CGDSNP: A High Quality Integrated Mouse SNPs Database

Lucie N. Hutchins, Yueming Ding, Jin P. Szatkiewicz, Randy Smith, Fernando P.M. de Villena, Gary Churchill, Joel H. Graber
The Jackson Laboratory , Bar Harbor, Maine 04609 USA

The Center for Genome Dynamics at The Jackson Laboratory

Abstract

Single nucleotide polymorphisms (SNPs) are an important tool to study genetic variations. We present CGDSNP (http://cgd.jax.org/cgdsnpdb/), a high quality open source Mouse SNP database with more than 8 Million SNPs from 74 inbred strains of laboratory mice, drawn from several sources.

These SNPs have been quality checked and curated, integrated with nearby gene annotations and various functional characteristics and implications of the SNPs, such as amino acid changes and conservation , base-pair substitution types, Amino acid codon frequency , and location within CpG dinucleotide sites. CGDSNP is accessible online via the CGDSNP-viewer (web-based query tool). You can also access our database via SQL using our mySQL public login.

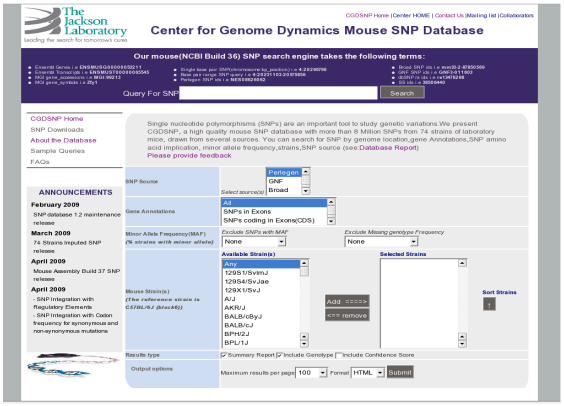
CGDSNP also serves as the interface tool to both the Mouse Diversity Genotyping Array and the "imputed SNP resource" in which a Hidden Markov Model (HMM) was used to assess local haplotypes and the most probable base assignment at several million genomic loci in tens of strains of mice. The imputed SNP calls may be searched, retrieved, and analyzed identically to the experimentally verified SNPs, with the additional information, such as HMM likelihood score provided in the query return.

The search engine facilitates a number of different queries, including search by chromosome region(s), nearby gene annotations, or SNP accession ID. The query tool can return its results in multiple popular formats. CGDSNP was implemented using open source LAMP technologies.

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/ gene_id
- SNP accession Id (SS, RS, Perlegen)

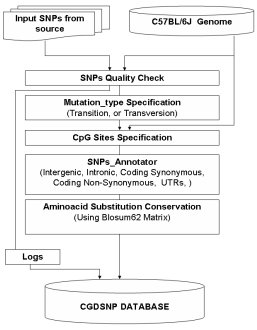
You can filter query results by:

- SNP source
- SNP location
- SNP function class
- Strains
- Minor Allele Frequency (MAF)
- Fraction of missing genotype

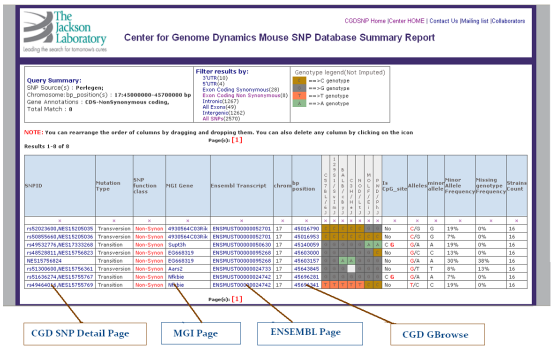
You can view query results in:

- HTML format
- CSV format

The Process



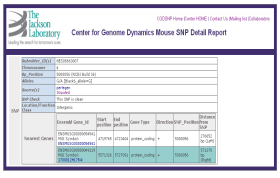
Query Summary Page



We provide the following information :

- Minor Allele Frequency (MAF)
- Missing Genotype fraction
- SNP Conflict(s)
- CpG site
- Mutation Type (Transition/Transversion)
- Link to CGD SNP Detail page
- Link to MGI Gene Detail page
- Link to ENSEMBL Transcript page
- Link to CGD Genome Browse

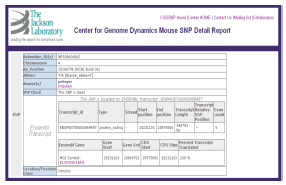
Intergenic SNP Detail Page



We provide Nearest genes information:

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

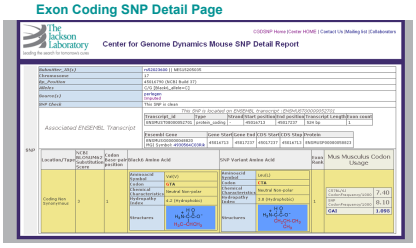
Intronic and UTR SNP Detail Page



We provide gene information:

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

Exon Coding SNP Detail Page



CGDSNP Brief Summary

Total SNPs	**8,306,008**
Total Transition SNPs	**5,696,796 (~68.6%)**
Total Transversion SNPs	**2,609,212 (~31.4%)**
Total Intergenic SNPs	**4,722,768 (~56.9%)**
Total SNPs On Transcripts	**3,583,240 (~43.1%)**
Total Strains	**97**
SNPs Sources Count	**5**

Future Plans and Upgrades

- **Additional search capabilities**
- **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**