 2014.3

Track Description: This track contains the location of post-translational modification sites, for example phosphorylation or ubiquitination sites. The sites have been manually curated from the scientific literature in PROTEOME™ edition, taken from a publiation by Olsen et. al, or parsed from the Uniprot knowledgebase. Coordinates encompass the three nucleotides which code for the affected amino acid, unless the nucleotides are separated by intron sequences, in which case the coordinates will be split.

Benefit: These data are manually curated, experimentally determined PTM sites from the scientific literature, which might have a functional impact in pathways.

Track Name: ptms

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	PROTEOME™ gene accession (if taken from BIOBASE PROTEOME™ or Uniprot Protein ID if taken from UniProt or Olsen et al. Cardinality: 1 MySQL Type: TEXT Examples: `015541`
	Cardinality: 1

aminoacid (ptm_aminoacid)	MySQL Type: TEXT Examples: `S`
aminoacidposition (ptm_aminoacidposition)	Cardinality: 0..1 MySQL Type: BIGINT
brief (feature)	Reaction name. Cardinality: 0..1 MySQL Type: TEXT Examples: `Y:N/A:phosphorylation`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT
modification (ptm_modification)	Cardinality: 1 MySQL Type: TEXT Examples: `phosphorylation`
molecule_acc (ptm_moleculeAcc)	Cardinality: 1 MySQL Type: TEXT
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT

source (ptm_source)	Specifies from which data source the PTM information was taken.
	Cardinality: 1
	MySQL Type: VARCHAR (18)
	Accepted Values:
	`Olsen et. al. 2010`
	`BIOBASE`
	`UniProt`

uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

miRNA

Version: miRBase 20

Track Description: This track contains microRNA sequences from miRBase[1]. Each entry in the miRBase database represents a predicted hairpin portion of a miRNA transcript, with information on the location and sequence. Data are provided courtesy of miRBase. MiRNAs are first transcribed as primary transcriptes of longer sequence length, that then are processed into shorter, mature miRNAs.

Benefit: miRNAs are post-transcriptional regulators that bind to complementary sequences on target messenger RNA transcripts (mRNAs), usually resulting in translational repression or target degradation and gene silencing.

Track Name: mirna

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	miRBase accession Cardinality: 1 MySQL Type: TEXT Examples: `MIMAT0005890`
brief (feature)	miRNA approved name Cardinality: 0..1 MySQL Type: TEXT Examples: `hsa-mir-1302-2` These identifiers are derived from the HGNC gene symbol (eg. MIR1302-2) and contain the species for disambiguation, in the case of mature miRNAs, also a tag for disambiguation of mulitple mature miRNAs that stem from the same primary transcript. Same miRNAs can originate from multiple sites in the genomes, in which case each site the gene symbols contain an appended hypen and differentiating number.
derives_from	If the miRNA is a mature miRNA that derives from a primary miRNA transcript, then this contains the accession of that primary miRNA transcript. Cardinality: 1 MySQL Type: TEXT

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	miRBase microRNA report. Cardinality: 0..1 MySQL Type: TEXT
name	The name of the mature miRNA product (also called miRNA id in miRBase) Cardinality: 1 MySQL Type: TEXT Examples: `hsa-miR-514a-5p`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

References:

1. miRBase: integrating microRNA annotation and deep-sequencing data. Kozomara A, Griffiths-Jones S. NAR 2011 39(Database Issue):D152-D157

Gene Functional Assignments

"Gene Functional Assignments" correspond to functional relationships that are mapped to the entire gene from the first nucleotide of the first exon to the last nucleotide of the last exon (inclusive of all introns). This is manually curated functional information from PROTEOME, and includes Disease Associations, Drug Targets and Pathway Membership for the genes. These assignments make it possible to filter for variation hitting genes that are known to be involved in a given disease, pathway or associated with a given compound, and can serve as candidate genes for ranking novel SNVs.

Disease associations

Version: PROTEOME™ 2014.3

Track Description: This track contains literature derived disease biomarker associations for genes. These disease associations for genes have been manually curated from the scientific literature as being, and can be based on other sources than mutations, for example on gene expression or changes in protein level. The diseases are linked to the corresponding PROTEOME™ Disease Report which lists further information on the disease and its associated genes.

Benefit: These disease association data are manually curated, experimentally determined associations from the scientific literature, mapped to coordinates. They allow you to identify novel SNPs that may be associated with a disease due to the gene on which they fall being implicated to be related to the disease, sometimes called "guilt-by-association".

Track Name: disease

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	Proteome Gene accession. Cardinality: 1 MySQL Type: TEXT Examples: `GN000000001`
brief (feature)	Human readable list of disease names associated with the gene, as in disease. Cardinality: 0..1 MySQL Type: TEXT

disease	Human readable disease name.
	Cardinality: 1..*
	MySQL Type: TEXT
	Primary Field: mesh_id
	Examples:
	`Schizophrenia`
	`Breast Neoplasms`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc

hgnc	symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	PROTEOME™ Gene Report (a PROTEOME™ license is required to access those data). Cardinality: 0..1 MySQL Type: TEXT
mesh_id (disease_mesh_id)	MeSH ID for the disease name. Cardinality: 1..* MySQL Type: TEXT Examples: `D012559`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Drug targets

Version: PROTEOME™ 2014.3

Track Description: This track contains contains DrugBank derived drug associations for proteins from PROTEOME™. The drug names are are linked to the corresponding PROTEOME™ Gene Report which lists details on each of the associated drugs.

Benefit: These drug association data are manually curated (by DrugBank), experimentally determined associations from the scientific literature, mapped to coordinates.

Track Name: drug

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	PROTEOME™ Gene Accession. Cardinality: 1 MySQL Type: TEXT Examples: `GN000000004`
brief (feature)	The drug name list, as given in drug. Cardinality: 0..1 MySQL Type: TEXT Examples:

	`B-Octylglucoside`
drug	Cardinality: 1..* MySQL Type: TEXT Primary Field: drug_acc Examples: `Bleomycin`
drug_acc	The PROTEOME™ accession numbers for the individual drugs. Cardinality: 1..* MySQL Type: TEXT Examples: `DR000000537`
drugbank_acc (drugbank)	Accession number of the drug in DrugBank Cardinality: 1..* MySQL Type: TEXT Primary Field: drug_acc Examples: `DB00290`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	PROTEOME™ Gene Report (a PROTEOME™ license is required to access those data). Cardinality: 0..1 MySQL Type: TEXT
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
	FDA approval status of the compound Cardinality: 1..*

status	MySQL Type: TEXT Primary Field: drug_acc Examples: `small molecule,approved`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Pathway membership

Version: PROTEOME™ 2014.3

Track Description: This track contains literature-derived pathway membership information for proteins from PROTEOME™. Pathway names are linked to the corresponding PROTEOME™ Pathway Report which lists each of the associated reactions and genes.

Benefit: These pathway association data are manually curated, experimentally determined associations from the scientific literature, mapped to coordinates.

Track Name: pathway

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	PROTEOME™ Gene Accession. Cardinality: 1 MySQL Type: TEXT Examples: `GN000000004`
brief (feature)	The pathway name list, as given in pathway. Cardinality: 0..1 MySQL Type: TEXT Examples: `leptin signaling`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature.

hgnc	Cardinality: 0..* MySQL Type: TEXT
hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT

pathway	The pathway name. This may be a descriptive name for "canonical" pathways, such as 'biosynthesis of hemoglobin and cytochromes', or it will be composed of some prominent molecule or process names along the pathway, such as 'PI3K ---AKT-1---/ FOXO4'. In these, ---/ indicates inhibition, ---> indicates activation, and the gene within the arrow, is an intermediary for mediating this effect.
	Cardinality: 1..*
	MySQL Type: TEXT
	Primary Field: pathway_acc
	Examples:
	`leptin signaling`
	`leptin ---PI3K%2C AKT-1---> AMP`

pathway_acc	PROTEOME™ accession number for the pathway. Cardinality: 1..* MySQL Type: TEXT Examples: `CH000004582`
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

HGMD® disease genes

Version: HGMD® professional 2014.3

Track Description: This track contains literature derived disease associations for genes from HGMD®. It includes mutations in HGMD® that have been reported for the gene in question, but for which no exact genomic coordinates can be provided because the original literature did not report the exact location of the mutation. It also associates the entire sequence of a gene with the diseases for which individual mutations with exact coordinates have been reported. The relationship of the gene to the disease is due to the overall function of the gene, and novel mutations to a known disease gene that also disrupt the gene may have a higher propensity to cause the same disease.

Benefit: These disease association data are manually curated, experimentally determined associations from the scientific literature, mapped to coordinates. They allow you to identify novel SNPs that may be associated with a disease due to the

gene on which they fall being implicated to be related to the disease, sometimes called "guilt-by-association".

Track Name: hgmd_disease_genes

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	HGMD® gene accession (identical to HGNC gene symbol) Cardinality: 1 MySQL Type: TEXT Examples: `LIPI`
brief (feature)	The gene and disease(s). Cardinality: 0..1 MySQL Type: TEXT Examples: `BRCA1:Breast Cancer`

confidence	Cardinality: 1..*
	MySQL Type: VARCHAR (4)
	Accepted Values:
	`High`
	`Low`
	The curators had some reservation about the strength of the evidence for the mutation/disease relationship.

disease	The associated disease or phenotype.
	Cardinality: 1
	MySQL Type: TEXT
	Examples:
	`Haemophilia B`
	`Thalassaemia beta`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT

hyperlink	HGMD® professional gene report. To get access to this report and full HGMD professional functionality and content, a separate license to HGMD is required. Cardinality: 0..1 MySQL Type: TEXT

mutation_type (mutationType)	A one-letter code determining which class (and table in HGMD®) the mutation belongs to. If there are several mutations of different type they are given as a comma-delimted list, such as *D,I,M* for a gene with at least one deletion, insertion and point mutation.
	Cardinality: 1..*
	MySQL Type: VARCHAR (1)
	Accepted Values:
	`D`
	Deletion.
	`E`
	Amplet.
	`G`
	Gross deletion - refers to lesions covering more than 20 nucleotides.
	`I`
	Insertion.
	`M`
	Mutation (mis-sense or non-sense single nucleotide).
	`N`
	Gross Insertion/Deletion.
	`P`
	Complex Rearrangement.
	`R`
	Promoter mutation.
	`S`
	Splice site mutation.
	`X`
	Indel.

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
supporting_variants	The number of variants in HGMD® associating the gene to the disease Cardinality: 1..* MySQL Type: BIGINT Examples: `6`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..*

	MySQL Type: TEXT

variant_type (variantType)	The severity category of the variant.
	Cardinality: 1..*
	MySQL Type: VARCHAR (3)
	Accepted Values:
	`DP`
	Disease-associated polymorphism - A polymorphism reported to be in significant association with a disease/phenotype (p<0.05) that is assumed to be functional (e.g. as a consequence of location, evolutionary conservation, replication studies etc), although there may as yet be no direct evidence (e.g. from an expression study) of a functional effect.
	`DFP`
	Disease-associated polymorphism with additional supporting functional evidence - A polymorphism reported to be in significant association with disease (p<0.05) that has evidence of being of direct functional importance (e.g. as a consequence of altered expression, mRNA studies etc).
	`FP`
	In vitro/laboratory or in vivo functional polymorphism - A polymorphism reported to affect the structure, function or expression of the gene (or gene product), but with no disease association reported as yet.
	`FTV`
	Frameshift or truncating variant - A polymorphic or rare variant reported in the literature (e.g. detected in the process of whole genome/exome screening) that is predicted to truncate or otherwise alter the gene product (i.e. a nonsense or frameshift variant) but with no disease association reported as yet. Please note that any variant affecting the obligate donor/acceptor splice site of a gene will not be included in this category unless there is evidence for an effect on the splicing phenotype. Variants occurring in pseudogenes will also be excluded unless evidence for a functional effect is present for both the pseudogene itself and the variant in question.
	`CNV`
	Copy number variations are DNA segments >1 kb in length that present with variable numbers of copies in a given population. These variants are being reported in the literature with an ever increasing frequency. CNVs are potentially functionally significant and should therefore in principle be treated by HGMD in a similar manner to any other polymorphism.
	`DM`
	Disease causing mutation - Pathological mutation reported to be disease causing in the corresponding report.
	`DM?`
	Disease causing mutation (report questionable) - mutation reported to be disease causing in the corresponding report, but where the author has indicated that there may be some degree of doubt, the curator had doubts about the validity of the claim given the data presented or subsequent evidence has come to light in the literature, calling the deleterious nature of the variant into question.
	`R`
	Removed - mutations that were removed from the database, for example because the report was erroneous or has been retracted. To allow users to track these changes, the records were not actually removed, but flagged as R, retaining all their other characteristics. These variants should not be used for annotation purposes.

Orphanet (Beta)

Version: Downloaded on 11th September 2014

Track Description: Orphanet is the reference portal for information on rare diseases and orphan drugs, for all audiences. Orphanet's aim is to help improve the diagnosis, care and treatment of patients with rare diseases.

Benefit: Allows you to associate known patterns of inheritance (dominant, recessive) with rare diseases and the genes implicated in them. Togehter with the observed zygosity, and the disease causing mutations in HGMD, this can help you to focus only on dominant disease causing variants, or on recessive disease causing variants that are homozygous in the patient sample.

Track Name: orpha

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	The numerical part of the 'Orpha number'. Cardinality: 1 MySQL Type: TEXT Examples: `79314` associated with the 'Orpha number' ORPHA79314
avg_age_of_death	Cardinality: 0..1 MySQL Type: TEXT
avg_age_of_onset	Cardinality: 0..1 MySQL Type: TEXT
brief (feature)	Inheritance:Disease. If nothing is known about the inheritance modes, this can be empty. Multiple values will be listed for inheritance. Cardinality: 0..1 MySQL Type: TEXT
disease	Textual name for the disease Cardinality: 1 MySQL Type: TEXT Examples: `L-2-hydroxyglutaric aciduria`
ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
	List of HGNC gene symbols, a symbol for each gene associated with a particular Record. Please note that Orphanet entries associated with more than 1 gene lead to several Records in this Track. Please also note that not all indicated genes are necessarily relevant to the genomic

hgnc	interval associated with a particular Record. Cardinality: 0..* MySQL Type: TEXT Primary Field: ensembl_id
hyperlink	A link to the associated entry in the portal for rare diseases and orphan drugs (orpha.net). Cardinality: 0..1 MySQL Type: TEXT

inheritance (orpha_inheritance)	Cardinality: 0..*
	MySQL Type: VARCHAR (25)
	Accepted Values:
	`Autosomal dominant`
	`Autosomal recessive`
	`Mitochondrial inheritance`
	`Multigenic/multifactorial`
	`Sporadic`
	`X-linked dominant`
	`X-linked recessive`
	`Unknown`
	`No data available`

omim_acc	OMIM external ID Cardinality: 0..* MySQL Type: BIGINT
pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
prevalence	Cardinality: 0..1 MySQL Type: TEXT

sign	Observable traits associated with the disease, and their frequency.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`Short Neck:Very Frequent`
	The observed frequency is separated by a colon.
	`Short Neck:Very Frequent,Splenomegaly:Occasional`
	Multiple sign/frequency pairs are allowed

uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

OMIM

Version: September 2014

Track Description: This track contains data from the OMIM [1]. OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.

Benefit: OMIM is a catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression.

Track Name: omim

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	MIM id Cardinality: 1 MySQL Type: TEXT
brief (feature)	Names of disorders that have been linked to a particular gene. If a disorder has its own record in OMIM, the OMIM number to that record is provided. Cardinality: 0..1 MySQL Type: TEXT Examples: `Chronic granulomatous disease, X-linked, 306400 (3)`
comments (omim_comments)	Additional gene information. Some comments may point out similarities or differences a gene has with other genes. Cardinality: 0..1 MySQL Type: TEXT Examples: `11kb from CLCNKB, simultaneous mutation in CLCNKA and CLCNKB`

disorders (omim_disorders)	Names of disorders that have been linked to a particular gene. If a disorder has its own record in OMIM, the OMIM number to that record is provided. Brackets, "", indicate "nondiseases" mainly genetic variations that lead to apparently abnormal laboratory test values (e.g., dysalbuminemic euthyroidal hyperthyroxinemia). Braces, "{ }", indicate mutations that contribute to susceptibility to multifactorial disorders (e.g., diabetes, asthma) or to susceptibility to infection (e.g., malaria). A question mark, "?", before the disease name indicates an unconfirmed or possibly spurious mapping.
	Cardinality: 0..*
	MySQL Type: TEXT
	Examples:
	`Chronic granulomatous disease, X-linked, 306400 (3)`
	`Hyperprolinemia, type II, 239510 (3)`

	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic

ensembl_id (ensembl)	available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT
entry_date (omim_entry_date)	Date of entry of the OMIM record in YY-MM-DD format. Cardinality: 1 MySQL Type: TEXT Examples: `12-11-21`
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT
location (omim_location)	Describes the location of the gene on the chromosome. Cardinality: 1 MySQL Type: TEXT Examples: `1p36.13`
	The methods for mapping genes. Cardinality: 1..* MySQL Type: VARCHAR (21) Accepted Values: `A` in situ DNA-RNA or DNA-DNA annealing (`hybridization') `AAS` deductions from the amino acid sequence of proteins `C` chromosome mediated gene transfer (CMGT) `Ch` chromosomal change associated with particular phenotype and not proved to represent linkage (Fc), deletion (D), or virus effect (V) `D` deletion or dosage mapping (concurrence of chromosomal deletion and phenotypic evidence of hemizygosity), trisomy mapping (presence of three alleles in the case of a highly polymorphic locus), or gene dosage effects (correlation of trisomic state of part or all of a chromosome with 50% more gene product). Includes "loss of heterozygosity" (loss of alleles) in malignancies

method (omim_method)

EM	exclusion mapping
F	linkage study in families
H	based on presumed homology
HS	DNA/cDNA molecular hybridization in solution (`Cot analysis')
L	lyonization
LD	linkage disequilibrium
M	Microcell mediated gene transfer (MMGT)
OT	ovarian teratoma (centromere mapping)
Psh	PCR of somatic cell hybrid DNA
R	irradiation of cells followed by `rescue' through fusion with nonirradiated (nonhuman) cells (Goss-Harris method of radiation-induced gene segregation)
RE	Restriction endonuclease techniques
REa	combined with somatic cell hybridization
REb	combined with chromosome sorting
REc	hybridization of cDNA to genomic fragment (by YAC, PFGE,microdissection, etc.)
REf	isolation of gene from genomic DNA; includes 'exon trapping'
REl	isolation of gene from chromosome-specific genomic library (see Pcm)
REn	neighbor analysis in restriction fragments
S	segregation (cosegregation) of human cellular traits and human chromosomes (or segments of chromosomes) in particular clones from interspecies somatic cell hybrids
T	TACT = telomere-associated chromosome fragmentation
V	induction of microscopically evident chromosomal change by a virus
X/A	X-autosome translocation in female with X-linked recessive disorder
Fb	Unknown
Fd	Unknown
Fc	Unknown

	`TM`
	Unknown
	
	`Ld`
	linkage disequilibrium
	
	`Pcm`
	PCR of microdissected chromosome segments (see REI)
	
	`REA`
	combined with somatic cell hybridization
	
	`FD`
	Unknown
	
	`HZ`
	Unknown
	
	`LOH`
	Unknown
	
	`ch`
	chromosomal change associated with particular phenotype and not proved to represent linkage (Fc), deletion (D), or virus effect (V)
	
	`REC`
	hybridization of cDNA to genomic fragment (by YAC, PFGE,microdissection, etc.)
	
	`REN`
	neighbor analysis in restriction fragments
	
	`Rec`
	hybridization of cDNA to genomic fragment (by YAC, PFGE,microdissection, etc.)
	
	`Re`
	Restriction endonuclease techniques
	
	`Rn`
	Unknown
	
	`fused with MOZ in AML`
	Unknown

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
status (omim_status)	Describes the certainty with which assignment of loci to chromosomes or the linkage between two loci has been established has been graded into classes C, P, I and L. Cardinality: 1 MySQL Type: VARCHAR (1) Accepted Values: `C` confirmed - observed in at least two laboratories or in several families `P` provisional - based on evidence from one laboratory or one family `I` inconsistent - results of different laboratories disagree `L` limbo - evidence not as strong as that provisional, but included for heuristic reasons (Same as `tentative')  `B` Unknown

title (omim_title)	The complete name of a gene. Cardinality: 1 MySQL Type: TEXT Examples: `Ribosomal protein L11`
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Novel Variants

Mutation effect prediction using snpEff

Version: 3.6 (for Ensembl 75)

Track Description: This track contains information on the genic region of variants and on the transcriptional or translational effect of variants. It includes for example changes that result in frameshifts, residue level changes, and introduction or skipping of a stop codon. All of these are calculated on the fly, and therefore are not available in the download version of Genome Trax. The effects are calculated using the software snpEff[1], based on the gene models available in ENSEMBL v65.

Variations that fall within in introns, exons, coding, regulatory and intergenic regions are mapped and results displayed.

Benefit: Frameshift and other non synonymous mutations frequently result in severe genetic diseases. This track will help in identification of novel plus annotated frameshift mutations in the input set

Track Name: snpeff

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
aa_change (snpeff_aa_change)	The change in the protein with the residue number as predicted by snpEff Cardinality: 1 MySQL Type: TEXT
aa_length (snpeff_aa_length)	Length of protein in amino acids. Cardinality: 1 MySQL Type: BIGINT
accession (transcript)	Ensembl transcript ID (ENST) Cardinality: 1 MySQL Type: TEXT
biotype (snpeff_biotype)	The transcript classification as reported by ENSEMBL. Cardinality: 1 MySQL Type: VARCHAR (14) Accepted Values:

	`protein coding`
	`pseudogene`
	`non-coding RNA`

brief (feature)	Dynamically generated HGVS notation for the frameshift Cardinality: 0..1 MySQL Type: TEXT

coding_status
(snpeff_coding_status)
	Indicates if the transcript is protein coding or not.
	Cardinality: 1
	MySQL Type: VARCHAR (9)
	Accepted Values:
	`CODING`
	`NONCODING`

codon_change (snpeff_codon_change)	Codon change: old_codon>new_codon. Cardinality: 1 MySQL Type: TEXT

	Effect of this variant. A detailed documentation of the effect is described here.
	Cardinality: 1
	MySQL Type: VARCHAR (33)
	Accepted Values:
	`INTERGENIC`
	The variant is in an intergenic region
	`UPSTREAM`
	Upstream of a gene (default length: 5K bases)
	`UTR_5_PRIME`
	Variant hits 5'UTR region
	`UTR_5_DELETED`
	The variant deletes and exon which is in the 5'UTR of the transcript
	`START_GAINED`
	A variant in 5'UTR region produces a three base sequence that can be a START codon.
	`SPLICE_SITE_ACCEPTOR`
	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).
	`SPLICE_SITE_DONOR`
	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).
	`START_LOST`
	Variant causes start codon to be mutated into a non-start codon. aTg/aGg, M/R
	`SYNONYMOUS_START`
	Variant causes start codon to be mutated into another start codon. Ttg/Ctg, L/L (TTG and CTG can be START codons)
	`CDS`
	The variant hits a CDS.
	`GENE`
	The variant hits a gene.
	`TRANSCRIPT`
	The variant hits a transcript.
	`EXON`

effect (snpeff_effect)	The variant hist an exon.
	
	`EXON_DELETED`
	A deletion removes the whole exon.
	
	`NON_SYNONYMOUS_CODING`
	Variant causes a codon that produces a different amino acid Tgg/Cgg, W/R
	
	`SYNONYMOUS_CODING`
	Variant causes a codon that produces the same amino acid Ttg/Ctg, L/L
	
	`FRAME_SHIFT`
	Insertion or deletion causes a frame shift An indel size is not multiple of 3
	
	`CODON_CHANGE`
	One or many codons are changed An MNP of size multiple of 3
	
	`CODON_INSERTION`
	One or many codons are inserted An insert multiple of three in a codon boundary
	
	`CODON_CHANGE_PLUS_CODON_INSERTION`
	One codon is changed and one or many codons are inserted An insert of size multipleof three, not at codon boundary
	
	`CODON_DELETION`
	One or many codons are deleted A deletion multiple of three at codon boundary
	
	`CODON_CHANGE_PLUS_CODON_DELETION`
	One codon is changed and one or more codons are deleted A deletion of size multiple of three, not at codon boundary
	
	`STOP_GAINED`
	Variant causes a STOP codon Cag/Tag, Q/*
	
	`SYNONYMOUS_STOP`
	Variant causes stop codon to be mutated into another stop codon. taA/taG, */*
	
	`STOP_LOST`
	Variant causes stop codon to be mutated into a non-stop codon Tga/Cga, */R
	
	`INTRON`
	Variant hist and intron. Technically, hits no exon in the transcript.
	
	`UTR_3_PRIME`
	Variant hits 3'UTR region
	
	`UTR_3_DELETED`
	The variant deletes and exon which is in the 3'UTR of the transcript
	
	`DOWNSTREAM`
	'Downstream of a gene (default length: 5K bases)'
	
	`INTRON_CONSERVED`
	The variant is in a highly conserved intronic region
	
	`INTERGENIC_CONSERVED`
	The variant is in a highly conserved intergenic region
	
	`INTRAGENIC`
	The variant hits a gene, but no transcripts within the gene
	
	`RARE_AMINO_ACID`
	The variant hits a rare amino acid thus is likely to produce protein loss of function
	
	`NON_SYNONYMOUS_START`
	Variant causes start codon to be mutated into another start codon (the new codon produces a different AA).

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
entrez_gene_id (entrez)	External identifier. Entrez gene ID for the gene Cardinality: 0..* MySQL Type: BIGINT

function (snpeff_function)	Functional class
	Cardinality: 1
	MySQL Type: VARCHAR (8)
	Accepted Values:
	`NONE`
	`SILENT`
	`MISSENSE`
	`NONSENSE`

gene (snpeff_gene)	Ensembl gene ID (ENSG) Cardinality: 1 MySQL Type: TEXT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hgvs (snpeff_hgvs)	Dynamically generated HGVS notation for the frameshift Cardinality: 1 MySQL Type: TEXT
hyperlink	Ensembl transcript page Cardinality: 0..1 MySQL Type: TEXT

impact (snpeff_impact)	Effect impact. The method that the snpEff categorizes the impact of the variation is listed here.
	Cardinality: 1
	MySQL Type: VARCHAR (8)
	Accepted Values:
	`High`
	`Moderate`
	`Low`
	`Modifier`

	Pubmed ID of the reference from which the information was taken.

pmid	Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

References:

1. Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M. and others. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, Fly, 6, 2, 2012

BIOBASE Trio Analysis

Version:

Track Description: Trio analysis identifies variants that might be disease causing in the offspring of unaffected parents, such as variants for which both parents are heterozygous and the child is homozygous or compound heterozygous, or De-novo mutations in the child not inherited from either parent. The input for trio analysis is a VCF file with the genotype information from all three family members.

Benefit: This track will help in identification of inherited mutations that are potentially disease causing, in the offspring if variations from parents are also available.

Track Name: trio

Annotation Fields

NAME (LEGACY NAME)	DESCRIPTION
accession	Unique number from input file. Cardinality: 1 MySQL Type: TEXT
brief (feature)	Dynamically generated HGVS notation for the frameshift. Cardinality: 0..1 MySQL Type: TEXT

consequence	The class of the inherited mutation.
	Cardinality: 1..*
	MySQL Type: VARCHAR (12)
	Accepted Values:
	`hom-from-het`
	`De-novo`
	`compound het`

ensembl_id (ensembl)	External identifier. Ensembl gene ID for the gene. There may be several identifiers. If there are Locus Reference Genomic available for the gene, in addition to the ENSG identifiers there may be LRG_ identifiers. Also, if there are fix patches for the gene or the gene is variable, there may be additional identifiers. Cardinality: 0..* MySQL Type: TEXT
	External identifier. Entrez gene ID for the gene

entrez_gene_id (entrez)	Cardinality: 0..* MySQL Type: BIGINT
hgnc	HGNC gene symbol for the gene. If the track describes features that are not directly linked to a gene, and a hgnc symbol is present, it refers to the gene closest downstream or overlapping with the feature. Cardinality: 0..* MySQL Type: TEXT
hyperlink	Link to a report or web-page with more detailed information. Cardinality: 0..1 MySQL Type: TEXT

pedigree_status	The disease status, the sample identity and sex of the individual encoded.
	Cardinality: 1
	MySQL Type: VARCHAR (17)
	Accepted Values:
	`M_FATHER_NORMAL`
	`M_FATHER_AFFECTED`
	`F_MOTHER_NORMAL`
	`F_MOTHER_AFFECTED`
	`M_CHILD_NORMAL`
	`M_CHILD_AFFECTED`
	`F_CHILD_NORMAL`
	`F_CHILD_AFFECTED`

pmid	Pubmed ID of the reference from which the information was taken. Cardinality: 0..* MySQL Type: BIGINT
uniprot_acc (uniprot)	External identifier. UniProt accession number for the protein. If there are several possible sequences, the canonical sequence and accession number is used. Cardinality: 0..* MySQL Type: TEXT

Protocol used to find the closest gene

Some tracks, especially the regulatory tracks describe elements outside the boundaries of a gene. Others describe elements that fall into the boundary of one, or several genes, but the data source does not specify a gene. We try to provide a HGNC gene symbol for all records. Only one gene symbol is provided, even if there are several genes in close proximity. The closest gene is selected using the following protocol, based on gene definitions from Ensembl, where the start and end of the gene are defined as the first and last nucleotide of the first and last exon.

1. If the datasource specifies a gene symbol, we provide this symbol.
2. If a track does not contain strand information (eg: Microsatellites and CpG Islands), consider both strands and return the gene which is closest on any of the two strands according to the following rules. Otherwise, we only consider the strand on which the elemnt is defined.
3. In case the element is not part of a gene, then the neighboring gene with the closest start or end position is selected.
4. In case the given position is part of one or more genes, then the smallest enclosing gene is selected.
5. In case the given position is part of only one gene, then that enclosing gene is selected.

Flatfile documentation

BIOBASE Genome Trax provides selected data from the proprietary TRANSFAC® Professional, HGMD® Professional, PROTEOME™ and PGMD™ databases. It also provides annotation from a wide variety of relevant public data sources. The data is updated quarterly and corresponds to data within the most current BIOBASE Knowledge Library or HGMD Professional release.

This documentation describes the file formats for these annotation tracks. The formats are used for the download version of Genome Trax™, for downloading result sets from the online version of Genome Trax™, and for results from the web service version of Genome Trax™.

Genome Trax™ is available in three standard flat file formats, VCF, BED and GFF, as well as a relational MySQL database dump. GFF and VCF files as well as the MySQL database contain much richer annotation content than the BED files, and are recommended if you want to access the full data content.

Genome Trax™ provides one file set for each supported genome build, hg18/NCBI36 and hg19/GRCh37. Each track is available for these builds in all flat file formats. The relational version of the tracks contains data for all the builds in the same schema.

BED file format

Genome Trax™ provides two kinds of (UCSC) Genome Browser-optimized BED-format files for each track. For a definition of the BED format, see here.

A *description* BED file provides a brief description of the feature of interest. A *linking* BED file contains accession numbers to more detailed reports for a particular feature. Please note that for some tracks, the access to detail records from the BIOBASE Knowledge Library requires a subscription to the relevant BIOBASE database. This is mentioned in the individual track description, and usually applies to tracks that link to highly connected, deep annotation like pathways, for which the full power of the BIOBASE Knowledge Library interface is advantageous. For some tracks, where all of the information on the feature is provided in the description, there are no links to detail reports. Again, this is mentioned in the individual track description.

Tech Note: In standard BED files, only one column can have a textual description of the feature, the fourth or “Name” column. The UCSC browser requires the entire content of this field to substitute placeholders in links to additional information. If one wants to have such links, no human readable annotation can be provided in addition to an accession identifier in this field. Therefore we provide two versions of BED files, one with accession numbers, and one with a human readable description.

Columns for BED files (UCSC optimized)

Columns within BED files are uniform regardless of the data track presented. However, there are differences in the content of the “Name” column, which varies by track.

Each BED file is preceded by a header defining the feature track, description, and settings for rendering the track within the genome browser. An example is shown below:

```
track name=HGMD® Mutations description=HGMD® Mutations color=176,23,31
visibility=3 hyperlink=https://genometrax.biobase-
international.com/static/hgmd_reports/$$
```

By default, the **visibility** parameter is set to 3. The values for the other rendering fields for each track are listed under the track description (or you can just look at the line in the files). The **hyperlink** parameter is only required for the linking BED files, and may be absent in the human readable BED files or the BED files that are exported from the online interface.

Columns contained within UCSC-optimized BED files are listed below in the order in which they appear within the BED files, along with a description of the information described within the column:

chromosome

The number of the chromosome on which the feature is located. This is given as the three letters `chr`, followed by the integer number of the chromosome (without leading

zeroes), or an uppercase X , Y or M (for mitochondrial DNA), e.g. `chrX`, or `chr7`. Please note that the `chr` part is case sensitive, `Chr7` will not be accepted.

start

The genome coordinate corresponding to the start of the feature, this value uses 0-based scoring where 1 is subtracted from the actual start site (which is a peculiarity required by the UCSC browser). If you wish to extract the actual start coordinate for a set of features without converting back to 1-based scoring, please use the GFF files.

end

The genomic coordinate corresponding to the end of the feature.

name

The contents of this column vary by individual track, and are listed in detail in the track description. They also depend on the nature of the BED file:

In the UCSC-optimized *linking* file, this column provides an accession used for linking to external database reports. Together with the base URL in the header line this allows you to follow hyperlinks from the UCSC browser to pages with more detail, if you upload the track as a custom annotation track. The detailed kind of accession number is listed in the track description under “accession”.

In the UCSC-optimized *description* file, this column provides a short, human readable description of the feature, for direct display in the genome browser. If there are several values to be listed in the field, they are separated by semicolons. If values contain spaces, these are replaced by underscores. The detailed description of the contents of this field is listed in the track description under “brief”.

Please note that a single row in the linking file may correspond to multiple rows in the corresponding descriptive file. For example, a single TRANSFAC binding site may be shared by multiple transcription factors. The descriptive file will contain one row for each transcription factor, with each row containing the same start and end coordinates, while the linking file will contain a single row providing the binding site accession. If this is the case, it is mentioned in the track description.

score

0 indicates a default setting of no score. This column is present only when strand column is also present. The current release does not have data for scores, other than the default.

strand

The DNA strand corresponding to the start and end of the feature. This column is only present when strand data is available. "+" indicates forward strand and "-" indicates reverse strand.

GFF file format

Genome Trax™ provides one GFF format file for each track. This is optimized for the CLC Genomics Workbench, and compatible with other genome analysis platforms such as Galaxy, which utilize GFF format files. For a definition of the GFF3 format see http://www.sequenceontology.org/gff3.shtml

The GFF files are designed to be parseable by bioinformaticians interested in incorporating Genome Trax™ data into custom workflows and applications.

Columns for GFF files

Columns within GFF files are uniform regardless of the data track. However, there are track-specific differences in the content especially of the "Attributes" column, which holds annotations. Columns may also contain no data.

Columns contained within GFF files are listed below in the order in which they appear within the GFF files, along with a description of the information described within the column:

chromosome

The number of the chromosome on which the feature is located. This is given as the three letters `chr`, followed by the integer number of the chromosome (without leading zeroes), or an uppercase X, Y, or M (for mitochondrial DNA), e.g. `chrX` or `chr7`.

source

The name of the track, as listed in the individual track description for each track under "Track Name". Spaces in the name are replaced by underscores.

description

A short, human readable description of the feature. This is listed in the individual track description under "brief".

start

A genomic coordinate corresponding to the start of the feature.

end

A genomic coordinate corresponding to the end of the feature.

score

A period (.) indicates default setting of no score. Currently we do not assign a score.

strand

The DNA strand corresponding to the start and end of the feature. "+" indicates forward strand, "-" indicates reverse strand, and period (.) indicates no strand information.

frames

The open reading frame corresponding to a feature. A period (.) indicates a default setting of no frame.

attributes

This column provides annotations in the form of key=value pairs. Every track has its own set of annotation fields, which may contain hyperlinks to detailed report pages, to PROTEOME or TRANSFAC (which requires a subscription), or to external databases.

We provide extensive detail in the attributes of the GFF files to allow you to use these flat files for parsing. You can easily identify supporting evidence for the relevance of the annotation, without following links out to further pages.

The format for attributes is key=value. Multiple key=value pairs are separated by semicolons. URL escaping rules are used for keys or values containing the following characters: `,=;`. Spaces are allowed in this field, but tabs are replaced with the %09 URL escape. Multiple attributes of the same type are indicated by separating the values with the comma `","` character.

VCF file format (beta)

Genome Trax™ provides one VCF format file for each track. The VCF files are designed to be parseable by bioinformaticians interested in incorporating Genome Trax data into custom workflows and applications. VCF also is compatible with many genome analysis tools.

INFO column in VCF files

All columns within VCF follow the VCF specification, except for the INFO column. The annotation data is provided as key=value pairs in this INFO column, in double quotes (`" "`) for values that contain white-space, semi-colons, equal-signs or commas (both white space and commas are relatively common in annotation strings).

We provide extensive detail in the INFO column of the VCF files to allow you to use these flat files for parsing. You can easily identify supporting evidence for the relevance of the annotation, without following links out to further pages.

Tech Note: VCF does not currently allow white-space, semi-colons, or equals-signs in the INFO field; commas are permitted only as delimiters for lists of values. The VCF specification does not currently specify how such values should be encoded, if they happen to be part of the content of the info field. One could encode them by defining ##INFO keys in the header, but this would make the file much less directly readable. Since there is a proposal to allow these characters in dounble-quoted strings in the info field, instead of inventing ad-hoc encoding, we double quote values that contain such characters.

Index of Files and Space Requirements

The following tar.gz packages are included within the Genome Trax product download, which include sets of GFF, BED and VCF files containing hg18/NCBI36 or hg19/GRCh37 genomic coordinates for each feature, as well as the MySQL relational dump. Files are bundled into archives using tar, and compressed using gzip:

```
Genome_Trax_hg18_bed.tar.gz
Genome_Trax_hg18_gff.tar.gz
Genome_Trax_hg18_vcf.tar.gz
Genome_Trax_hg19_bed.tar.gz
Genome_Trax_hg19_gff.tar.gz
Genome_Trax_hg19_vcf.tar.gz
GenomeTrax_2014.3.sql.gz
```

You will need approximately the following amount of space for downloading and installing these files:

Compressed Files: 50 GB

Uncompressed Flatfiles: 500 GB

Uncompressed DB Dump (SQL File): 350 GB

MySQL DB: 350 GB

As we continuously add to Genome Trax™, we recommend you reserve at least 2 TB of disk space.

Migration Note: We migrated GFF for all tracks files to use unified names for a number of fields containing the same kind of data, such as PMIDs, Uniprot-IDs, HGNC symbols, to make it easier to aggregate across tracks. The 2013.3 release is the last release to contain in addition to the new gff files, gff files with the legacy gff file format. These are provided in the compressed archives `Genome_Trax_legacy_hg18_gff.tar.gz` and `Genome_Trax_legacy_hg19_gff.tar.gz`, and each of the files also has the postfix _legacy inserted after the track name.

File names

Each track has its own base file name, as listed under "Track Name" in the individual track descriptions below. In the case of BED files, _linking is appended for linking files that contain an accession number, or _description is appended for files that contain a human-readable description. The track names then are postfixed by an underscore and the genome build. This is followed by the filetype extension, .bed, .gff or .vcf.

For example, the TRANSFAC® experimentally verified TFBS track has the base file name *transfac_sites*. The BED file mapped to the genome build hg19 containing accession numbers for linking would be named *transfac_sites_linking_hg19.bed*.

Whats New? (Release Notes)

Release 2014.3

Dataset from Online Mendalian Inheritance in Man

OMIM is a catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression. The new OMIM track allows you to associate known patterns of inheritance (dominant, recessive) with diseases or phenotypes and the genes implicated in them. Together with the observed zygosity, and the disease causing mutations in HGMD, this can help you to focus only on dominant variants, or on recessive variants that are homozygous in the patient sample. This track is available only for download users.

Data update

The release contains new content for several major tracks. See the statistics page for more detail.

Release 2014.2

Notify by e-mail for long running jobs

It is now possible to set a notify option, selecting which will send an e-mail to your registered account once the Genome Trax™ annotations are completed. There is also facility to view the completed results for a limited period of time. Similar alerts are sent for export requests also.

Data update

The release contains new content for several major tracks. See the statistics page for more detail.

Release 2014.1

Side-by-side variant summary

The variant summary now makes it easier for you to compare samples. It now displays information such as genotype for all samples in VCF file side by side, on a common line. Annotations for these samples are aggregated. The summary now displays information more densely and by default shows displays columns that carry annotation, again making it easier to compare data on-screen.

New severity scale

The severity assesment now follows the five step scale from pathogenic to likely pathogenic to variant of unknown significance to likely not pathogenic and finally not pathogenic. The highest class is reserved to HGMD disease causing variants. Please note the Genome Trax is NOT A DIAGNOSTIC TOOL.

Export Queueing (beta)

When exporting large result sets, generating the export can take a while. We now indicate that an export is running, and you can pick up the export files, when they are ready.

Data update

The release contains new content for several major tracks. See the statistics page for more detail.

Release 2013.4

Data update

The release contains new content for several major tracks. See the statistics page for

more detail.

Release 2013.3

Gene Panels

The new gene panel feature allows you to only focus on variants in genes that are known to be involved with a given disease, and not report results for other variants. You can directly select the panel by disease, or you can upload your own tailored gene list.

Population-specific Allele Frequencies

To enable better filtering and selection if the population background of the sample is known, in addition to general allele frequencies, this build now includes allele frequncies from the 1000 Genomes project that are specific for European (CEU), African (YRI), or Asian (JPT/CHB) heritage.

Haplotype matching

The pharmacogenomic data from PGMD, the Pharmaco-Genomic Data Base provides detailed genotypes and haplotypes for it's associations. The new version of Genome Trax is able to do haplotype- and genotype-exact matching. If you rather would like to see a broad match, instead of an specific one, as with HGMD it is possible to just match based on genomic positions.

OrphaNet inheritance mode and trait information

The new OrphaNet annotation track allows you to associate known patterns of inheritance (dominant, recessive) with rare diseases and the genes implicated in them. Togehter with the observed zygosity, and the disease causing mutations in HGMD´, this can help you to focus only on dominant variants, or on recessive variants that are homozygous in the patient sample. It also provides prevalence information and observable traits ("signs") associated with the diseases.

Data update

The release contains new content for several major tracks. See the statistics page for more detail.

Release 2013.3

Gene Panels

The new gene panel feature allows you to only focus on variants in genes that are known to be involved with a given disease, and not report results for other variants. You can directly select the panel by disease, or you can upload your own tailored gene list.

Population-specific Allele Frequencies

To enable better filtering and selection if the population background of the sample is known, in addition to general allele frequencies, this build now includes allele frequncies from the 1000 Genomes project that are specific for European (CEU), African (YRI), or Asian (JPT/CHB) heritage.

Haplotype matching

The pharmacogenomic data from PGMD, the Pharmaco-Genomic Data Base provides detailed genotypes and haplotypes for it's associations. The new version of Genome Trax is able to do haplotype- and genotype-exact matching. If you rather would like to see a broad match, instead of an specific one, as with HGMD it is possible to just match based on genomic positions.

OrphaNet inheritance mode and trait information

The new OrphaNet annotation track allows you to associate known patterns of

inheritance (dominant, recessive) with rare diseases and the genes implicated in them. Togehter with the observed zygosity, and the disease causing mutations in HGMD, this can help you to focus only on dominant variants, or on recessive variants that are homozygous in the patient sample. It also provides prevalence information and observable traits ("signs") associated with the diseases.

Data update

The release contains new content for several major tracks. See the statistics page for more detail.

Release 2013.2

Trio analysis

Trio analysis identifies variants that might be disease causing in the offspring of unaffected parents, such as variants for which both parents are heterozygous and the child is homozygous or compound heterozygous, or 'De-novo' mutations in the child not inherited from either parent. The input for trio analysis is a VCF file with the genotype information from all three family members.

Enhanced filtering based on allele frequency

It is now possible to select the allele frequency of for removing common input variants. The larger the allele frequency cutoff, the fewer data will be filtered out, and more variants from the input file will be considered for annotation.

dbSNP annotations

Genome Trax™ now offers dbSNP as an annotation track. This track will allow one to assess if a variant in your file is a known rsID in dbSNP. This also provides minor allele frequencies for all variants that are known from dbSNP..

VCF formatted track files

We now provide the VCF formatted files for tracks where the nucleotide changes are known in the download version.

Data update

The release contains new content for several major tracks. See the statistics page for more detail.

Release 2013.1

ClinVar

We integrate the clinically significant variants from NCBI's ClinVar resource as a new track. These phenotype - variant maps are derived from various studies and are expert reviewed.

Asynchronous procesing for exports

When the user annotates variants with low-level annotation like the genomic region, or with predicted outcomes, instead of just known, literature reported findings, a large number of annotated variants typically results. Exporting those to Excel or other download files takes time. Long running exports are now asynchronously processed, so that you can continue to work with the tool while the export is being prepared in background

Aggregated results for gene-based variants

We now match annotations for gene based variants (e.g. known disease annotations for the gene) specifically for the coding sequence. Furthermore, we aggregate all annotations for a category (pathway, gene, disease) into a single result, and report this for all matching submitted variants. This corrects the slightly unexpected behavior, where such annotations were only reported for the first variant for which they were

observed.

Data update

The release contains new content for major tracks. See the statistics page for more detail.

Release 2012.4

Filtering on common SNPS from HGMD®

Some mutations are disease causing, though they have a population frequency of >5%. The common SNP filter track ignores the HGMD® mutations to maximise the HGMD® results, while this track allows users to filter out all the SNPs that have a higher population frequency.

HGMD Imputed mutations

Some disease causing mutations in HGMD® are a result of amino acid changes. This track collects all such mutations falling within an exon, and compute all possible nucleotide changes within the codon that would result in the original amino acid mutation as described in HGMD®. This track will help to identify novel mutations.

Predicted TFBS sites within hypersensitive DNAse sites

Hypersensitivity to nuclease clevage is an indication of active transcription and potential transcription factor binding sites. From the DNAse fragments from 142 ENCODE data sets, this track lists the predicted potential transcription factor binding sites. This track will help to identify the potential regulatory effects of the mutations that falls in this region.

Enhanced filters for HGMD® mutations

The filter tab now also has the ability to restrict the result based on HGMD® variant type.

Data update

The release contains new content for major tracks. See the statistics page for more detail.

Release 2012.3

Uniprot Post-Translational Modifications (PTMs)

The revamped PTM track now includes annotation from the Uniprot database, greatly increasing the coverage of known modification sites.

Zygosity support

Genome Trax now reads, infers and reports zygosity calls for variants from VCF or Complete Genomics Mastervar files. Zygosity can be used as an additional evidence in support of an identified disease-causing variant.

Gene region and Translation effect

GenomeTrax now integrates SNPeff to support annotation of the genic region and translational effect of variants, for example splice-site variants, frameshift and truncating variants, and non-synomymous variants. This is especially useful to evaluate high-impact novel variants that do not match curated annotation. Note that we do not report low impact predictions unless they also match some other evidence.

Gene Summary

The new gene summary tab provides a quick overview over the number of variants and relevant annotation found for each gene, which you can use as a starting point to drill down on annotation details. It also ranks genes by an overall severity assessment for the gene that is based on all found annotation. Additional in-tool filters help you to

make better use of this and the Variant Summary tab.

User Track Support

It now is possible to upload your own annotation and filter track, for example to provide in-house collection of commonly seen variants for filtering, or to annotate for features from UCSC that are not included. Tracks need to be provided in BED format. This feature only applies to the online version of GenomeTrax.

Data update

The release contains new content for major tracks. See the statistics page for more detail.

Release 2012.2

Common Variant Filtering

A new track of common variants from dbSNP and the 1000 Genomes project allows you to filter out variants that are commonly (more than 1% allele frequency) observed in the general population. You can recover the filterd variants in a separate tab in the results page.

Allele Frequencies

A new track from the EVS (Exome Variant Server) provides detail information on Allele Frequencies for different populations for exome single nucleotide variants.

OMIM disease genes

A new track of disease genes from OMIM adds further support from another annotation source to the already existing disease gene tracks from HGMD and Proteome.

Variant Summary and focus tools

The new variant summary tab provides a quick overview over relevant annotation found for each input variant, which you can use as a starting point to drill down on annotation details. You can now focus on annotations in the results page by HGNC symbol and genomic coordinate. You can also focus on annotations in the results page by HGNC symbol and genomic coordinate.

Data update

The release contains new content for major tracks. See the statistics page for more detail.

Release 2012.1

Pharmacogenomic Variants

This new track contains variants that are associated with pharmacogenomic effects, including detail annotation about the supporting evidence and studies. Please note that this track is a beta-release version and still may change. Let us know if you have suggestions for improvement.

Data update

The release contains new content for major tracks. See the statistics page for more detail.

Release 2011.4

Personal Genome Interpretaion

Genome Trax supports now direct input of variation files from 23andMe to enable you to learn what is already known about the SNPs from your personal genome sequence.

Use of Genes or rsIDs as input

To make it easier to look up annotation information that is known for individual SNPs or Genes, Genome Trax now accepts HGNC gene identifiers or rsIDs, in addtion to genomic intervals, as input data.

miRNA

This new track contains known miRNA sites from miRBase. This is available in all versions.

HGNC

We make an effort to support HGNC identifiers for all tracks, so that you can aggregate findings on a per gene level across all annotation types.

Exon, Intron, UTR annotation

Genome Trax allows you to characterize variants based on gene structure information (from Ensbembl).

Frameshift mutations

Genome Trax now also identifies frame shift mutations.

Data update

The release contains new content for major tracks. See the statistics page for more detail.

Release 2011.3

Data update

The release contains new content for major tracks. See the statistics page for more detail.

GWAS mutations

We added a new track on disease associated risk mutations from GWAS studies to enable even more comprehensive mutation screening. This is available in all versions.

Novel Mutation filtering

The new version is able to identify novel variants by removing all known (annotated) matches.

UCSC direct export

You now can visualize the results of a search directly in UCSC genome browser, at a single button.

Complete Genomics support

Genome Trax is now able to natively read Complete Genomics var files.

Release 2011.2

Data update

The release contains new content for major tracks. See the statistics page for more detail.

COSMIC somatic mutations

We added a new track on somatic mutations from COSMIC to enable even more comprehensive mutation screening. This is available in all versions.

Exact mutation change matching

In cases where you provide the exact observed change for a single nucleotide variant, and we also know the exact nucleotide change in our annotation (for example, for point mutations in HGMD), we will only predict a match if the changes also exactly match. The change has to be provided in the fourth input column.

VCF format support

Genome Trax is now able to natively read VCF formatted input files.

Generic tab separated input support

You can now upload any kind of TAB separated file. Additional columns that might have been prepared by preprocessing software will be retained. While you may not see all of them in the online interface, if you re-export the data, your original input lines will contain the Genome Trax annotation as additional columns, saving you the effort to reconcile the annotations.

Richer Track Annotation

Individual tracks contain more annotation directly in the flat files, not just in the auxiliary reports. For example, most tracks now come with HGNC gene identifiers. The regulatory SNP tracks contain information by which criteria the SNPs were found to be regulatory.

Release 2011.1

Data update

The release contains new content for major tracks. See the statistics page for more detail.

Dropped Histone Region Track

This track was one of the largest by number of intervals, but one of the least useful in finding relevant variation. We therefore decided to discontinue it which increases the performance of the online version of the tool.

Initial Release 2010

Unique content relevant to functional genomics

- 3,600+ regulatory sites from TRANSFAC®
- 80,000+ disease linked mutations from HGMD® Professional
- 600,000+ ChIP-Seq fragments with best binding site predictions
- Disease biomarkers, drug targets, and pathway memberships from PROTEOME™
- Single Nucleotide Polymorphisms from dbSNP and Ensembl which overlap with promoter features and sites of regulation
- Post Translational Modifications
- Additional genome features such as microsatellites, transcription start sites (TSSs), and CpG islands

Packaged for ease of use in NGS applications

- Search for genome features using sequence coordinates
- Find genome feature data mapped to human reference genomes hg18/NCBI36 and hg19/GRCh37
- Export data for visualization on independent genome browsers

Questions? support@biobase-international.com