# Database schema documentation for SNPdbe

Changes
02/27/12:

- seqs_containingSNPs.taxid removed
- dbSNP_SNP.tax_id renamed to dbSNP_SNP.taxid

General information:
Data in SNPdbe is organized on several levels. Protein sequence specific information is stored in tables starting with **seqs_** , information on nsSNP level is stored in **SNPs_** tables.

The central table is **geno2func**. It collects nsSNPs from all sources and stores prediction results and evolutionary information.

Most important fields in the schema are:

- the **md5** hashsum of a sequence. It serves as the only unique sequence identifier.
- the **mut_id**. It designates a unique nsSNP throughout all tables in terms of sequence (through its md5), position and mutant residue.

Both implicitly serve as foreign keys and connect information between tables.

In the following, primary keys are underlined and highlighted in bold.

Important: We store protein sequence positions **0-based** throughout the whole schema!

| seqs_SP<br>(contains sequence information on SwissProt proteins) | | |
|---|---|---|
| md5 | char(32) | md5 hashsum of the sequence |
| seq | mediumtext | wild-type protein sequence |
| **AC** | varchar(10) | primary SwissProt accession |
| ID | varchar(20) | SwissProt identification |
| gene | varchar(20) | gene identifier |
| taxid | int(11) | NCBI taxid |
| description | mediumtext | Information on the protein |

| seqs_RefSeq<br>(contains sequence information taken from RefSeq; used to map nsSNPs from dbSNP and 1KG) | | |
|---|---|---|
| md5 | char(32) | md5 hashsum of the sequence |
| seq | mediumtext | wild-type protein sequence |
| **id** | varchar(15) | RefSeq protein identifier |
| **ver** | tinyint(3) | RefSeq protein version |
| gi | int(11) | GenBank identifier |

| | | |
|---|---|---|
| comment | text | Information on the protein |
| taxid | int(11) | NCBI taxid |

| seqs_PMD<br>(contains sequence information taken from PMD) | | |
|---|---|---|
| md5 | char(32) | md5 hashsum of the sequence |
| **pmd_id** | char(10) | PMD identifier |
| **nr** | tinyint(3) | PMD sub-identifier |
| seq | mediumtext | wild-type protein sequence |
| taxid | int(11) | NCBI taxid |

| seqs_containingSNPs<br>(stores all sequences that contain at least one nsSNP and flags to the source of nsSNPs) | | |
|---|---|---|
| **md5** | char(32) | md5 hashsum of the sequence |
| seq | mediumtext | wild-type protein sequence |
| in_dbSNP | enum('0','1') | '1' if sequence contains at least one SAAS from nsSNP |
| in_SP | enum('0','1') | '1' if sequence contains at least one SAAS from SwissProt |
| in_PMD | enum('0','1') | '1' if sequence contains at least one SAAS from PMD |
| in_1KG | enum('0','1') | '1' if sequence contains at least one SAAS from 1KG |

| seqs_equalities<br>(relates two md5s if their sequences differ only by one residue or the leading methionine; in both cases two sequences are considered as being equal) | | |
|---|---|---|
| **md5_1** | char(32) | md5 hashsum of the first sequence |
| **md5_2** | char(32) | md5 hashsum of the second sequence |
| by_missing_M | enum('0','1') | '1' if both differ only by the leading methionine |
| by_one_mismatch | enum('0','1') | '1' if both differ by one residue |
| pos_mismatch | smallint(5) | position of that mismatch residue |

| md52keywords<br>(stores gene names and descriptions for each sequence) | | |
|---|---|---|
| **md5** | char(32) | md5 hashsum of the sequence |
| gene_symbol | mediumtext | Gene symbol |
| gene_SP | varchar(11) | SwissProt identification |
| description | mediumtext | Information on the protein's function |

| geno2func<br>(central table that stores predictions and evolutionary information for each nsSNP) | | |
|---|---|---|
| mut_id | int(11) | Unique nsSNP identifier |
| **md5** | char(32) | md5 hashsum of the first sequence |
| wt | char(1) | wild type residue |
| **pos** | smallint(5) | mutation position in the sequence (0-based) |
| **mt** | char(1) | mutant residue |
| in_dbSNP | enum('0','1') | '1' if that mutation is derived from dbSNP |
| in_SP | enum('0','1') | '1' if that mutation is derived from SwissProt |
| in_PMD | enum('0','1') | '1' if that mutation is derived from PMD |
| in_1KG | enum('0','1') | '1' if that mutation is derived from 1000Genomes |
| SNAP_status | enum('0','1','2') | '0': SNAP failed, '1': internal use, '2': SNAP succeeded |
| SNAP_bin | enum('+','-') | SNAP prediction: '+' non-neutral, '-' neutral |
| SNAP_score | tinyint(4) | SNAP raw score from -100 (strongly neutral) to 100 (strongly non-neutral) |
| SNAP_ri | tinyint(1) | SNAP prediction reliability index from 0 (low reliability) to 9 (high reliability) |
| SNAP_acc | tinyint(3) | SNAP expected accuracy |
| SIFT_bin | enum('+','-') | SIFT prediction: '+' deleterious, '-' neutral |
| SIFT_score | float | SIFT raw score from 0 (strongly deleterious) to 1 (strongly neutral), decision threshold 0.05 |
| PERC_wt | tinyint(3) | Relative frequency of wildtype residue in PSI-BLAST alignments |
| PERC_mt | tinyint(3) | Relative frequency of mutant residue in PSI-BLAST alignments |
| PSSM_wt | tinyint(4) | PSI-BLAST's PSSM value for wildtype residue |
| PSSM_mt | tinyint(4) | PSI-BLAST's PSSM value for mutant residue |
| PSIC_wt | float | PSIC conservation score for wildtype residue |
| PSIC_mt | float | PSIC conservation score for mutant residue |
| pph2 | enum('possibly damaging','probably damaging','benign','unknown') | PolyPhen2 prediction for a subset of dbSNP mutants |

| dbSNP2OMIM_verified<br>(maps dbSNP mutants to disease associations in omim) | | |
|---|---|---|
| **mut_id** | int(11) | mut_id of the dbSNP mutation (ref to SNPs_dbSNP) |
| **omim_id** | int(11) | major omim entry id (ref to OMIM2disease) |
| **var_id** | varchar(5) | omim variant identifier (ref to OMIM2disease) |

| OMIM2disease<br>(relates omim ids, disease and the mutation in a given protein sequence) | | |
|---|---|---|
| **omim_id** | int(11) | major omim entry id |
| **var_id** | varchar(5) | omim variant identifier |
| phenotype | text | disease associated with that variant |
| gene | varchar(30) | gene symbol |
| mutation_raw | text | for internal use only |
| wt | char(1) | wildtype residue found at the disease causing mutation |
| pos | int(11) | protein sequence position (0-based) |
| mt | char(1) | mutant residue |

| taxid2names<br>(maps a NCBI taxid onto organism names; see ftp://ftp.ncbi.nih.gov/pub/taxonomy/) | | |
|---|---|---|
| **taxid** | int(11) | NCBI's taxid |
| misspelling | varchar(255) | Columns below denote different alternatives for the organism |
| genbank_anamorph | varchar(255) | name. |
| scientific_name | varchar(255) | |
| synonym | varchar(255) | |
| blast_name | varchar(255) | |
| unpublished_name | varchar(255) | |
| genbank_synonym | varchar(255) | |
| equivalent_name | varchar(255) | |
| includes | varchar(255) | |
| acronym | varchar(255) | |
| in_part | varchar(255) | |
| anamorph | varchar(255) | |
| authority | varchar(255) | |
| genbank_common_name | varchar(255) | |
| genbank_acronym | varchar(255) | |
| common_name | varchar(255) | |
| misnomer | varchar(255) | |
| teleomorph | varchar(255) | |

| taxid_merged<br>(relates identical taxids that got merged in the past) | | |
|---|---|---|
| **taxid_old** | int(11) | Taxid that was deleted and merged with... |
| taxid_new | int(11) | ...new taxid. |

| dbSNP_ValCode<br>(Defines the validation status of a snp in dbSNP; taken as is from table dump provided by dbSNP) | | |
|---|---|---|
| **code** | tinyint(3) | Validation status code |
| abbrev | varchar(50) | Abbreviation of the validation status code |
| descrip | varchar(250) | Description of the validation status |
| create_time | varchar(50) | Datetime when the row was inserted into the table; unused |
| last_updated_time | varchar(50) | Datetime when the row was updated; unused |

| dbSNP_SNP<br>(collects additional info on each dbSNP rs; taken as is from dbSNP's organism specific SNP tables) | | |
|---|---|---|
| **snp_id** | int(11) | dbSNP rs id; ref to SNPs_dbSNP |
| avg_heterozygosity | float | average heterozygosity of a snp |
| het_se | float | standard error of average heterozygosity |
| create_time | varchar(30) | first time this refSNP cluster is created (from dbSNP); unused |
| last_updated_time | varchar(30) | last time this row of SNP data is updated (from dbSNP); unused |
| CpG_code | tinyint(4) | currently unused |
| taxid | int(11) | NCBI taxid of the organism |
| validation_status | tinyint(4) | ref to dbSNP_ValCode; definition of each validation status |
| exemplar_subsnp_id | int(11) | currently unused |
| univar_id | int(11) | currently unused |
| cnt_subsnp | tinyint(3) | number of ss in that rs; unused |
| map_property | tinyint(3) | currently unused |

| SNPs_SP<br>(collects all VARIANT,MUTAGEN,CONFLICT entries from SwissProt plus annotations) | | |
|---|---|---|
| mut_id | int(11) | Unique mutant id; ref to geno2func |
| **md5** | varchar(32) | Md5 hashsum of the protein sequence |
| AC | varchar(10) | Primary SwissProt accession |
| taxid | int(11) | NCBI taxid |
| **pos** | smallint(5) | Mutant position (0-based) |
| **mt** | char(1) | Mutant residue |
| **kind** | enum('VARIANT | Kind of the amino acid change; one of VARIANT (observed |

| | | |
|---|---|---|
| | ','MUTAGEN','CONFLICT') | variation), MUTAGEN (mutagenesis), CONFLICT (sequencing conflicts) |
| site_features | text | Features observed at the mutant position; currently:<br>Active sites (ACT_SITE), Metal binding (METAL), PTMs (MOD_RES), general binding (BINDING), Lipid (LIPID), Glycosylation (CARBOHYD), disulfide bond (DISULFID), located in a domain? (DOMAIN) |
| functional_effect | text | effect of the variant; MUTAGEN only |
| var_id | varchar(7) | SwissVar id, VARIANT only |
| disease | text | Disease association with the VARIANT |

| SNPs_dbSNP<br>(collects all nsSNPs from dbSNP; merges all organism-specific SNPContigLocusId tables from dbSNP plus additional information) | | |
|---|---|---|
| mut_id | int(11) | Unique mutant id; ref to geno2func |
| **build** | smallint(6) | dbSNP build for that organism |
| **assembly** | varchar(10) | version of the reference genome |
| verified | enum('0','1','2','3') | 0: OK, 1: mutant position out of sequence, 2: reported mt found at given position, 3: no RefSeq sequence available;<br>note, only mutants with status '0' inserted into geno2func! |
| md5 | varchar(32) | Sequence md5 hashsum |
| **taxid** | int(11) | NCBI taxid |
| **snp_id** | int(11) | dbSNP rs id |
| **contig_acc** | varchar(32) | contig accession; unused |
| contig_ver | tinyint(4) | contig version; unused |
| asn_from | int(11) | starting position on contig (0-based); unused |
| asn_to | int(11) | end position on contig (0-based); unused |
| locus_id | int(11) | NCBI locus id; unused |
| locus_symbol | varchar(64) | locus name |
| mrna_acc | varchar(32) | accession of mrna refseq; unused |
| **mrna_ver** | int(11) | version of mrna refseq; unused |
| **protein_acc** | varchar(32) | protein accession |
| protein_ver | int(11) | protein accession version |
| **fxn_class** | int(11) | dbSNP internal variation code; only contains 42; i.e. nsSNP |
| reading_frame | int(11) | base position in codon; unused |
| **allele** | varchar(255) | snp allele in the orientation of the mrna_acc; unused |
| mt | varchar(8) | mutant amino acid |
| **pos** | int(11) | protein sequence position of the mutant (0-based) |
| ctg_id | int(11) | contig id; unused |

| | | |
|---|---|---|
| mrna_pos | int(11) | position in mRNA; unused |
| mrna_start | int(11) | start position in mRNA; unused |
| mrna_stop | int(11) | stop position in mRNA; unused |
| codon | varchar(257) | dDbSNP internal usage; unused |
| protRes | varchar(8) | 3-letter code for mutant amino acid |
| contig_gi | int(11) | unused |
| mrna_gi | int(11) | unused |
| mrna_orien | tinyint(4) | unused |
| cp_mrna_ver | int(11) | unused |
| cp_mrna_gi | int(11) | unused |
| verComp | varchar(7) | unused |

| SNPs_PMD<br>(collects all SAASs from PMD (ProteinMutantDatabase); also contains effects on function and disease) | | |
|---|---|---|
| mut_id | int(11) | unique mutant id; ref to geno2func |
| md5 | varchar(32) | sequence md5 hashsum |
| pmd_id | char(10) | PMD primary identifier |
| nr | tinyint(3) | PMD subidentifier |
| authors | text | unused |
| journal | varchar(70) | unused |
| title | text | unused |
| medline | varchar(10) | unused |
| crossref | varchar(75) | unused |
| uniprot_id | varchar(10) | unused |
| ensembl_id | varchar(20) | unused |
| other_ref | varchar(20) | unused |
| protein | text | unused |
| source | varchar(255) | unused |
| expression_sys | text | unused |
| mut_PMD | varchar(20) | unused |
| mut_real | varchar(10) | the mutant in the form XposY |
| function | text | effect on protein function |
| fb | tinyint(4) | unused |
| structure | text | effect on protein structure |
| strB | tinyint(4) | unused |
| stability | text | effect on protein stability |
| staB | tinyint(4) | unused |

| | | |
|---|---|---|
| expression | text | unused |
| eB | tinyint(4) | unused |
| transport | text | unused |
| tB | tinyint(4) | unused |
| maturation | text | unused |
| mB | tinyint(4) | unused |
| disease | text | disease associated with that mutant |
| dB | tinyint(4) | unused |
| uni_real | varchar(20) | unused |
| uni_realid | varchar(20) | unused |
| uni_start | int(11) | unused |
| uni_finish | int(11) | unused |
| uni_loc | int(11) | unused |
| ens_real | varchar(20) | unused |
| ens_organism | int(11) | unused |
| ens_start | int(11) | unused |
| ens_finish | int(11) | unused |
| ens_loc | int(11) | unused |
| pos_real | smallint(5) | mutant position, 0-based |
| mt_real | char(1) | mutant residue |
| taxid | int(11) | NCBI taxid |

| SNPs_1KG<br>(collects all SAASs mapped from 1KG onto RefSeq) | | |
|---|---|---|
| mut_id | int(11) | unique mutant id; ref to geno2func |
| **md5** | varchar(32) | sequence md5 hashsum |
| protein_acc | varchar(32) | RefSeq protein id |
| protein_ver | int(11) | RefSeq protein ver |
| mrna_acc | varchar(32) | RefSeq mrna id |
| exon | int(10) | Number of exon |
| locus_symbol | varchar(64) | locus |
| chromosome | int(10) | Number of chromosome |
| mrna_wt | char(1) | Wild type nucleotide in the mRNA |
| mrna_pos | int(11) | Mutant position in the mRNA (0-based) |
| mrna_mt | char(1) | Mutant nucleotide in the mRNA |
| chrom_wt | char(1) | Wild type nucleotide in the chromosome |
| chrom_pos | int(11) | Mutant position in the chromosome (0-based) |

| | | |
|---|---|---|
| chrom_mt | char(1) | Mutant nucleotide in the chromosome |
| LDAF | float | MLE Allele Frequency Accounting for LD (from 1KG vcf) |
| AC | int(10) | Alternate Allele Count (from 1KG vcf) |
| AN | int(10) | Total Allele Count (from 1KG vcf) |
| wt | char(1) | Wildtype amino acid |
| **pos** | int(11) | Mutant position in protein sequence (0-based) |
| **mt** | char(1) | Mutant amino acid |

| SP_AC_map<br>(maps the primary SwissProt accession with alternatives) | | |
|---|---|---|
| **first** | varchar(10) | primary accession |
| **additional** | varchar(10) | alternative accession |