



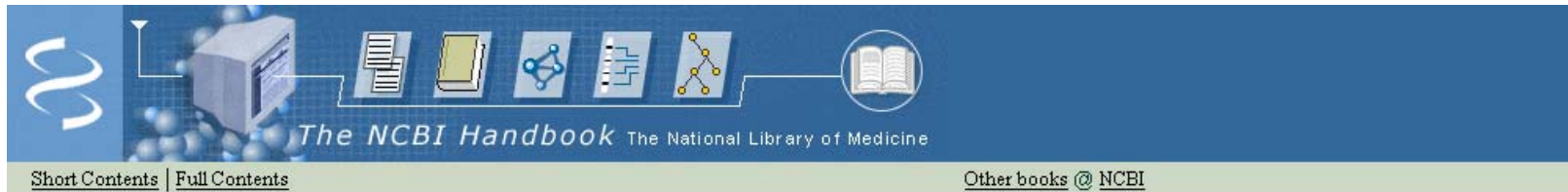
## Outline

- What is dbSNP?
- What is the scope of dbSNP?
- How does dbSNP differ from HapMap?

Is it limited to representing single nucleotide polymorphisms?

Is the primary data storage a location on an annotated chromosome?

How does an rs identifier compare to an ss identifier?



<b>Navigation</b>
<a href="#">About this book</a>
<b>Part 1. The Databases</b>
<a href="#">5. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation</a>
<a href="#">Introduction</a>
<a href="#">Searching dbSNP</a>
<a href="#">Submitted Content</a>
<a href="#">Computed Content (The dbSNP Build Cycle)</a>
<a href="#">dbSNP Resource Integration</a>

**The NCBI Handbook** → **Part 1. The Databases**

Created: October 09, 2002

Updated: September 13, 2006

## 5. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation

by Adrienne Kitts and Stephen Sherry

### Summary

The Single Nucleotide Polymorphism database (dbSNP) is a public-domain archive for a broad collection of simple genetic **polymorphisms**. This collection of polymorphisms includes single-base nucleotide substitutions (also known as single nucleotide polymorphisms or SNPs), small-scale multi-base deletions or insertions (also called deletion insertion polymorphisms or DIPs), and retroposable element insertions and microsatellite repeat variations (also called short tandem repeats or STRs). Please note that in this chapter, you can substitute any class of variation for the term SNP. Each dbSNP entry includes the sequence context of the polymorphism (i.e., the surrounding sequence), the occurrence frequency of the polymorphism (by population or individual), and the experimental method(s), protocols, and conditions used to assay the variation.

<b>Search</b>
<input type="text"/>
<input type="button" value="Go"/>
<input checked="" type="radio"/> This book <input type="radio"/> All books
<input type="radio"/> PubMed

polymorphism (by population or individual), and the experimental method(s), protocols, and conditions used to assay the variation.

dbSNP accepts submissions for variations in any species and from any part of a genome. This document will provide you with options for finding SNPs in dbSNP, discuss dbSNP content and organization, and furnish instructions to help you create your own (local) copy of dbSNP. [↑ TOP](#)

Is it limited to representing single nucleotide polymorphisms?

Is the primary data storage a location on an annotated chromosome?

How does an rs identifier compare to an ss identifier?

## Submitted Content

The SNP database has two major classes of content: the first class is submitted data, i.e., original observations of sequence variation; and the second class is computed content, i.e., content generated during the dbSNP “build” cycle by computation on original submitted data. Computed content consists of refSNPs, other computed data, and links that increase the utility of dbSNP.

A complete copy of the SNP database is publicly available and can be downloaded from the SNP [FTP](#) site (see the section *How to Create a Local Copy of dbSNP*). dbSNP accepts submissions from public laboratories and private organizations. (There are online [instructions](#) for preparing a submission to dbSNP.) A short tag or abbreviation called Submitter HANDLE uniquely defines each submitting laboratory and groups the submissions within the database. The 10 major data elements of a submission follow.

## Flanking Sequence Context DNA or cDNA

The essential component of a submission to dbSNP is the nucleotide sequence itself. dbSNP accepts submissions as either genomic DNA or cDNA (i.e., sequenced mRNA transcript) sequence. Sequence submissions have a minimum length requirement to maximize the specificity of the sequence in larger contexts, such as a reference genome sequence. We also structure submissions so that the user can distinguish regions of sequence actually surveyed for variation from regions of sequence that are cut and pasted from a published reference sequence to satisfy the minimum-length requirements. [Figure 1](#) shows the details of flanking sequence structure. [↑ TOP](#)

[http://www.ncbi.nlm.nih.gov/SNP/snp\\_ref.cgi?rs=7765803](http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=7765803)

Is it limited to representing single nucleotide polymorphisms?

Is the primary data storage a location on an annotated chromosome?

How does an rs identifier compare to an ss identifier?

## Computed Content (The dbSNP Build Cycle)

### Submitted SNPs and Reference SNP Clusters

Once a new SNP is submitted to dbSNP, it is assigned a unique submitted SNP ID number (ss#). Once the ss number is assigned, we align the flanking sequence of each submitted SNP to its appropriate genomic contig. If several ss numbers map to the same position on the contig, we cluster them together, call the cluster a “reference SNP cluster”, or “refSNP”, and provide the cluster with a unique RefSNP ID number (rs#). If only one ss number maps to a specific position, then that ss is assigned an rs number and is the only member of its RefSNP cluster until another submitted SNP is found that maps to the same position.

on the [FLIP](#) site, and deliver them as sets of results when a user conducts a dbSNP batch query. We also maintain both refSNPs and submitted SNPs in FASTA databases for use in [BLAST](#) searches of dbSNP. [↑ TOP](#)

# dbSNP and HapMap

Attribute	HapMap	dbSNP
Species	Human	Multiple species
Coverage	Validated, $\geq 5\%$ polymorphism (subset of dbSNP)	Archive of all submissions
Genotype Source	Probe / primer based assays Illumina, Perlegen, etc. Encode region resequencing	Genotype assays as well as resequencing from projects like PGA EGP
Human Individuals	270 $\rightarrow$ 4 sample sets	2000 Individuals /Cell lines $\rightarrow$

# Getting started: Multiple ways to search

Entrez

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Snps&cmd=Limits>

dbSNP

<http://www.ncbi.nlm.nih.gov/SNP/>

BLAST

[http://www.ncbi.nlm.nih.gov/SNP/snp\\_blastByOrg.cgi](http://www.ncbi.nlm.nih.gov/SNP/snp_blastByOrg.cgi)

e-Utilities

<http://www.ncbi.nlm.nih.gov/SNP/SNPeutils.htm>

# What is new?

Genotype query form

[http://www.ncbi.nlm.nih.gov/projects/SNP/snp\\_gf.cgi](http://www.ncbi.nlm.nih.gov/projects/SNP/snp_gf.cgi)

Pedigree viewer:

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Snp&list\\_uids=13318299&dopt=GEN](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Snp&list_uids=13318299&dopt=GEN)

Genome WorkBench (**VERY BETA, Bleeding edge!**)

<http://www.ncbi.nlm.nih.gov/projects/gbench/>

# Genome Workbench and dbSNP

1. Launch GBench
2. Click on Search
3. Query LPA from Entrez Gene
4. Select the human record
5. Right click on the selection and use the default (Add to project)
6. Select the default (New project)
7. Select the query, right click, Add new view, Graphic view
8. At the 5' end of the gene, zoom in
9. Drag your mouse over the SNPs, and select several (you should see vertical lines appear that project to the sequence)
10. Tools->LinkOuts->dbSNP
11. Generate a genotype report
12. Generate an LD graph/download files for haploview