CGDSNP: A High Quality Integrated Mouse SNPs Database

Lucie N. Hutchins, Gary Churchill, Joel H. Graber

The Jackson Laboratory , Bar Harbor, Maine 04609 USA

Abstract

Single nucleotide polymorphisms (SNPs) are an important tool to study genetic variations. We present CGDSNP, a SNP database with more than 8 Millions SNPs from 74 strains of laboratory mice, drawn from several sources. These SNPs have been quality checked , integrated with nearby gene annotations and various functional characteristics and implications , such as amino acid changes, base-pair substitution types, and location within CpG dinucleotides. CGDSNP is accessible online via a web-based query tool.

CGDSNP also serves as the interface tool to a recently developed "imputed SNP resource" , in which a Hidden Markov Model (HMM) was used to assess local haplotypes and the most probable base assignment at several million genomic loci in several tens of strains of mice (Szatkiewicz et al. 2008).

The CGDSNP database consists of:

- **The core tables, containing all pertinent data for the SNP, including all data from the source download; all such data is quality checked for consistency prior to loading.**
- **Gene related tables that allow linking of SNPs, Genes and functional category**
- **Tables holding SNP functional implications, produced from our own analysis of SNP features such coding changes, CpG dinucleotides, etc.**

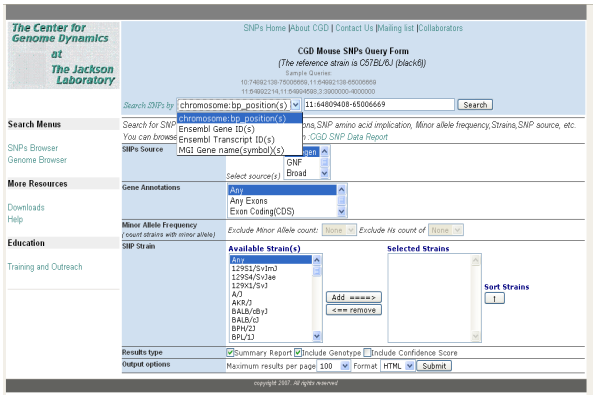
The web-based interface facilitates a number of different queries, including search by chromosome region(s) or nearby gene annotations. The query tool can return its results in multiple popular formats.

Jin P. Szatkiewicz , Glen L. Beane, Yueming Ding, Lucie Hutchins , Fernando Pardo-Manuel de Villena and Gary A.Churchill (2008)

An imputed genotype resource for the laboratory mouse. Mammalian Genome 19 (3):199-208.

The Web Tool

http://cgd.jax.org/cgdsnpdb/



You can query for SNPs using:

- Chromosome location ,range
- Ensembl Gene IDs
- Ensembl Transcript IDs
- MGI Gene symbol

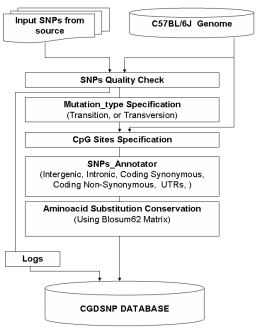
You can filter query results by:

- SNP source
- SNP location
- SNP function class
- Strains

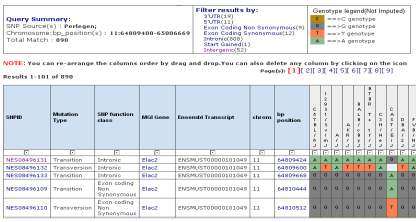
You can view query results in:

- HTML format
- CSV format

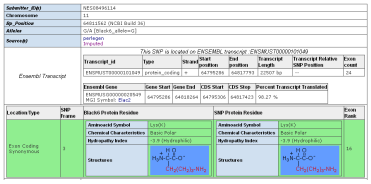
The Process



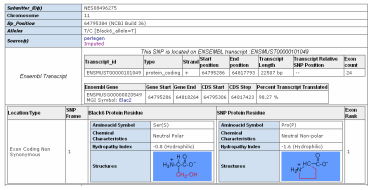
Query Summary Page



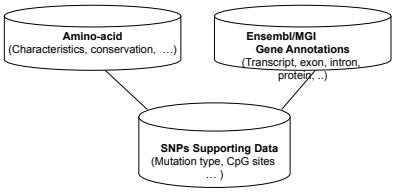
SNPs Coding Synonymous Amino Acids



SNPs Coding Non-Synonymous Amino Acids

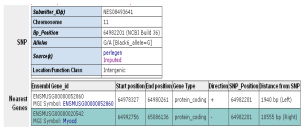


Data Integration



SNPs Brief Summary	
Total SNPs	8,316,925
Total Transition SNPs	5,702,603 (68.6%)
Total Transversion SNPs	2614322 (31.4%)
Total Intergenic SNPs	4,990,628 (60%)
Total SNPs On Genes	3,326,297 (40%)
Total Strains	74
Number of Sources	4

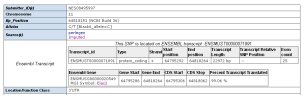
Intergenic SNPs



We provide Nearest genes information:

- Nearest gene left and right of SNP
- Distance in basepairs from SNP
- Gene orientation

Intronic and UTR SNPs



We provide gene information:

- Gene coordinates
- Transcript coordinates
- CDS start and stop
- Transcript exon count
- Transcript length

Future Plans and Upgrades

- **Further details of coding SNPs**
  - **Comparison of codon frequencies for both synonymous and non-synonymous substitutions**
  - **Display the amino acid Blosum62 substitution score**
- **Additional search capabilities**
  - **SNP RS number**
  - **Entrez Gene**
  - **Unigene ID**
- **Further details of SNP context**
  - **Display intron rank for intronic SNPs**
  - **Overlay SNPs with predictions of functional elements such as splice sites, promoters, polyA signals, etc.**
- **Additional features as requested by users**
- **Additional sources of SNPs**